Frege Puzzles in Metaphysical Debates

Elisabetta Sassarini

Jesus College

University of Oxford

Thesis Submitted for the Degree of

Doctor of Philosophy

July 2025

Supervised by Professor Timothy Williamson and Dr. Matthew Parrott

To my brother, Michele

Contents

Abstract

Frege puzzles arise in *opaque* contexts—linguistic contexts in which sentences that differ only in co-referential expressions appear to differ in truth value. These cases challenge the principle of compositionality, which holds that the semantic value of a complex expression is determined by the semantic values of its parts, and thus expressions with the same semantic value should be freely substitutable without affecting truth value. Propositional attitude verbs such as "know" and "believe", as well as explanatory terms like "because", are taken to generate opacity: both attitude ascription and explanation appear to be sensitive not only to the referents of the relevant expressions, but also to the manner in which those referents are represented.

This dissertation examines a range of philosophical problems through the framework of Frege puzzles. Chapter 1 discusses verbal disputes; Chapter 2 addresses the Knowledge Argument against physicalism; Chapter 3 analyses reductive accounts of metaphysical explanation. The dissertation aims to show, first, that the semantics of propositional attitude ascriptions plays a central role in these debates, and that the intuitions driving them frequently arise from the presence of expressions that give rise to apparently opaque contexts. Second, it argues that potential solutions to the problems in question—including the relevant error theories—often amount, in effect, to familiar strategies for resolving Frege puzzles.

By drawing out these connections, the dissertation aims to offer a unifying framework that brings together seemingly unrelated issues, while providing insights into the scope and significance of Frege puzzles and the conceptual tools involved in their formulation. More broadly, this dissertation seeks to motivate an anti-exceptionalist stance

to the problems discussed. Recognizing these issues as manifestations of a familiar phenomenon reduces the need for ad hoc theoretical devices tailored to each individual case.

Acknowledgments

Completing this thesis and my DPhil would not have been possible without the support and encouragement of many people, to whom I owe my heartfelt thanks.

First and foremost, I am deeply grateful to my supervisors, Tim Williamson and Matt Parrott, for invaluable discussions, for their guidance and patience, and for being incredibly generous with their time, all while giving me the space to develop my own ideas. I have learned a great deal from them—philosophically, professionally, and personally.

I owe special thanks to my viva examiners, Nick Jones and Robbie Williams, for their thoughtful engagement with my work. Their questions and comments were extremely helpful in refining my ideas, and they made the viva a genuinely stimulating experience.

For insightful comments and discussions on parts of this thesis, I also thank Riccardo Baratella, Kenneth Black, Mattia Cecchinato, Matti Eklund, Filippo Ferrari, James Glover, Jonas Hertel, Daniel Kodsi, Boaz Laan, Olle Lövgren, Mariona Miyata-Sturm, Bernhard Salow, Giuseppe Spolaore, Alessandro Torza, Victor Verdejo, Kathie Zhou, as well as the audiences at the Oxford DPhil seminar, the Open Minds XVI Postgraduate Conference in Manchester, the 15th SIFA Conference, the Harvard-MIT Graduate Conference, the 1st Parma Workshop in Analytic and Scientific Metaphysics, the ENFA 9th National Meeting in Analytic Philosophy in Lisbon, the International Society for the Philosophy of the Sciences of the Mind Annual Conference, the Seminario di Filosofia Analitica in Bologna, and the 16th SIFA Conference.

To my Oxford friends—Mattia Cecchinato, James Glover, Mariona Miyata-Sturm, Boaz Laan, Dominik Ehrenfels, Kassandra Dugi, Jonas Hertel, Amit Karmon, Valentino Gargano, and Corine Besson—thank you for your friendship, and for the countless much-needed pub nights and coffee breaks. You truly helped make Oxford feel like home.

To my friends in Italy—Letizia Emultifiori, Arianna Nicolini, Ilaria Salvi, Margherita Mattioni, Ines Zampaglione, Cosimo Ferrari, Michele Banelli, and Matteo Cervetti—thank you for being a steady presence in my life despite the distance. It's a rare and precious kind of friendship that stays just as strong across years and miles.

I'm especially grateful to Filippo Ferrari for his unconditional support and for the many philosophical conversations we've shared over the past seven years. Without him, I might never have considered pursuing a PhD in the first place.

To my partner, Andrea, thanks for standing by me through the highs and lows alike, and for making everything lighter and brighter.

Finally, I am deeply thankful to my family—especially my mum and dad—for their love, care, and strength even through the toughest of times. Thanks for being such extraordinary parents.

Introduction

Philosophical inquiry is often structured around puzzles that expose tensions between widely accepted general principles and our intuitive judgements about certain problematic cases. Among these, Frege puzzles stand out for their profound impact across multiple research areas, including philosophy of language, epistemology, linguistics, and cognitive science. Formulated by Gottlob Frege in 1892, these puzzles have played a central role in philosophical discussions on meaning, cognitive significance, and the ascription of propositional attitudes for over a century. This dissertation argues that various classic problems in philosophy can be fruitfully understood as instances of Frege puzzles. This perspective not only clarifies the debates in question by revealing their structural analogies and offering a unified approach to their potential solutions, but also deepens our understanding of Frege puzzles themselves, underscoring their generality and highlighting the need for further work on central notions.

§1 Frege Puzzles

In *Über Sinn und Bedeutung* ("On Sense and Reference", 1892), Gottlob Frege introduced two puzzles purporting to show that meaning cannot be reduced to reference alone. To solve these puzzles, Frege suggested that the terms of a language have both a sense ("Sinn") and a denotation, or reference ("Bedeutung"), where sense is a *mode of presentation* of the reference. The first puzzle concerns the meaning of identity statements. Consider:

1. Hesperus is Hesperus.

2. Hesperus is Phosphorus.

Both “Hesperus” and “Phosphorus” refer to the planet Venus. However, (1) is trivial and, supposedly, knowable a priori, whereas (2) is informative, and, allegedly, only knowable a posteriori. Unlike (1), (2) seems to express an astronomical discovery which took substantial empirical work to make. Yet, if meaning were simply reference, (1) and (2) would have the same meaning, and thus convey the same information. In general, identity statements of the form “a=a” potentially differ in cognitive significance from statements of the form “a=b”, even though “a” and “b” are co-referential. The idea is that meaning should reflect cognitive significance, but reference alone does not.

The second puzzle involves the ascription of propositional attitudes such as belief, desire, and knowledge. Suppose Jane has a belief that she would express with “Hesperus is visible in the evening”, but has never heard the name “Phosphorus”, and is not disposed to assent to “Phosphorus is visible in the evening”. Consider the following pair of statements:

3. Jane believes that Hesperus is visible in the evening.

4. Jane believes that Phosphorus is visible in the evening.

In this case, our ordinary heuristics for belief ascription yield that (3) is true, whereas (4) is false. Thus, although “Hesperus” and “Phosphorus” refer to the same object, substituting one for the other seems to change the truth value of the belief report.

These cases highlight a tension between, on the one hand, our intuitions concerning meaning and propositional attitude ascriptions and, on the other hand, the semantic principle of compositionality. The principle of compositionality states that the meaning of a complex

expression is determined by the meanings of its constituent expressions and the way these are put together. Thus, sentences which only differ in expressions with the same semantic value cannot express different propositions. This also entails that terms with the same semantic value can be substituted in any sentence *salva veritate*—that is, without altering the truth-value of the sentence.[1] In general, if S is a sentence, "a" and "b" are names, and S(a) differs from S(b) only by the fact that at least one occurrence of "a" replaces "b", S(a) and a=b entail S(b). This principle captures the idea that if we say something true about an object, then even if we change the name by which we refer to that object, we should still be saying something true about it (Zalta 2024).

Frege's cases challenge compositionality, as sentences like (1) and (2) only differ in co-referential expressions, yet they appear to differ in meaning, or the proposition they express. Likewise, (3) and (4) differ only in co-referential expressions but seem to have different truth-values. Similarly, if "Hesperus" and "Phosphorus" share the same semantic value, then "Hesperus is visible in the evening" and "Phosphorus is visible in the evening" should have the same meaning—that is, they should express the same proposition. If so, it is unclear how "Jane believes that Hesperus is visible in the evening" and "Jane believes that Phosphorus is visible in the evening" could differ in truth-value, given that the object of Jane's belief is the same proposition in both cases.

Frege's solution was that while the names "Hesperus" and "Phosphorus" share the same reference, they have different senses, as they present their referent, Venus, in different ways. Specifically, the familiar story is that "Hesperus" presents Venus as the evening star, whereas "Phosphorus" presents it as the morning star. Sense is a *semantic* notion serving

[1] This does not apply to quotational contexts. Of course, the co-referentiality of "Hesperus" and "Phosphorus" is irrelevant to whether they can be substituted *salva veritate* in the sentence "The name 'Hesperus' has eight letters".

multiple purposes. First, a term's sense is meant to be a mode of presentation of its referent and to *uniquely* determine it. The sense of a term is sometimes understood as a sort of descriptive condition that its referent uniquely satisfies. Second, any difference in sense should correspond to a difference in cognitive significance, meaning that sense supervenes on cognitive significance. Finally, although senses are "perspectival" ways of thinking about objects, they, unlike what Frege calls "ideas", are meant to be sharable—different individuals can think of an object in the same perspectival way.

Frege extended his analysis to entire sentences: in his view, the reference of a complete sentence is its truth value, while its sense is the thought it expresses. This accounts for how two sentences can share the same truth value while differing in meaning. In this case, the idea is that due to the difference in sense between "Hesperus" and "Phosphorus", (1) and (2) express different thoughts—or propositions—without violating compositionality. Moreover, Frege argued that within the scope of a propositional attitude verb, terms refer to their customary senses rather than to their customary referents. Thus, sentences like (3) and (4) can differ in truth value without violating compositionality because compositionality only requires that terms *with the same semantic value* be substitutable *salva veritate*. In other words, due to the difference in semantic value between "Hesperus" and "Phosphorus", (3) and (4) express different propositions, and may therefore differ in truth value: the difference in meaning between the two terms is reflected in a difference in the propositions expressed at the sentential level.

Frege's puzzles have been taken to highlight the limitations of purely referential theories of meaning and have played a central role in debates within the philosophy of language and epistemology. Later philosophers—including Russell (1905), Putnam (1975), Kripke (1979, 1980), and Kaplan (1989)— have built upon or challenged Frege's

framework, shaping developments in theories of reference, rigid designation, and direct reference.

For independent reasons, extensional semantic frameworks are now widely rejected, along with the idea that a term's meaning can be identified with its referent. In particular, such frameworks proved inadequate in cases involving contingently co-extensional terms. Suppose that the predicates "has a heart" and "has a liver" are contingently co-extensional (Quine 1970). If the semantic value of a predicate were merely its extension, then "has a heart" and "has a liver" would have the same semantic value—that is, the same meaning. However, the properties of having a heart and having a liver are, arguably, distinct: there are possible organisms that have a heart but no liver, and vice versa. This inadequacy—along with the independent development of possible-world semantics—led to the replacement of extensional semantics with intensional frameworks. In these frameworks, a term's meaning is understood as a function from possible worlds to extensions, and the meaning of a sentence—the proposition it expresses—is understood as the set of worlds where the sentence is true. However, this shift does not resolve Frege's puzzles, as the puzzle also arises for co-intensional terms and sentences. For example, since "Hesperus" and "Phosphorus" are co-intensional, sentences (1) and (2) above are likewise co-intensional. Thus, the original puzzle remains unsolved.

§2 Proposed Solutions

Contemporary approaches to Frege puzzles can be divided into two broad camps, encompassing a wide range of proposals. On one side, Fregean frameworks maintain that differences in cognitive significance entail differences in meaning. Proponents of these

views agree that terms like "Hesperus" and "Phosphorus" pick out Venus via different *modes of presentation,* and that these modes of presentation play a semantic role, determining a level of meaning beyond reference and intension (see Forbes 1990, Chalmers 2011b). As a result, in line with Frege's original view, "Hesperus" and "Phosphorus" make different semantic contributions to the relevant sentences, despite being co-referential and co-intensional. At the sentential level, this means that "Hesperus is bright" expresses a proposition that differs from the proposition expressed by "Phosphorus is bright", even though the two are co-intensional—that is, true at the same possible worlds.

Anti-Fregeans, by contrast, take Fregeans to mistakenly project onto semantics fine-grained distinctions which pertain to the cognitive and linguistic domains (Braun 1998, Salmon 1986, Williamson 2021a). Anti-Fregeans favour coarser-grained semantic frameworks than those of Fregeans, such as Russellianism and intensionalism. Russellianism, broadly speaking, holds that propositions are structured entities whose constituents are the very objects and relations to which the terms involved refer. Since "Hesperus" and "Phosphorus" are co-referential, "Hesperus is visible in the evening" and "Phosphorus is visible in the evening" express the same proposition—one composed of Venus and the property of being visible in the evening. Similarly, "Hesperus is Hesperus" and "Hesperus is Phosphorus" express a proposition whose constituents are Venus and the identity relation. Intensionalism, as mentioned, treats propositions as sets of possible worlds. Given that Hesperus and Phosphorus are identical, all worlds in which Hesperus is visible in the evening are also worlds in which Phosphorus is visible in the evening. Consequently, in an intensionalist framework, "Hesperus is visible in the evening" and "Phosphorus is visible in the evening" express the same proposition—namely, the set of worlds where Venus is visible in the evening. Likewise, "Hesperus is Hesperus" and "Hesperus is Phosphorus" correspond to the same set of worlds: the set of all possible worlds. Of course, anti-Fregeans

acknowledge that the content of linguistic expressions can be presented under different modes of presentation, or guises, which also reflect their cognitive significance. However, unlike Fregeans, they argue that modes of presentation do not play a *semantic* role and are therefore irrelevant to determining the meaning of terms and sentences.

When it comes to propositional attitude ascriptions, however, we must distinguish between two versions of anti-Fregeanism (Williamson 2021a). On *contextualist* anti-Fregean views (e.g. Crimmins & Perry 1989, Richard 1990, Schiffer 1992), propositional attitude ascriptions are sensitive to the contextually relevant modes of presentation of the proposition in question. In the case of belief, the idea is that "S believes that p" is true in a context c if and only if S believes the proposition that p under some mode of presentation relevant in c.[2] This view can take the form of a hidden-indexical theory (e.g. Crimmins & Perry 1989, Schiffer 1992), according to which an utterance of "S believes that p" is true just in case the utterer is implicitly referring to a mode of presentation *m* and S believes the proposition referred to by "p" under *m*. Alternatively, Richard has proposed a contextualist account in terms of RAMs (Russellian Annotated Matrices). For Richard, RAMs are structured complexes pairing word-types or mental representations with semantic values. For a belief ascription to be true, the RAM it involves must *correlate* with a RAM believed by the subject of the ascription, where what correlates with what is context-dependent.

On this account, "Jane believes that Hesperus is visible in the evening" is true in c because Jane believes the proposition p expressed by "Hesperus is visible in the evening" under the mode of presentation (or sentential guise) "Hesperus is visible in the evening". However, Jane may not believe p under the mode of presentation "Phosphorus is visible in

[2] One way to articulate this idea is to say that "S believes that p" is true in a context just in case S believes that p and is disposed to assent to some contextually relevant sentence expressing p. This is what Williamson (2021a) refers to as a "language-sensitive" account of belief-ascription.

the evening"—perhaps because she is not disposed to assent to the relevant sentence. If so, "Jane believes that Phosphorus is visible in the evening" is false, whereas "Jane believes that Hesperus is visible in the evening" is true, even though "Hesperus is visible in the evening" and "Phosphorus is visible in the evening" both express p.[3]

Finally, we turn to *anti-contextualist* anti-Fregeanism (Braun 1998, Salmon 1986, Williamson 2021a). Proponents of this view adopt a "coarse-grained" anti-Fregean semantics (e.g. Russellianism or intensionalism), according to which "Hesperus is visible in the evening" and "Phosphorus is visible in the evening" express the same proposition, just like "Hesperus is Hesperus" and "Hesperus is Phosphorus". Combined with an anti-contextualist account of propositional attitude ascription—on which the ascription of propositional attitudes is *not* sensitive to the mode of presentation (or guise) of the proposition believed—this account entails that sentences like (3) and (4) cannot differ in truth-value: Jane's believing that Hesperus is visible in the evening entails that she believes that Phosphorus is visible in the evening, regardless of whether she has ever heard the term "Phosphorus" at all. In fact, even if Jane were convinced that Phosphorus is *not* visible in the evening—and were disposed to assent to "Phosphorus is not visible in the evening"—she would still, on this account, believe that Phosphorus is visible in the evening, simply by virtue of believing that Hesperus is.

This is because, on the one hand, proponents of this view maintain the ordinary heuristic for belief ascription, whereby (a disposition to) assent to a sentence S entails belief

[3] There are other ways one might attempt to restrict an account of belief ascription. For example, an alternative view could hold that, in order to be ascribed a belief in a proposition, one must believe that proposition under every guise one possesses. On such a view, for Jane to be ascribed the belief that Phosphorus is visible in the evening, she must believe that proposition under every guise she possesses, including "Phosphorus is visible in the evening". Thus, if Jane has never heard the name "Phosphorus", she cannot be said to believe that Phosphorus is visible in the evening. In this dissertation, I will set aside similar proposals.

in the proposition p expressed by S. On the other hand, contrary to contextualists, anti-contextualists hold that a lack of any disposition to assent to S does not entail a failure to believe p. In this case, failure to assent to "Phosphorus is visible in the evening" does not entail a failure to believe that Phosphorus is visible in the evening, as that proposition might be believed under a different guise. Thus, anti-contextualist anti-Fregeanism entails that one may rationally believe both p and not-p under different guises.

Many consider this outcome a significant problem for anti-contextualist anti-Fregeanism, as ascriptions of inconsistent beliefs are typically viewed as attributions of irrationality. In response, proponents of this view claim that believing a proposition and its negation is not, by itself, sufficient for irrationality: irrationality arises only when one believes a proposition and its negation *under the same guise, or mode of presentation* (see Nelson 2024).

Assessing the advantages and disadvantages of the various competing solutions to Frege puzzles is beyond the scope of this introduction and, indeed, of this dissertation. Nonetheless, we should briefly note that the main advantage of Fregeanism is that it seems to accommodate our intuitive judgements about attitude ascriptions and differences in meaning.

However, Fregean sense is problematic for several reasons. Kripke (1980) offers powerful arguments to the effect that definite descriptions can neither be synonymous with names nor serve to fix their reference. For example, the name "Gödel" would still refer to Gödel even if it turned out that he was not, in fact, the person who proved the incompleteness of arithmetic, and someone else was. Therefore, Kripke argues, the description "the man who proved the incompleteness of arithmetic" cannot serve as either the meaning of "Gödel" or as a means of fixing its reference. Moreover, Fregean senses face additional challenges when applied to indexical terms. For example, the sense of indexical terms like "I", "here",

or "now" cannot be what Kaplan (1989) calls "character"—a function from contexts to contents. In the case of "I", for instance, character is a function mapping each context to the agent of that context. One might thus be tempted to say that the sense of "I" is captured by the description "the person who utters this token". Yet, this description does not *uniquely* determine a referent—it only does so within a specific context. More importantly, as Kaplan notes, this descriptive condition cannot be the meaning of the term "I". If "I" meant "the person who utters this token", then in a situation where no one utters the token, the speaker would not exist (Kaplan 1989).[4]

Additionally, on the usual formulations of the Fregean theory, the locution "a believes that b is φ" attributes to the referent of *a* a belief whose content is a proposition, or "thought", made up in part of the sense of the singular term *b.* But the sense, or conceptual content, attached to a proper name, as used with a particular reference, may vary significantly from speaker to speaker. The same is true of demonstratives and indexicals such as "I", "you", and "here". For example, it is extremely unlikely that I use the name "Socrates" with Plato's sense, attaching to it Plato's conceptual content for the ancient Greek version of "Socrates". I use the name with my own conceptual content. If the singular term *b* occurring in "a believes that *b* is φ" is a proper name, it is used there to refer to the speaker's sense for the name. The conceptual content which the subject of the attribution (the referent of *a*) happens to attach to *b* is entirely irrelevant to the attribution. Hence, the Fregean theory

[4] It is also worth saying that there is no consensus on how to understand Fregean senses, particularly on whether they should be analysed as descriptions (e.g. Dummett 1973, Evans 1982). Some contemporary Fregeans adopt two-dimensional semantic frameworks, in which so-called "primary intensions"—which can be modelled as functions from worlds considered as actual to extensions—play the role of Fregean senses (e.g. Chalmers 2002). Among two-dimensionalists, some hold that primary intensions should be statable as descriptions, while others reject this view. Non-descriptivist versions of two-dimensionalism might avoid some of the anti-descriptivist arguments mentioned above. These issues have sparked an extensive and complex debate, which goes beyond the scope of this introduction.

seems to entail that if I utter the sentence "Plato believed that Socrates is wise" I attribute to Plato a belief made up in part of *my* concept of Socrates, the sense *I* attach to the name "Socrates". Almost certainly, Plato had no such belief (Salmon 1986).

On the other hand, as explained above, anti-Fregeanism (especially in its anti-contextualist variant) is in tension with our intuitive judgements about meaning and attitude ascriptions. Thus, proponents of this view need to provide error theories for our intuitions in the relevant cases. Some appeal to pragmatics, arguing that speakers systematically confuse the semantic content of the ascriptions in question with the information pragmatically conveyed by their utterance (e.g. Salmon 1986). By contrast, Williamson (2021a, 2024) argues that our intuitions in Frege cases are the product of our ordinary heuristics for attitude ascriptions, which lead us astray in such cases. Despite its counterintuitive consequences, one of the advantages of anti-Fregeanism is that it complies with compositionality without requiring a further fine-grained and representation-sensitive layer of meaning, such as Fregean sense.

Another general principle challenged by Frege puzzles is Leibniz's Law, which states that two objects are identical just in case they share the same properties. That is, for every property *F* and objects x and y:[5]

$$\forall x \forall y(x{=}y \leftrightarrow \forall F(Fx \leftrightarrow Fy))$$

Frege puzzles for propositional attitude ascription challenge this general principle. For instance, the case of (3) and (4) suggests that Hesperus might be such that Jane believes it is visible in the evening while Phosphorus is not such that Jane believes it is visible in the

[5] Leibniz's Law can be formulated without second-order quantification using a schema, where for every predicate ϕ, the following is asserted: $x{=}y \rightarrow (\phi(x) \leftrightarrow \phi(y))$.

evening, even though Hesperus is identical to Phosphorus. Fregean accounts that validate this judgement may require an apparently ad hoc restriction of Leibniz's Law in cases of propositional attitude ascriptions. Frege's own strategy, which involves reference-shifting within such contexts, might avoid this issue, though it may itself be ad hoc. In contrast, on anti-contextualist anti-Fregeanism, believing that Hesperus is visible in the morning entails believing that Phosphorus is visible in the morning, precisely because Hesperus and Phosphous are the same thing. This issue also intersects with debates on *de re* vs *de dicto* propositional attitudes, which for the sake of brevity I will set aside in this introduction (e.g. Quine 1956, Kaplan 1968, Sosa 1970).

There are other ways in which characterizing meaning as supervenient on cognitive significance is problematic. As Kripke (1979) showed, differences in cognitive significance can arise even between terms that are, by hypothesis, synonymous, or even between different tokens of the same term. For example, "Londres" and "London" may differ in cognitive significance: one may rationally have conflicting attitudes towards "Londres is pretty" and "London is pretty". However, positing a difference in meaning between these terms would, arguably, make meaning excessively fine-grained. The same applies to synonyms like "furze" and "gorse". We can easily imagine someone who is certain that furze is furze but still rationally doubts that furze is gorse, or who attaches different descriptive or perceptual information to the two terms (Kripke 1979, Williamson 2021a). Yet, these terms are synonymous. Moreover, the same kind of case—where two synonyms differ in cognitive significance for a rational and linguistically competent subject—can be constructed for any pair of synonymous terms. Therefore, an account on which terms like "furze" and "gorse" differ in meaning would essentially entail that no two words are synonymous. Furthermore, even distinct tokens of the same term might differ in cognitive significance. Suppose Peter believes that Paderewski, a pianist, has musical talent, while also believing that Paderewski,

a politician, lacks musical talent. However, unbeknownst to Peter, both occurrences of "Paderewski" refer to the same person (Kripke 1979). Here, the Fregean view seems committed to attaching different meanings to the two occurrences of the term. This case also shows that modes of presentation cannot always be characterized as linguistic guises: they have to do with how something is represented, but not necessarily how it is *linguistically* represented.

§3 The Generality of Frege Puzzles

The previous section highlighted some of the implications of Frege puzzles for a range of philosophical debates, including those on meaning, belief, knowledge, and rationality. Moreover, it showed how Frege puzzles also raise methodological issues concerning the trade-off between preserving general principles—such as compositionality and Leibniz's Law—and accommodating our intuitive judgements in problematic cases.

Beyond their implications for these debates, Frege puzzles are general and pervasive in another important respect: they arise across a wide variety of phenomena. These include the ascription of a priori and a posteriori knowledge, epistemic necessity, knowledge of one's evidence, epistemic probability, and metalinguistic knowledge and belief (see Williamson 2021a). All these phenomena generate allegedly "opaque" contexts, namely linguistic contexts where substituting terms or phrases that refer to the same object appears to alter the truth-value of the relevant sentences (Schwitzgebel 2024). Allegedly opaque contexts are, in other words, those in which the correctness of our assertions about certain things seems to depend on how these things are represented. Consider the following pair of sentences:

5. We know a priori that Hesperus is Hesperus.

6. We know a priori that Hesperus is Phosphorus.

This constitutes just another instance of Frege's puzzle: we intuitively judge (5) as true and (6) as false. Yet, "Hesperus is Hesperus" and "Hesperus is Phosphorus" express the same *coarse-grained* proposition. Therefore, (5) and (6) cannot differ in truth-value unless we adopt a finer-grained semantic framework. Similar considerations apply to statements about epistemic necessity:

7. It is epistemically necessary that Hesperus is Hesperus.

8. It is epistemically necessary that Hesperus is Phosphorus.

Again, we tend to judge (7) as true and (8) as false, although the two sentences only differ in co-referential (and co-intensional) terms. Analogous issues arise for statements concerning one's access to one's own evidence:

9. Jane knows that her evidence includes that Hesperus is visible in the evening

10. Jane knows that her evidence includes that Phosphorus is visible in the evening

Frege puzzles also arise for metalinguistic knowledge. For example, we might judge the following two sentences—though they express the same coarse-grained proposition—as potentially differing in truth-value:

11. Jane knows that "Hesperus" refers to Hesperus.

12. Jane knows that "Hesperus" refers to Phosphors.

Explanation contexts are also sometimes said to be opaque: the way something is represented seems to affect whether the explanation is intelligible, appropriate, or informative. Terms like "because" are thus said to generate opaque contexts:

13. Hesperus is visible in the morning, as well as in the evening, because Hesperus is Phosphorus.
14. Hesperus is visible in the morning, as well as in the evening, because Hesperus is Hesperus.

It is easy to see how (13) may, in certain contexts and given the appropriate background knowledge (for example, that Phosphorus is visible in the morning), constitute a good explanation, while it is hard to see how (14) could be helpful in any context.

Significantly, Frege puzzles can arise even in the absence of language, since differences in modes of presentation need not be *linguistic* differences. This is illustrated not only by Kripke's Paderewski case, but also by cases involving, for instance, distinct occurrences of demonstrative terms which differ only perceptually— e.g. two uses of "that house" uttered while looking at apparently distinct buildings which are in fact connected and thus parts of the same house. In general, modes of presentation can reflect perceptual or imaginative ways of representing the relevant entities. They appear to be tied to representation itself, rather than to *linguistic* representation specifically. Indeed, certain contexts might require positing additional types of modes of presentation. For example, in debates concerning the relationship between propositional and practical knowledge, philosophers have introduced the notion of *practical* modes of presentation (Stanley & Williamson 2001, Glick 2015, Pavese 2020). Additionally, physicalist accounts of

consciousness often appeal to *phenomenal* modes of presentation—or phenomenal concepts—of physical properties (Loar 1997, Perry 2001, Papineau 2002, Levine 2007).

In this sense, while some of the examples discussed so far may appear quite specific, Frege puzzles in fact bring to light broad and fundamental issues concerning the relationship between representation, cognition, meaning, and facts. Given certain assumptions about the relation between semantics and metaphysics, the core issue may be understood as concerning the relation between the fine-grained structure of language—and of representation more generally—and the underlying reality it aims to represent. At their core, the theoretical and methodological challenges revolve around striking a balance: on the one hand, modelling reality in a way that accommodates how it appears to us; on the other, avoiding the projection onto reality of distinctions that may be merely cognitive or linguistic.

§4 Treating Philosophical Problems as Frege Puzzles

This dissertation argues that Frege puzzles and opacity play a central role in various classic philosophical debates. Viewing philosophical problems through the lens of Frege puzzles involves, first, drawing attention to the role of semantics and propositional attitude ascriptions in debates that appear to concern non-semantic issues—such as consciousness, disagreement in metaphysical disputes, and fundamentality. Second, this perspective highlights how some of the central intuitions driving these debates are those generated by Frege puzzles, and how proposed solutions to the relevant problems often reduce to familiar responses to Frege puzzles. Third, this approach offers a unifying framework that connects seemingly unrelated issues by revealing a shared underlying structure, without necessarily dissolving the problems themselves. Finally, tracing these issues back to underlying Frege

puzzles provides not only new insights into their possible resolution, but also a deeper understanding of Frege puzzles themselves and their scope. For example, extending Frege puzzles beyond purely linguistic contexts highlights the need for a general notion of modes of presentation encompassing phenomenal or practical modes of presentation.

Related projects have already been pursued. For instance, several philosophers have attempted to reduce problems concerning indexicality—especially so-called "*de se* puzzles" (Perry 1979)—to particular instances of Frege's puzzle (Cappelen & Dever 2013, Magidor 2015, see Torre & Weber 2019 for discussion). The problem of consciousness has also been treated, at least indirectly, as an instance of Frege's puzzle (e.g. Loar 1990/1997, Tye 2000). Chapter 2 examines this approach, assessing the analogies and putative differences between familiar Frege cases and the scenario presented in Frank Jackson's Knowledge Argument against physicalism (1986). If successful, these reductions contribute to "deflating" the problems in question, tracing them back to a familiar phenomenon. In doing so, they support an anti-exceptionalist stance: rather than treating these philosophical puzzles as sui generis or as requiring radically new theoretical tools, they are understood as manifestations of a common and well-studied semantic phenomenon.

Chapters 1 and 3 explore the role of Frege puzzles in other metaphysical debates—specifically, the debate over whether certain metaphysical disputes are merely verbal, and the debate over whether fundamentality is a feature of "worldly" items, such as facts or properties, or rather of our representations of them. Strikingly, in these debates, the very distinction between representation and reality often appears to be at stake. In the debate on consciousness, a central issue is whether there is a real distinction between the appearance of certain phenomenal states and the states themselves. For example, realists and anti-realists about consciousness disagree over whether the notion of an illusory feeling of pain is even coherent. Issues related to representation and reality also arise in the debate on metaphysical

explanation and fundamentality. The very idea of metaphysical explanation seems to impose explanatory structure onto reality, even though explanation is at least partially an epistemic notion. Finally, in debates about verbal disputes, the central question is precisely whether competing views in metaphysics truly diverge on what reality is like or merely diverge in their ways of (linguistically) representing it.

In all these cases, we can identify key terms that appear to generate putatively opaque contexts. Metaphysical disputes are said to be verbal in the sense that the parties involved are taken to *agree* on all factual issues and merely disagree about language. The Knowledge Argument claims that one can *know* all physical facts about colour perception yet still *learn* something new upon experiencing colour. In both cases, a central role is played by the apparent opacity introduced by the propositional attitude verbs "agree", "disagree", "know", and "learn". In the final chapter, I extend this approach to issues concerning metaphysical explanation. As noted earlier, explanation contexts are said to be opaque, due to allegedly opacity-inducing terms like "because". For example, the statement "The member of {Socrates} is wise *because* Socrates is wise" seems true, whereas "Socrates is wise because the member of {Socrates} is wise" seems false, even though "Socrates" and "the member of {Socrates}" are co-intensional. This apparent asymmetry requires intensionalists to provide an error theory for our intuitions about metaphysical explanation. I argue that a similar account can explain our judgements about perspicuity, or representational fundamentality—that is, our intuitions concerning objectively better or worse ways of representing reality.

§5 Clarifications

A few clarifications are in order. As the reader may notice, I tend to favour coarse-grained semantic frameworks and, accordingly, an anti-Fregean approach to Frege puzzles. However, this dissertation aims to remain neutral on questions concerning semantics and propositional attitude ascription, and each chapter considers and discusses at least some of the main alternative views.

Precisely because my goal has not been to defend a particular semantic theory, but rather to highlight how the central issues in each case *are* semantic—and hinge on which semantic framework one adopts—I have not examined in detail any specific account of the semantics of propositional attitude ascriptions, nor have I attempted to weigh their respective advantages and drawbacks. In particular, following Williamson (2021a), I have divided the space of possible views on Frege puzzles into two broad camps, Fregean and anti-Fregean, specifying that the orthogonal distinction between contextualist and anti-contextualist accounts of propositional attitude ascriptions further subdivides the anti-Fregean camp.

There are, of course, some relatively influential proposals that I do not discuss, such as "relationist" accounts based on the notion of coordination, typically associated with Fine (2007). On this view, coordination is a relation between representations: roughly, two representations are coordinated when they represent two objects as the same. The idea is that asserting the identity between two objects is representing them *to be* the same, whereas *presupposing* their identity is representing them *as* the same (Gray 2017).

Within an individual's attitudes at a time, coordination affects which inferences are licensed, and whether the subject counts as irrational. For instance, if Jane has coordinated beliefs that Hesperus is the morning star and Hesperus is the evening star, she is entitled to infer that something is both the morning star and the evening star. By contrast, if she has uncoordinated beliefs that Hesperus is the morning star and Phosphorus is the evening star, she is not entitled to that inference. Moreover, she is not irrational in believing that

Phosphorus is the evening star while also believing that Hesperus is *not* the evening star: in the absence of coordination, this does not amount to believing an explicit contradiction. On this account, coordination explains differences in cognitive significance between representations with the same referential content. In Fine's view, coordination, or lack thereof, affects the semantic content of the relevant statements, but non-semantic accounts of coordination are possible (e.g. see Bonardi 2020, Gray 2017). This view can thus fall within the Fregean or anti-Fregean camp, depending on whether coordination plays a role in determining semantic content.

For similar reasons, I will not discuss accounts of cognitive significance in terms of mental files (Recanati 2012). On Recanati's view, a mental file is a repository for Mentalese predicates that a subject (correctly or incorrectly) takes to be true of a certain object. Two files respectively concerning two objects *o* and *o*' are said to be *linked* if their bearer judges (again, correctly or incorrectly) that *o* is identical to *o*' (Recanati 2012). There is no consensus on the identity conditions for mental files, but the issue ultimately mirrors that of the identity conditions for modes of presentation.

Both the relationist and the mental files approaches provide accounts of differences in cognitive significance, and aim to explain how such differences affect attitudes, inference, and rationality. However, what ultimately differentiates Fregean from anti-Fregean views is whether these cognitive differences are taken to have a semantic import. Insofar as both relationist and mental files frameworks can be formulated in Fregean or anti-Fregean terms—depending on whether semantic significance is attributed to the relevant notions—they can be unproblematically situated within one of the two broad camps.

§6 Overview of the Three Chapters

Chapter 1—Verbal Disputes and Frege Puzzles

Some disputes arising in various domains of inquiry have been labelled as merely verbal. These include disputes as to whether whales are fish, free will is compatible with determinism, the Pope is a bachelor, and so forth. The idea is that if the disputants appear to be equally informed about relevant details of the cases at issue but still find themselves stuck in seemingly intractable disagreement, then they must be talking past each other in a way that involves some sort of linguistic misunderstanding. For example, if the parties agree that whales are warm-blooded aquatic vertebrates with lungs but still dispute whether whales are fish, they must be using the term "fish" differently. In general, in these cases one party asserts the disputed sentence "o is F" and the other asserts its negation ("o is not F"), but both agree that o is G, H, J, etc. Most accounts of verbal disputes fall into two main categories. On the one hand, on "semantic consistency" accounts (e.g. Sider 2006), the speakers' apparently conflicting utterances "o is F" and "o is not F", as well as the corresponding beliefs, are in fact consistent, and the disputants fail to disagree on whether o is F. On the other hand, "metalinguistic" accounts (e.g. Chalmers 2011a, Plunkett & Sundell 2021) take linguistic divergence and agreement on the relevant facts to warrant the ascription of conflicting (normative or descriptive) metalinguistic beliefs to the speakers—for example, about whether "F" means G.

In this chapter, I argue that paradigmatic cases of allegedly verbal disputes can be interpreted as involving neither semantic consistency nor primarily metalinguistic disagreement. The parties can instead be taken to disagree about whether o is F in virtue of disagreeing about whether to be F is to be G. So interpreted, these disputes are not verbal, as they arise from disagreement about factual matters—typically, higher-order identities. For example, the parties might engage in a dispute over "Whales are fish"—while agreeing that

whales are warm-blooded aquatic vertebrates with lungs—in virtue of disagreeing about the *non-metalinguistic* question of whether to be a fish is to be an aquatic vertebrate. These factual identity disputes come out as verbal in the semantic consistency sense only under specific semantic assumptions, and in the metalinguistic sense only given some form of deflationism about identity claims.

Although these factual identity disputes are not necessarily verbal, it might be objected that they involve agreement on all the relevant non-metalinguistic information, and are thus not fully substantive either: if to be F is to be G, and the disputants agree that o is G, aren't they, in a sense, in complete agreement? These questions, however, are analogous to those arising in paradigmatic Frege puzzles: assuming that to be F is to be G, do "o is F" and "o is G" express the same proposition? If so, does believing that o is G entail believing that o is F? The answers to these questions, I argue, depend on the semantics of propositional attitude ascriptions. Indeed, only given an anti-contextualist anti-Fregean framework does agreement that o is G entail agreement that o is F. However, this does not mean the disputes in question are defective: under anti-contextualist anti-Fregean assumptions, many perfectly legitimate disputes involve agreement on all the relevant propositions. Moreover, the metalinguistic account is subject to the same objection, as the problem of complete agreement arises with metalinguistic propositions as well. Thus, the disputes in question are neither verbal nor defective.

Chapter 2—The Knowledge Argument as a Frege Puzzle

Imagine Mary, who knows every physical fact about colour perception but has spent her life in a black-and-white environment. According to Jackson's (1986) Knowledge Argument (KA), Mary gains knowledge of a new non-physical fact when she leaves the room and sees red, meaning that physicalism is false. Against Jackson, physicalists deny that Mary gains

knowledge of a new fact, but concede that KA poses an epistemic problem for physicalism, as it requires it to account for the intuition that there is something that Mary learns when she sees red (Tye 1995, Loar 1990, Perry 2001).

In this chapter I argue that, within a reductive physicalist framework, the intuition that Mary learns something new is not due to anything unique to phenomenal consciousness, but rather to the apparent opacity of propositional attitude ascriptions in play in familiar Frege puzzles. Whether Mary can be said to gain new knowledge entirely depends on the semantics of propositional attitude ascription.

I begin by distinguishing three steps in Mary's putative epistemic progress: t1, when Mary is in her room, t2, when Mary sees red, and t3, when she is in a position to describe her experience *as* an experience of red. I then argue that the intuition that Mary gains new knowledge at t3, like the intuition that Jane gains new knowledge when she comes to know that Hesperus is Phosphorus under the informative guise "Hesperus is Phosphorus", is due to the (alleged) opacity of knowledge ascriptions.

Yet, Mary's real progress arguably takes place at t2, when she comes to know *what it's like* to see red. After examining several accounts of knowing what it's like—including Lewis' ability hypothesis (1990) and acquaintance accounts (Conee 1994)—I argue that the semantics of "know" followed by wh-clauses supports the view that knowing what it's like involves propositional knowledge. Once again, whether Mary gains new knowledge at t2 depends on which semantic framework and account of propositional attitude ascriptions are correct. In other words, Mary's knowledge of what it's like to see red introduces a new Frege puzzle.

I then address the objection that, in paradigmatic Frege puzzles, one comes to know a further contingent proposition when grasping the relevant identity claim under an informative guise. For instance, only when one comes to know the informative claim

"Hesperus is Phosphorus" does one learn that there is something which is both the brightest planet in the evening and the brightest planet in the morning. According to the objection, it is unclear whether anything analogous applies to KA. In response, I argue that in both cases, whether new knowledge of contingent propositions is gained depends on which semantic framework is correct and whether certain principles of epistemic closure hold.

Finally, I discuss a challenge to physicalist accounts of KA as a Frege puzzle, namely that either phenomenal modes of presentation introduce new *non-physical* properties or Mary's new phenomenal knowledge seems insufficiently "substantial" (Levine 2007, Block 2007, Tye 2008, Schroer 2010). I argue that it is unclear whether this argument can be formulated without begging the question against physicalism and that, in any case, under a sufficiently broad understanding of phenomenal modes of presentation, a posteriori physicalism can respond to the challenge.

Chapter 3—On Representational Grounding

Contemporary grounding theorists adopt two different approaches to the question of metaphysical explanation, or grounding. On the "worldly" generative approach, grounding connects *distinct* facts. On the "representational" reductive approach, grounding connects distinct truths representing the same fact: perspicuous truths representationally ground less perspicuous truths, but only one fact is involved in each case (Rubenstein 2024, Jones 2022)

In this chapter, I focus on the reductive approach, arguing that representational grounding needs to satisfy two constraints: Independence and Structure. First, representational and worldly approaches are mutually inconsistent in specific cases, and correspond to competing views in metaphysics. Thus, I argue, the relation of representational grounding needs to be defined *independently* of worldly grounding (Independence). Second, representational grounding needs to have certain structural features: despite being sometimes

labelled as a "reductive" relation, it gives rise to a hierarchy of *distinct* truths, which requires it to be asymmetric and irreflexive (Structure).

After considering various candidate relations of perspicuity, I contend that none satisfies both requirements. On the one hand, it seems difficult to account for the difference in perspicuity between two truths without positing a difference in *worldly* fundamentality between the referents of their constituents. But this just reintroduces the worldly hierarchy that representational grounding was meant to avoid, failing to satisfy Independence. On the other hand, a familiar relation of entailment does satisfy Independence, but does not satisfy Structure: representational grounding is irreflexive and asymmetric, while entailment is not. Defining representational grounding in terms of a restricted notion of "one-way" entailment, I argue, does not work either.

I conclude by suggesting that the perceived asymmetry in representational fundamentality between the truths in question may not reflect any objective difference between them. Just as, in cases of opacity generated by terms like "because", anti-Fregeans offer error theories that appeal to pragmatic features of explanation (e.g. Williamson 2024), I argue that our judgements about the apparent asymmetry in perspicuity between the truths in question can be similarly explained away in terms of pragmatic and cognitive factors.

CHAPTER 1

Verbal Disputes and Frege Puzzles

The view that some apparently substantive disputes are, in fact, merely verbal is quite popular among metaphysicians. If the parties appear to be equally well-informed about relevant details of the cases at issue, but remain stuck in apparently intractable disagreement, their dispute is assumed to arise from an implicit linguistic misunderstanding. For instance, if we agree that whales are warm-blooded aquatic vertebrates but still dispute whether whales are fish, we must be *meaning* different things by "fish". On the Semantic Consistency account of verbalness, the disputants literally express different things by the relevant term and therefore fail to genuinely disagree. On the Metalinguistic Disagreement account, the disputants merely disagree about the meaning of the term in question. I argue that paradigmatic cases of allegedly verbal disputes might instead arise from disagreement over non-metalinguistic identity claims (e.g. "To be a fish is to be a cold-blooded creature with gills"). These factual (non-metalinguistic) identity disputes (FIDs) qualify as verbal only under specific controversial assumptions. It may be objected that, though not verbal, FIDs are nonetheless "non-substantive", since they involve agreement on all the relevant propositions. In replying to this objection, I argue that FIDs are structurally similar to Frege puzzles: the question of whether the disputants fully agree depends on the semantics of propositional attitude ascriptions. The upshot is that paradigmatic examples of allegedly verbal disputes can instead be characterized as factual identity disputes, which are neither verbal nor otherwise defective.

§ Introduction

The view that certain apparently "substantive", or "heavyweight", metaphysical disputes are, in fact, merely verbal is popular among contemporary metaphysicians (Hirsch 2005, 2009; Sidelle 2007; Chalmers 2011a; Jenkins 2014). The idea is that if the disputants appear to be equally well-informed about relevant details of the cases at issue but still find themselves stuck in seemingly intractable disagreement, then they must be talking past each other in a way that involves some sort of linguistic misunderstanding.

The idea that disputes with these features are merely verbal is not confined to metaphysics: debates arising in a wide range of domains of inquiry—such as taxonomy and astronomy, as well as everyday disputes—have been labeled as merely verbal. These include disputes over whether whales are fish, Pluto is a planet, the Pope is a bachelor, tomatoes are fruit, and so forth. For example, before Linnaeus, the term "fish" was used to refer to any aquatic vertebrate, whereas in Linnaeus' taxonomy, the category of fish is mutually exclusive with the category of mammal: fish have cold blood and gills, while mammals have warm blood and lungs. So, in Linnaeus' classification, aquatic vertebrates with warm blood and lungs, such as whales and dolphins, are not fish, but mammals. Suppose a fisherman and a biologist agree that

(1) Whales are aquatic vertebrates

(1b) Whales are warm-blooded and have lungs

but still dispute whether

(2) Whales are fish

The idea is that the disputants must *mean* different things by "fish": one must be using "fish" in Linnaeus' sense, the other in the folk sense. Likewise, if two disputants agree that

(3) The Pope is an unmarried man

(3b) The Pope is ineligible for marriage

but still dispute whether

(4) The Pope is a bachelor

they must *mean* different things by "bachelor"—plausibly, one means unmarried man, whereas the other means unmarried man *eligible for marriage*. Another example is the discussion over whether Sedna is a planet. Sedna counts as a planet by some taxonomies but not others. If two astronomers who know everything about Sedna's size, shape, structure, and location still disagree about whether it is a planet, then they are likely to mean different things by the term "planet".

This idea is spelled out differently in two of the main accounts of verbalness. On what I will call "Semantic Consistency" accounts, the speakers mean different things by the relevant terms in the sense that they literally *express* different things by using such terms. Therefore, on this view, the disputants' apparently conflicting assertions—e.g. "Whales are fish" and "Whales are not fish"—are in fact consistent. On the other hand, "Metalinguistic Disagreement" accounts hold that, regardless of what is literally expressed by the terms in question, the speakers mean different things by such terms in the sense that they *disagree* about their meaning. On this view, then, the source of the dispute is an underlying metalinguistic disagreement. After introducing these two clusters of views, I will also

discuss the proposal that verbalness should be accounted for in pragmatic terms (e.g. in terms of speaker meaning), arguing that it, in fact, reduces to the two broader frameworks above.

Importantly, paradigmatic examples of putatively verbal disputes are cases in which the parties agree on a relevant set of claims. In the cases above, for instance, the disputants agree on (1), (1b), and (3), (3b)—they agree that whales are warm-blooded aquatic vertebrates with lungs, and that the Pope is an unmarried man ineligible for marriage. Such "uncontroversial" claims are normally specified on a case-by-case basis, but it is typically not required that the disputants agree on all the details about the case at issue in order for their dispute to count as verbal—for example, whether or not the disputants disagree about the Pope's age, or nationality, will plausibly be irrelevant in determining whether their dispute about whether the Pope is a bachelor is verbal.

The plan for the chapter is as follows: after introducing the two main frameworks for verbalness, Semantic Consistency and Metalinguistic Disagreement (§1), I argue that paradigmatic examples of allegedly verbal disputes can instead be taken to arise from genuine disagreement about factual (i.e. non-metalinguistic) identity claims (§2). For example, the dispute as to whether whales are fish can be taken to arise from *non-metalinguistic* disagreement about whether to be a fish is to be an aquatic vertebrate, and the dispute as to whether the Pope is a bachelor can be taken to arise from disagreement as to whether to be a bachelor is to be an unmarried man eligible for marriage. These factual identity disputes (henceforth, FIDs) qualify as verbal only under specific (and controversial) assumptions. In §3, I argue that they only count as verbal$_{MD}$ (i.e. verbal in the metalinguistic disagreement sense) under deflationist assumptions whereby identity claims are themselves interpreted as metalinguistic. In §4, I argue that FIDs only come out as verbal$_{SC}$ (i.e. verbal in the semantic consistency sense) given specific semantic assumptions. Finally, in §5, I consider the objection that, though not verbal, FIDs are still somehow defective, as the

parties agree on all the relevant non-metalinguistic information. In replying to this objection, I argue that FIDs are structurally similar to Frege's puzzle (1892), and that the issue of the disputants' full agreement depends on the semantics of propositional attitude ascriptions. The upshot is that paradigmatic examples of allegedly verbal disputes can instead be characterized as factual identity disputes (FIDs), which are neither verbal nor otherwise defective.

A brief note on the distinction between *disputes* and *disagreements,* as I use these terms. Disputes have to do with the disputants' *speech acts*: simplifying a bit, if A utters a sentence S and B replies by uttering not-S, then A and B are engaging in a dispute. Disagreement, on the other hand, involves the disputants' *beliefs* and their contents: A and B disagree iff A believes p and B believes q, where q and p are inconsistent propositions. From these definitions, it follows that two speakers may disagree without engaging in a dispute—they may simply hold inconsistent beliefs without performing any speech acts—and, at least in some cases, engage in a dispute without disagreeing, for example when context-dependent terms are involved (more on this in later sections).

Finally, it is worth noting that I will couch the discussion in terms of words and their meanings, but most of the relevant points about metalinguistic disagreement and metalinguistic beliefs could equally be reformulated in terms of metaconceptual disagreement and metaconceptual beliefs.

§1 Verbal Disputes

Most accounts of verbal disputes, as noted, can be broadly divided into two main categories. The primary distinction between them concerns how the parties' "meaning different things" by some relevant terms is characterized.

§1.1 Semantic Consistency

On what I will refer to as "Semantic Consistency" accounts (see Hirsch 2005; Sider 2006, 2009; Sidelle 2007; Bennett 2009; Chalmers 2011a; Jenkins 2014 for discussions of this view), the disputants are taken to literally express different meanings when using the relevant terms. As Sider explains, this entails that the disputants' seemingly conflicting utterances are, in fact, consistent:

> To say that an apparent dispute over sentence ϕ is merely verbal is to say that the disputants do not mean the same thing by the sentence ϕ, and that what one says by uttering ϕ is consistent with what the other says by uttering ¬ϕ.
>
> (Sider 2006, p.76)

For example, in a dispute over (2), the idea is that, *given their agreement on (1) and (1b),* the disputants are most charitably interpreted as meaning different things by the term "fish", so that the disputed sentence (2) expresses different propositions as used by each party. Thus, when the fisherman asserts (2) and the biologist denies it, they are actually expressing consistent propositions p and not-q, rather than p and not-p. The same idea is often extended to mental content: although the disputants appear to disagree, the content of their beliefs is likewise constituted by the mutually consistent propositions p and not-q. Accordingly, the parties fail to genuinely disagree. We may thus define verbalness in terms of semantic consistency as follows:

> Semantic Consistency—A dispute over a sentence S is verbal iff when one party asserts S and the other asserts not-S, they are in fact asserting two *mutually consistent* propositions p and not-q. Likewise, their apparently conflicting beliefs are in fact consistent: one party believes p and the other believes not-q—hence they fail to disagree.

One charitable interpretation of the dispute about whales runs as follows: as used by the fisherman, "fish" means aquatic vertebrate, so by saying "Whales are fish", the fisherman expresses the proposition p that whales are aquatic vertebrates. In contrast, as used by the biologist, "fish" means cold-blooded creature with gills, so by denying "Whales are fish", the biologist expresses the proposition not-q that whales are not cold-blooded creatures with gills. In line with principles of charity which aim to minimize attributions of error, p and not-q are not only consistent but also both true. In other words, each disputant expresses a truth "in their own idiolect".

It is worth noting that, strictly speaking, the Semantic Consistency account does not require both parties to express a truth for a dispute to be classified as verbal. What the account demands is just that their relevant utterances and beliefs are consistent. Nevertheless, the idea that both parties are right in their respective idiolects often plays a central role in accounts of *metaphysical* disputes as verbal, given that the participants in such disputes are typically equally well-informed about the relevant details of the cases under discussion. I will use the label "verbal$_{SC}$" to refer to any dispute that qualifies as verbal according to the Semantic Consistency account.

§1.2 Metalinguistic Disagreement

It is now widely acknowledged in the literature that not every paradigmatic case of verbal dispute qualifies as verbalSC. Several theorists (e.g. Chalmers 2011a, Balcerak Jackson 2014) have argued that the mere verbalness of a dispute cannot always be traced to a difference in the literal meaning of the disputed sentence for the two parties. In other words, verbal disputes do not always involve cases where the parties' assertions are consistent. As will be discussed further in §4, the Semantic Consistency account presupposes that a divergence in the parties' individual uses of a term determines a difference in the meaning the term expresses in their respective utterances, against widely influential externalist considerations due to Burge (1979) and Putnam (1975).[6]

An example from Balcerak Jackson (2014) illustrates this point. Consider the English term "billion", which denotes the number 10^9. Native speakers of German and Italian sometimes mistakenly assume that "billion" translates the German "Billion" or the Italian "bilione", both of which denote the larger number 10^{12}. Imagine the following exchange in English between Fozzy, a native English speaker, and Guido, a bilingual native German speaker:

Fozzy: There are currently more than seven billion people living on Earth.
Guido: No way! There are far fewer than seven billion people living on Earth right now.

Arguably, this exchange does not satisfy Semantic Consistency, as both Fozzy and Guido are speaking standard English and both intend to use "billion" with the meaning the term carries within the broader English-speaking community. Accordingly, Fozzy asserts the

[6] To avoid this objection, Hirsch (2005, 2009) treats the parties to verbal disputes as if they were literally speaking different languages—that is, as if they belonged to distinct linguistic communities. While this may avoid Burge-style objections, Hirsch's view still conflicts with other varieties of externalism, such as Sider's (2011) reference magnetism (more about this in §4).

proposition that there are more than $7x10^9$ people on Earth, while Guido asserts the proposition that there are fewer than $7x10^9$. Yet, many theorists would still characterize this dispute as merely verbal, as it seems to arise from a mere linguistic misunderstanding, and crucially involves a metalinguistic disagreement between the speakers, who hold mutually inconsistent beliefs about the meaning of the term "billion" in English.[7]

It is precisely to accommodate this kind of case that some theorists have proposed what I refer to as the "Metalinguistic Disagreement" account of verbalness (Chalmers 2011a; Plunkett & Sundell 2021). On this view, the biologist and the fisherman mean different things by "fish" in the sense that they *disagree* about the meaning of the term, *regardless of what the term actually expresses as used by each of them.* In other words, this account aims to remain neutral on the underlying semantics—and thus on whether the speakers express consistent propositions—and instead locates the source of the dispute's verbalness in the metalinguistic disagreement (normative or descriptive) between the speakers. Chalmers (2011a) defines verbalness in these terms:

> Metalinguistic Disagreement: A dispute over a sentence S is verbal iff the speakers disagree about the meaning of a term T in S, and the dispute arises in virtue of this disagreement regarding T.

Suppose the fisherman believes that "fish" means aquatic vertebrate, and the biologist believes that "fish" means cold-blooded creature with gills. The same applies to the proposition expressed by the disputed sentence "Whales are fish": the fisherman believes

[7] In fact, Balcerak Jackson also claims that the speakers do not genuinely disagree about the number of people on Earth. However, this claim also presupposes anti-externalist commitments concerning the individuation of mental content.

that the sentence expresses the (true) proposition that whales are aquatic vertebrates, while the biologist believes it expresses the (false) proposition that whales are cold-blooded creatures with gills. This offers a plausible explanation of the otherwise puzzling fact that the two disputants agree on most propositions about what whales are like—e.g. (1) and (1b)—but still dispute whether whales are fish. The metalinguistic explanation of this impasse is that one party believes the disputed sentence (2) expresses a true proposition p, while the other believes that (2) expresses a false proposition q. Hence, despite their agreement on (1) and (1b), one party asserts (2), and the other denies it.

Of course, the Semantic Consistency account and the Metalinguistic Disagreement account are not mutually exclusive. A dispute may be classified as verbal in both senses if the disputants both express different things by the relevant term T and disagree about its meaning. I will use the label "verbal$_{MD}$" to refer to any dispute that qualifies as verbal by the Metalinguistic Disagreement account.

Chalmers (2011a) remains largely neutral on what dimensions of meaning matter for verbalness$_{MD}$, aside from his claim that verbal$_{MD}$ disputes arise from disagreement over *non-merely-extensional* dimensions of meaning. His reasoning is that whenever two speakers disagree *substantively* about whether o is F, they will also disagree about the extension of the term "F". For example, he argues that if two speakers disagree on whether Joe is a murderer, they will likewise disagree about the extension of "murderer", but this alone does not render the dispute verbal. However, the problem with this example is not that the speakers merely disagree about extension, but that the "in virtue of" condition is not met: according to the metalinguistic account, verbal disputes arise *in virtue of* metalinguistic disagreement. In other words, when disagreement over a term's extension occurs *independently* of any substantive disagreement—unlike in Chalmers' example—it is unclear why the dispute would not count as verbal by the metalinguistic account. Indeed, other

theorists have provided examples of what they take to be verbal disputes arising in virtue of disagreement about a term's extension. Sidelle (2007) and Balcerak Jackson (2014) discuss the following example: Rolf and Scooter have just watched a movie starring Burt Lancaster and Kirk Douglas. Rolf, unfamiliar with old Hollywood stars, has Burt Lancaster's image in mind and is thinking about Lancaster's behaviour in the film when he says "Kirk Douglas was really menacing, wasn't he?". Scooter replies "No, Douglas wasn't menacing, but Lancaster was!". As in the billion case, there seems to be a sense in which the dispute here stems from a linguistic misunderstanding. Yet, the standard theory of proper names maintains that proper names have no layer of meaning beyond reference (and intension). Thus, Balcerak Jackson and Sidelle argue, the dispute between Rolf and Scooter arises from disagreement about the *extension* of the relevant term.

One desideratum for a theory of verbalness is that it should provide a plausible account of how disputes typically regarded as verbal—such as metaphysical disputes about composite objects, or disputes about whether a certain celestial body is a planet—often persist longer than a mere linguistic misunderstanding would. To explain this fact, Plunkett & Sundell (2021, 2023) have suggested that such disputes may be cases of "metalinguistic negotiation", in which speakers, by using the relevant terms in a certain way, implicitly—i.e. without making overtly metalinguistic claims—advance normative claims about how those terms *should* be used.

To consider a standard example, two speakers who agree that there are simples arranged table-wise but still disagree about whether tables exist, might just be implicitly disagreeing about how the term "exist" *should* be used. This is meant to explain why many paradigmatic cases of supposedly verbal disputes do not simply dissolve once the metalinguistic divergence is recognized. In general, in cases of metalinguistic negotiation, the dispute is unlikely to be resolved by making the underlying metalinguistic disagreement

explicit, precisely because the disagreement is *normative* in nature—it concerns how linguistic expressions *ought* to be used (Plunkett & Sundell 2021). For example, suppose that two speakers A and B are debating whether waterboarding constitutes torture. A may be aware that waterboarding is not classified as torture under American law, yet still assert "Waterboarding is torture" to implicitly advocate a revised notion of torture under which waterboarding would be included. Because it characterizes the relevant disputes as arising in virtue of (normative) metalinguistic disagreement, I will treat this account as a version of the metalinguistic account.

There is, however, a further challenge affecting Metalinguistic Disagreement accounts: in many cases, the disputants themselves would resist the idea that their claims and disagreements concern language, rather than the relevant object-level issues (see Plunkett & Sundell 2021, Abreu 2023). Admittedly, in *some* cases, the speakers may be content to conclude the dispute by saying "I guess we just mean different things by 'T'", where "T" is the term that appears to generate the dispute. However, this will not always be the case, especially in more theoretical cases such as metaphysical disputes. In this sense, it is arguably a shortcoming of the Metalinguistic Disagreement account that it not only reinterprets seemingly factual disputes as in fact metalinguistic, but also portrays the disputants themselves as, unbeknownst to them, merely disagreeing about linguistic matters.

§1.3 The Pragmatic Account

Some have argued, against Semantic Consistency and Metalinguistic Disagreement, that verbalness is in fact a pragmatic phenomenon. Balcerak Jackson (2014), for example, contends that verbal disputes are characterized by a kind of pragmatic defect, namely that there is no single question both disputants are trying to answer. He illustrates this point with the following example. There is a tendency among speakers of American English to use the

term “metaphysics” to refer to the study of supernatural phenomena, such as past lives and out-of-body experiences. This, of course, is not the way in which professional philosophers use the term. According to Balcerak Jackson, this divergence arguably reflects a genuine semantic ambiguity in the term “metaphysics” in American English. Suppose that Kermit and Gonzo are approached by a third person who asks “Is there anywhere I can buy books on metaphysics?”, and the following exchange ensues:

Gonzo: “The bookstore downtown sells books on metaphysics”
Kermit: “No, that bookstore doesn’t have any books on metaphysics”

Kermit and Gonzo might share the same beliefs about the meaning of “metaphysics”—for instance, both might give it the philosophers’ sense—and yet they could differ in their assumptions about what the third party means by the term. Gonzo might assume she is using it in the spiritualist sense, while Kermit might assume she has the philosophers’ use in mind. According to Balcerak Jackson, this exchange constitutes a verbal dispute, as the parties assert contradictory sentences merely because of their conflicting assumptions about what the questioner is asking. He maintains that the dispute is not verbal in the Semantic Consistency sense, nor in the Metalinguistic sense. Rather, it is verbal because we cannot identify a mutually agreed-upon question that both parties attempt to address.

However, I believe there are several problems with this example. First, we assumed that both Kermit and Gonzo understand “metaphysics” in the philosophers’ sense, and that both agree the bookstore sells spiritualist books but not philosophy books. As noted, Gonzo understands the question of whether there are any metaphysics books at the bookstore as asking whether there are any spiritualist books there, and responds affirmatively. Yet, it is not clear that Gonzo is genuinely assenting to the sentence “The bookstore sells books on

metaphysics" when he answers the question. In a sense, he is merely pretending that this sentence is true, since he takes "metaphysics" to refer to a branch of philosophy, and knows that the bookstore does not carry philosophy books. If a dispute requires *sincere* assent to the sentence uttered, it is not even clear that Kermit and Gonzo are engaging in a genuine dispute.

Second, it is unclear why Kermit and Gonzo should not be taken to disagree about what "metaphysics" means—at least in the *token* occurrence of the term in the question they are answering—given that the term (as a type) is, we have supposed, genuinely ambiguous. If they do disagree about what the relevant *token* occurrence of "metaphysics" means, then the dispute could arguably be verbal in the metalinguistic sense. Furthermore, since Balcerak Jackson himself assumes that the term is ambiguous in American English, it is unclear why Gonzo's and Kermit's responses should not be taken to express different propositions. Suppose I say that there is money in the bank (meaning the financial bank) and you say that there is no money in the bank (meaning the river bank). Since "bank" is genuinely ambiguous between the river bank and the financial bank, this kind of case may in fact be one of the few that do satisfy Semantic Consistency even for the externalist. Thus, a major issue with Balcerak Jackson's example is that it is unclear why it could not be accounted for in terms of verbalness$_{SC}$ or verbalness$_{MD}$.

A third problem is that many paradigmatic examples of allegedly verbal disputes do not involve polysemous or ambiguous terms at all. Consider the dispute about whales. The idea would be that this dispute is verbal because we cannot identify a mutually agreed-upon question that both parties attempt to answer: one party is attempting to answer the question of whether whales are aquatic vertebrates, while the other is addressing whether whales are cold-blooded aquatic vertebrates with gills. But it is not obvious why we should interpret the dispute this way. In fact, it seems quite clear that we *can* identify a question that both parties

attempt to address, namely whether whales are fish. The idea that different uses of "fish" cause the meaning of the question "Are whales fish?" to shift for each speaker seems to rely, once again, on the anti-externalist assumptions that underpin the Semantic Consistency account, whereby individual differences in usage result in semantic differences. In other words, claiming that "Are whales fish?" expresses different questions for each party is just like saying that "Whales are fish" expresses different propositions for them—which is exactly what the Semantic Consistency account claims. Although Balcerak Jackson rejects the Semantic Consistency account, it is difficult to make sense of his claim that there is no single question both parties are attempting to answer, except by presupposing that the term "fish" means different things in each speaker's idiolect, so that the question effectively translates into different questions for each of them.

As we will see in §4, there are indeed cases—involving, for instance, indexical terms used in different contexts—where it may be genuinely difficult to identify a single question that both parties are trying to address. But in most other cases, the surface subject matter of the dispute, individuated disquotationally, arguably constitutes a legitimate question in its own right.[8] In short, Balcerak Jackson's example seems to fall within one or both of the two main accounts of verbalness outlined earlier. Moreover, we are not given adequate reasons to believe that the disputes typically characterized as verbal satisfy his account.

Another version of the pragmatic account, proposed by Vermeulen (2018), relies on the notion of speaker meaning. According to her account:

[8] Although this does not seem to be what Balcerak Jackson has in mind, the idea could be that there is actually no fact of the matter as to whether whales are fish—the only questions that do have an answer are questions "in the vicinity", such as whether whales are aquatic vertebrates, or whether whales are cold-blooded creatures with gills. See §3.3 for discussion of this view.

> A dispute over a statement S—where one party utters S and the other not-S—is verbal when (1) the parties use the same utterance-type S with different speaker's meaning such that what A means by uttering S (p) does not conflict with what B means by uttering not-S (not q), but (2) each ascribes the negation of their own speaker's meaning to the other (not p and q respectively).
>
> Vermeulen (2018, p. 342)

Vermeulen defines the speaker meaning of an utterance in terms of what the speaker *intends* to convey with the utterance, as opposed to what the utterance actually means or expresses. For example, Vermeulen asks us to consider a speaker A who utters "There's a hippopotamus in the fridge", intending to convey that there is an orange in the fridge. The speaker meaning of this utterance, Vermeulen claims, is that there is an orange in the fridge. In response, a second speaker B says "Of course there is no hippopotamus in the fridge", intending to convey precisely that there is no hippopotamus in the fridge. According to Vermeulen, this constitutes a verbal dispute, because what A intends to communicate does not conflict with what B intends to communicate. Vermeulen maintains that her account is superior to Semantic Consistency and Metalinguistic Disagreement, because it involves fewer theoretical commitments. For example, she claims that her account does not rely on the semantic assumption that meaning is determined by individual usage.

However, I believe that semantic problems similar to those affecting the Semantic Consistency account also arise for Vermeulen's account. Following Putnam and Burge, externalists typically hold that not only the content of utterances, but also mental content is partially determined by external factors. This applies to the content of *any* propositional attitude, including the attitude of *meaning* something by an utterance in the sense of *intending* to convey a particular content. Speaker A from the example above, for instance, is disposed to assent to sentences such as "There is a hippopotamus in the fridge" and to affirm

internally “There is a hippopotamus in the fridge”. Even when asked what she intends to convey with her utterance, she may say “I intend to convey that there is a hippopotamus in the fridge”. Furthermore, speakers typically have a standing intention to use words with the meaning these have within the broader linguistic community, and to convey what their utterances actually express in the relevant language. Thus, an externalist would maintain that, by ordinary standards for attitude ascriptions, A believes that there is a hippopotamus in the fridge, and that this is what she intends to convey.

Of course, A will likely have some sort of mental representation of an orange in the fridge when she utters “There’s a hippopotamus in the fridge”. But the idea that these sorts of individualistic mental representations determine the content of utterances and mental states is essentially the view that meanings are “in the head”—again, contrary to widespread externalist positions. One could certainly adopt this view, but the point is that Vermeulen’s account does not, as she claims, involve a more “lightweight” theoretical apparatus than competing accounts of verbalness. In fact, it seems to be subject to the same kinds of problems that affect the Semantic Consistency account. These considerations closely parallel those raised in response to Balcerak Jackson above: the claim that a speakers’ individual usage or mental representations determine which question they are trying to answer requires specific semantic assumptions.

Returning to the main discussion, it is worth noting that the disputes discussed so far can be contrasted with the corresponding “substantive” cases, in which the dispute arises from disagreement about some of the relevant propositions that are uncontroversial in allegedly verbal cases. For instance, if the parties to the whales dispute were to disagree about (1) or (1b), and on that basis engaged in a dispute over (2), their dispute would no longer count as verbal. Intuitively, the most plausible interpretation of such a case is that both disputants mean the same thing by “fish” and simply disagree about whether whales

meet certain *shared* criteria for "fish" to be correctly applied to them. By contrast, in the verbal version of the dispute, the parties agree on what whales are like, but still debate whether whales are fish because they disagree on what "fish" means—that is, they disagree about the criteria that something must meet for "fish" to be correctly applied to it.

In the following sections, I will argue that paradigmatic examples of allegedly verbal disputes can instead be understood as arising from disagreement about factual (i.e. non-metalinguistic) identity claims. The dispute over whether whales are fish, for example, can be taken to stem from *non-metalinguistic* disagreement about whether to be a fish is to be an aquatic vertebrate. I will refer to disputes that ultimately arise from disagreement about such identity claims as Factual Identity Disputes (FIDs).

§2 Factual Identity Disputes

§2.1 Why the Distinction Matters

At first glance, the difference between disagreeing about a factual (higher-order) identity claim—such as "To be a fish is to be an aquatic vertebrate"—and a metalinguistic claim—such as "'Fish' means aquatic vertebrate"—might appear irrelevant. However, we should first note that these two issues come apart. For one thing, factual disagreement about what fish are is not tied to a specific language: two speakers may disagree about what fish are even if they have never encountered the word "fish". By contrast, metalinguistic disagreement about the meaning of "fish" concerns the *English* word "fish".

More importantly for our purposes, distinguishing between these two interpretations can, in fact, have significant implications for both the status of the dispute in question and its mode of resolution. Indeed, if the dispute over whales stems from disagreement about the

meaning of the term "fish", the natural way of resolving the dispute is by consulting a dictionary, or looking at how the term is used among English speakers—or, at most, by deciding how the term ought to be used. This, as we will see, may itself depend on non-metalinguistic considerations about the property of being a fish. By contrast, if the whales dispute is taken to arise from disagreement about what fish *are*, then resolving it arguably has nothing to do with semantics. Instead, it will involve theoretical inquiry, for example into which properties are most relevant to biological classification (e.g. whether evolutionary origin is more theoretically relevant than capacity for interbreeding).

On the other hand, what is typically considered the corresponding "substantive" dispute over whether whales are fish—which can be contrasted both with the verbal and the factual identity disputes above—requires yet another mode of resolution. If A and B disagree about whether whales are fish while meaning the same by "fish"—or while agreeing on what it is to be a fish—then they plausibly disagree about questions such as whether whales are cold-blooded. Thus, resolving a factual dispute of this kind will involve checking whether whales have the relevant properties.

In short, the kind of dispute we are dealing with determines the kind of question we must answer. What is typically labeled as a "substantive" dispute over whether whales are fish can be resolved by answering questions like "Are whales cold-blooded?"; the corresponding verbal dispute can be resolved by asking "What does 'fish' mean?", or "How should 'fish' be used?"; and a factual identity dispute (FID) can be resolved by asking "What is a fish?". On views that conceive the subject matter of a dispute as a question—which in turn is standardly treated as a set of propositions (its possible answers)—the subject matter itself of the whales dispute depends on whether the dispute is "substantive" in the sense above, verbal, or a FID. Thus, the distinction between metalinguistic disputes and FIDs is, in fact, of real theoretical importance.

§2.2 To be F is to be G

A core claim of this chapter is that the parties to an allegedly verbal dispute may disagree on the criteria for something to count as F—a non-metalinguistic disagreement—rather than (merely) on the criteria for the term "F" to apply to something. In such cases, the disputants disagree about whether some o is F in virtue of disagreeing about some factual identity claim about what counts as F, or what it is to be F. In the whales dispute, for example, the disagreement can be taken to concern what fish are, rather than what the term "fish" means.

The claims that I have called "factual (non-metalinguistic) identity claims"—such as the claim that to be a fish is to be an aquatic vertebrate—are typically expressible in the form of "To be F is to be G" statements and serve to answer prototypically metaphysical questions such as "What is it to be F?".[9] These statements have received considerable attention in recent metaphysical literature (e.g. Rosen 2015, Dorr 2016, Correia 2017) and it is, in fact, controversial whether they should be taken to express higher-order identity claims or so-called "real definitions", and how these, in turn, should be interpreted. Still, the relevant point is that these statements concern the relevant properties themselves, rather than the meanings of the corresponding terms.

Schematically, we can say that in paradigmatic examples of allegedly verbal disputes, the parties agree that an object o has certain properties G, H, J, but still dispute whether o is F. On the metalinguistic account, this is explained in terms of metalinguistic disagreement: the disputants disagree about whether "F" means G. My contention, instead, is that the disputants may simply be taken to disagree about whether to be F is to be G. Of course, ordinary speakers often lack fully developed theories of what it is to be F, comprising of full

[9] Though they may take various surface forms, such as "An F is a G", where "an F" is a generic, or "Fs are Gs".

necessary and sufficient conditions for something to qualify as an F. Instead, they are more likely to hold beliefs about *some* necessary and/or sufficient conditions. As a result, many actual identity disputes will not involve disagreements over fully articulated identity claims. What matters for our purposes is that, in the disputes in question, the speakers may be taken to disagree on at least *some* of the criteria for something to count as an F (a genuine and non-metalinguistic disagreement), and *in virtue of that*, disagree about whether something is an F. By contrast, in what are normally taken to be "substantive" disputes, the parties are supposed to agree on the criteria for something to count as an F and disagree on whether something is an F in virtue of disagreeing on whether that thing meets those shared criteria.

Another important clarification is that FIDs can also stem from disagreement about *first-order* identity claims. As discussed in §1.2, disputes that the metalinguistic account classifies as verbal may be taken to arise from disagreement about merely extensional aspects of meaning, as in the case of Rolf and Scooter. On the metalinguistic account, Rolf and Scooter disagree about whether the name "Douglas" refers to Lancaster. However, an alternative non-metalinguistic (first-order) identity interpretation is also possible: the disputants may simply disagree about whether Douglas *is* Lancaster. Of course—as will be discussed in more details in §5—in a sense Rolf and Scooter do, in fact, agree that Douglas is not Lancaster, as both know that Douglas and Lancaster are different actors. However, it is not unnatural to think that, upon realizing the misunderstanding, they might laugh it off by saying "You really thought Lancaster was Douglas!". In fact, the metalinguistic interpretation, whereby Rolf mistakenly believes that "Douglas" refers to Lancaster, has a similarly implausible reading: there's a clear sense in which Rolf knows that "Douglas" refers to Douglas, not Lancaster.

So far, I have mostly discussed "ordinary" disputes, such as those over whether whales are fish and whether the Pope is a bachelor. For the sake of simplicity, this chapter

will focus on these ordinary simpler cases, without delving into more complex metaphysical disputes. Still, metaphysical disputes are central to the literature on verbal disputes, as they are often the targets of accounts of verbalness. It is thus important to note that the same line of reasoning can be extended to metaphysical disputes.

For example, disputes about personal identity are sometimes treated as paradigmatic cases of verbal$_{MD}$ disputes, where the parties are taken to disagree about what "person" means (Burgess & Plunkett 2013, Thomasson 2017). In contrast, the identity account treats these disputes as arising from disagreement about what it is to be a person. This non-metalinguistic disagreement could concern theoretically significant and "heavyweight" issues, such as whether biological, psychological, or sociological properties are more relevant for the individuation of persons, or which properties the property of being a person supervenes on.

Mereological disputes are another common target of accounts of verbalness. Hirsch (2005, 2009) famously argued that in these disputes, the parties can be seen as using languages where the quantifiers express different meanings. For mereological nihilists—who deny that mereological composition ever occurs— quantifiers only range over simples. In contrast, mereological realists—who believe that mereological composition does occur—hold that quantifiers also range over composite objects. Mereological disputes can also be interpreted as identity disputes: mereological nihilists and mereological realists may be taken to disagree about what it is for something to exist, or under what conditions something can be said to exist. Such disagreement may concern, for example, whether being a simple (something that lacks proper parts) is a necessary condition for existing.

In sum, given plausible assumptions, we can interpret paradigmatic examples of supposedly verbal disputes as arising from disagreement about non-metalinguistic identity claims. In §3 and §4, I will argue that disagreement about these identity claims does not

entail metalinguistic disagreement or semantic consistency, meaning that factual identity disputes are not necessarily verbal$_{MD}$ or verbal$_{SC}$. In §5, I will then discuss how these disputes present structural similarities to Frege puzzles.

§3 Disagreement About Identities Without Metalinguistic Disagreement

§3.1 Disagreement About Identities Without Metalinguistic Beliefs

Disagreement about identity claims is often accompanied by metalinguistic disagreement. Two speakers who disagree about whether to be a fish is to be an aquatic vertebrate will likely also disagree about some semantic properties of the term "fish"—e.g. its intension, extension, or the conditions for its correct application. As a result, many disputes involving factual disagreement about identities will also involve some kind of metalinguistic disagreement.

However, the connection between factual disagreement about identity and metalinguistic disagreement falls short of entailment. There are ordinary cases where speakers who hold the relevant factual beliefs about identity simply do not form any corresponding metalinguistic belief. Indeed, they might even lack the conceptual resources to do so. For instance, young children may hold the relevant beliefs about identities without having acquired or mastered any metalinguistic or metasemantic notions. In general, the idea is that agents without the sophistication required to entertain metalinguistic propositions can nonetheless form and hold beliefs about identity. Chalmers specifies that on his account, one need not be able to articulate an expression corresponding to "such-and-such" in order to have the metalinguistic belief that a term "T" means such-and-such. Still, even on his view, holding a metalinguistic belief plausibly requires having a belief of the form "T *means* such-

and-such"—that is, a belief involving semantic notions. One might object that holding metalinguistic beliefs does not require the possession of semantic concepts, but merely, for instance, a disposition to use certain terms in a particular way—such that even young children or individuals who do not engage in explicit metalinguistic reasoning could be said to hold metalinguistic beliefs. As I argue in §3.2, however, this is arguably still insufficient to render a dispute merely verbal.

It is worth emphasizing that, in other cases in which both a metalinguistic reading and a factual reading are available, interpretation in terms of factual beliefs about identity is typically the default. In standard Frege cases, for example, agents are generally taken to lack knowledge of identity facts, rather than *primarily* metalinguistic information. If Jane believes that Hesperus is bright but does not assent to "Phosphorus is bright", the standard explanation is that she does not know that Hesperus is Phosphorus—at least under the relevant informative guise—and not that she does not know that "Hesperus" refers to Phosphorus, or that "Hesperus" and "Phosphorus" co-refer—precisely for the reasons outlined above (see Nelson 2024). In these cases, the core issue concerns whether the speakers believe or disbelieve the relevant factual identity claims, regardless of whether they also hold the corresponding metalinguistic beliefs.

A further complication for metalinguistic accounts is that the notion of "metalinguistic disagreement" is often left underspecified. Chalmers (2011a), for instance, writes that the relevant notion of meaning can be left intuitive. However, the issue is that two speakers may at the same time agree on some dimensions of a term's meaning while disagreeing on others. Consider a person on Earth and another on Putnam's Twin-Earth (Putnam 1975). They may agree that "water" refers to the clear liquid substance filling lakes and seas, while disagreeing about its intension or extension. Whether disputes involving the

term "water" between these two speakers qualify as verbal$_{MD}$ depends on whether they count as disagreeing on the meaning of "water".

That said, a proponent of the metalinguistic account might object that, while some language users may lack metalinguistic beliefs about the relevant terms, *competent* language users cannot. In response, we should note that it is not typically assumed that both parties in a verbal dispute are competent with the relevant terms. In fact, there are clear examples to the contrary. For instance, someone who uses the term "horse" to refer to chairs, in a linguistic community where "horse" is used to refer to horses, arguably does not count as competent with the term. Yet such a speaker may still engage in disputes that qualify as verbal on the accounts above—e.g. about how many horses are in her kitchen, or whether horses are made of wood. Some of Chalmers' own examples (2011a) involve speakers who are clearly not fully competent users of the relevant terms—such as someone who uses "lie" to refer to any false statement. In general, whether the disputants count as competent language users does not seem relevant to assessing the verbalness of a dispute.

Even if we wanted to restrict the discussion to competent speakers, it is implausible to claim that competence with a term "T" requires grasping a verbally articulated proposition such as "'T' means T", on pain of an infinite regress: to grasp "'T' means T" one would need to grasp "'"T' means T' means that 'T' means T", and so on. Moreover, the view that linguistic competence requires *tacit* metalinguistic knowledge is itself controversial (see Schiffer 2006, Williamson 2007).[10] For one thing, as Evans (1985) argues, tacit beliefs are inferentially insulated from an agent's other beliefs, so it is not even clear that inconsistent tacit metalinguistic beliefs would amount to metalinguistic disagreement. Moreover, some

[10] Although a detailed discussion of this issue falls beyond the scope of this chapter, it is plausible to think that language learning is primarily a matter of imitation, rather than the acquisition of metalinguistic beliefs—a view defended, for instance, by Moore (2013) and Tomasello (2008), and seemingly consistent with the empirical evidence from psycholinguistics.

have argued that tacit metalinguistic knowledge is in fact a form of *practical* knowledge—a disposition to apply a term correctly—rather than propositional knowledge (Schiffer 2006). But mere differences in such dispositions do not constitute metalinguistic disagreement.

In any case, most accounts of linguistic competence, whether externalist or internalist, do not take competence to require metalinguistic knowledge at all. On externalist views, linguistic competence tends to be relatively "cheap", as it is largely a matter of one's interaction with the environment and linguistic community. For example, Williamson (2007) argues that linguistic competence—or linguistic understanding—is a matter of participation in the social practice of using the relevant terms: full participation in the practice constitutes full understanding. Indeed, Williamson emphasizes that participation in the social practice is more relevant to competence than holding (correct) metalinguistic beliefs. A non-native English speaker might know that "gob" means the same as "mouth" but not be fully aware of which social contexts make the use of "gob" inappropriate. As Williamson puts it, "knowing the meaning of an expression doesn't automatically qualify one for full participation in the practice of using it. Someone who acquires the word 'gob' just by being reliably told that it is synonymous with 'mouth' knows what 'gob' means without being fully competent to use it" (2007, p. 129). Even by more demanding internalist standards, however, linguistic competence is generally not taken to require explicit metalinguistic beliefs. On popular internalist accounts—as will be discussed in §4—competence with a term requires instead that the term play the right sort of conceptual role for the speaker, and that the speaker be disposed to assent to certain "meaning-constitutive" sentences.

§3.2 Substantive Disputes and Metalinguistic Disagreement

Even setting these issues aside and focusing only on cases in which factual disagreement about "To be F is to be G" claims is accompanied by metalinguistic disagreement, it is clear

that the mere presence of metalinguistic disagreement is not sufficient to render a dispute verbal$_{MD}$. This is because there are many clearly substantive disputes in which metalinguistic disagreement arises merely as a byproduct of an underlying factual disagreement about the relevant identity claims. For example, a physicalist and a functionalist about consciousness are likely to disagree about some dimensions of the meaning of the term "pain". In particular, they probably disagree about both its extension and intension: the physicalist holds that "pain" refers to a physical property at every possible world, whereas the functionalist holds that it refers to a functional property at every world. Yet, this is because they disagree about whether pain *is* a physical property or a functional one. The presence of the metalinguistic disagreement does not make their dispute verbal$_{MD.}$

To characterize this dispute as metalinguistic merely because it involves metalinguistic disagreement would be akin to claiming that Newtonian physics and relativistic physics merely diverge on semantics just because their conflicting claims about mass might generate conflicting *metalinguistic* claims about the intension of "mass". Their divergence would also count as verbal, if the mere presence of metalinguistic disagreement were enough to make a dispute verbal$_{MD}$. Yet, these cases clearly do not match the usual characterization of verbal disputes, whereby "nothing is substantively at stake beyond the correct use of language" and "the proper way to resolve these questions is by appealing to common sense and ordinary language" (Hirsch 2005, p. 67), or "when we discount any disputes arising from language-related differences, there is no relevant residual dispute or disagreement between the parties" (Jenkins 2014, p 19-20), and "the two parties agree on the relevant facts about a domain of concern, and just disagree about the language used to describe that domain" (Chalmers 2011a, p. 515). Part of the point is precisely that determining the correct use of language may require substantive inquiry into what the world

is like. In this sense, metalinguistic disagreement may reflect, and be parasitic on, a deeper and entirely factual disagreement.

It is worth noting that, although considerations of charity—according to which attributions of false claims and false beliefs to the disputants should be minimized—may provide a reason to endorse the Semantic Consistency account,[11] charity does not make the Metalinguistic Disagreement account any more plausible than the identity account. This is because both the Metalinguistic Disagreement account and the identity account attribute false claims and false beliefs to at least one of the disputants in the relevant cases. And there seem to be no compelling reasons to favour the attribution of false metalinguistic beliefs over false factual ones. In fact, as noted above, there may be grounds for resisting the idea that the disputants are merely disagreeing about language rather than about object-level issues—namely that the disputants themselves would often reject such a characterization.

The central issue, in short, is that most disputes involving substantive disagreement about identities will also involve metalinguistic disagreement. Therefore, the mere presence of metalinguistic disagreement cannot, by itself, suffice to make a dispute verbal$_{MD.}$ To distinguish genuine cases of merely metalinguistic disagreement from cases in which the metalinguistic disagreement is simply a byproduct of a factual disagreement over identities, we would need to determine whether the metalinguistic disagreement stems from an underlying non-metalinguistic one. Yet, this is often a very difficult task. In many cases, a simple "third-person" description of the dispute will not allow us to establish whether the disagreement is, *at bottom,* metalinguistic or not. And even from a "first-person" perspective, the matter may remain unclear: I myself am unsure whether I believe that to be a fish is to be a cold-blooded creature with gills in virtue of holding the corresponding metalinguistic belief that "fish" means cold-blooded creature with gills, or vice versa. If I

[11] Reasons *not* to endorse Semantic Consistency will be provided in §4.

were to engage in a dispute about whether whales are fish with someone who uses "fish" to refer to all aquatic vertebrates, I probably would not be able to tell whether the dispute ultimately stems from disagreement about what fish are, or about what "fish" means.

§3.3 Deflationism

So far, I considered whether metalinguistic beliefs are *causally* prior to the corresponding identity beliefs. The most plausible conclusion, I think, is that this question must be answered case by case, and that in most instances there is no straightforward way to determine the direction of dependence.

A further source of resistance to the idea that the disputes in question arise from disagreement over factual identity claims is the broadly deflationist view that all apparently factual identity claims should be *reinterpreted as* metalinguistic claims—or at least that their truth depends on the truth of the corresponding metalinguistic claims. Unlike the causal issue just discussed, this concerns the source of the truth of the relevant non-metalinguistic claims. In examining disputes about whether whales are fish and bats are birds, Sidelle writes:

> Upon what can the truth of the issue in dispute depend? The most—I propose—is that they can depend upon the established use of the words "fish" and "bird"—which is not a matter of the nature of fish or birds, or any other biological matter, but a matter, at best, of semantic history.
>
> Sidelle (2007, p. 92)

Objections to this "deflationist" approach have been widely discussed (e.g., see Fine 1994, Rosen 2015, Williamson 2007). Still, it is worth stressing that many accounts of verbalness do not explicitly rely on these assumptions. In a slogan, accounts of verbalness are not meant "for deflationists only". Indeed, Sidelle's question—"Upon what can the truth of the issue

in dispute depend?"—need not be answered by appealing to metalinguistic facts, contrary to what Sidelle himself seems to assume. Philosophers with realist inclinations may appeal instead to considerations concerning the relevant properties—for example, concerning which properties matter the most for the purposes of biological classification. If identifying fish as cold-blooded aquatic vertebrates with gills yields a simpler and more explanatory taxonomical model than identifying them as aquatic vertebrates of all kinds, then we have a non-metalinguistic reason to endorse the claim that fish are cold-blooded creatures with gills, regardless of how the term "fish" is used.

A further, related consideration against Sidelle's framework is that it is unclear how it can account for taxonomic revision. For instance, if the term "fish" was historically taken to mean aquatic vertebrate and used to refer to all aquatic vertebrates, it is unclear how the claim that fish are, in fact, only the cold-blooded aquatic vertebrates with gills could derive from an analysis of the meaning of the term "fish". On one version of the metalinguistic account, as explained in §1.2, the relevant metalinguistic claims are normative—they concern how a term *should* be used—rather than descriptive of current usage (e.g. Plunkett & Sundell 2021, 2023). However, barring pragmatic reasons to prefer a certain usage (discussed below) it is unclear what the truth of the *normative* metalinguistic claim could depend on, apart from non-metalinguistic claims about the properties themselves.

To illustrate, the deflationist claims that the truth of "To be a fish is to be an aquatic vertebrate" depends on the truth of the metalinguistic claim that "fish" means aquatic vertebrate (or is used to refer to aquatic vertebrates). My objection is that this view cannot account for taxonomic revision: how can we establish that fish are, in fact, cold-blooded aquatic vertebrates with gills if the truth of the relevant claim about what fish are depends on what "fish" means or how it is used? Of course, the deflationist might respond that the truth of the claim that fish are cold-blooded creatures with gills relies on the truth of the

corresponding normative metalinguistic claim that "fish" *should* be used to refer to cold-blooded creatures with gills. Yet, it remains unclear what the truth of the normative claim could depend on, if not the fact that the property of being a cold-blooded creature with gills is somehow more taxonomically important than the property of being an aquatic vertebrate. In other words, normative metalinguistic disagreement is also plausibly derivative of non-metalinguistic disagreement about the relevant properties.

In a later passage, Sidelle seems to acknowledge that these general questions—such as what fish are (whether all aquatic vertebrates or only the cold-blooded ones) or what pain is (a physical property P or a functional property F)—may concern which properties are more theoretically interesting. However, he seems to treat this question as a pragmatic issue concerning which property "deserve the label" of pain or fish. Regarding the pain debate, for instance, he writes:

> It should be appreciated and discussed not as a conflict over the nature of mental states (or even what is true in the relevant possible worlds), but over the semantics of mental terms, *or what properties and states are psychologically interesting*.
>
> (Sidelle 2007, p. 97, *my emphasis*)

Similarly, Chalmers (2011a) contends that there is no objective, mind-independent answer to most "what-is" questions. On his view, for example, there may be multiple candidate notions of fish, such that to be a $fish_1$ is to be an aquatic vertebrate, while to be a $fish_2$ is to be a cold-blooded creature with gills. Chalmers suggests that the choice between the relevant notions (e.g. $fish_1$ and $fish_2$) may be merely pragmatic:

> On the picture I favor, instead of asking 'What is X?', one should focus on the roles one wants X to play and see what can play that role […] There are multiple interesting concepts (corresponding to multiple interesting roles) in the vicinity […] and not much of substance depends on which one goes with the term.
>
> (Chalmers 2011a, pp. 538-539)

However, neither Sidelle nor Chalmers seem to consider that what matters for purposes like speciation or explanatory adequacy in taxonomy may not be a matter of convention or linguistic practice, but a function of which properties yield *objectively* better (e.g. more systematic and explanatorily powerful) theories, models, or taxonomical systems.

Consider a simplified example involving polar bears (*Ursus maritimus*) and brown bears (*Ursus arctos*). Although the issue is somewhat controversial, evolutionary studies suggest that polar bears may have diverged from ancestral brown bears approximately 500,0000 years ago and have since adapted to different ecological niches. Their morphology, behavior, and genetics reflect this divergence. Yet, polar bears and brown bears can interbreed and produce fertile offspring. Whether polar bears *are* brown bears (at the level of species) seems to depend on whether interbreeding is taken to be decisive, or whether evolutionary history is prioritized. This, however, does not seem to be the sort of issue that can be settled by consulting dictionaries or looking at ordinary usage, and it is entirely plausible that one classificatory criterion might yield an objectively more explanatory or systematic taxonomical model than the other. Thus, questions about which properties are theoretically preferable may well concern non-metalinguistic and mind-independent theoretical issues.

Of course, in some cases, prioritizing one property or another may yield equally good theories or models. In such cases, there may indeed be no objective, mind-independent fact of the matter as to whether to be F is to be G. This is essentially a version of Sider's point

that a dispute is verbal when the terms involved have no maximally joint-carving candidate meaning. For example, two candidate meanings for "bachelor" are unmarried man and unmarried man eligible for marriage. If neither option yields an objectively better model than the other, then the dispute might, in a sense, be considered verbal—though this sense of verbalness does not reduce to semantic consistency or metalinguistic disagreement. Still, this construal significantly narrows the range of disputes that count as verbal compared to typical classifications. Most importantly, it places a greater burden on the proponents of verbalness: namely, to show that none of the candidate properties would yield an objectively better model.

Furthermore, even if there were no mind-independent answers to the relevant "what-is" questions, there could still be mind-dependent facts about which the disputants disagree. That is, arguably, two individuals might be taken to disagree about whether X is the case even when there is no objective fact of the matter as to whether X is the case. For example, you and I might disagree about whether carbonara is the best Italian dish even if this question lacks an objective answer. In this sense, it seems plausible that two people may disagree about whether to be a fish is to be a fish_1 or a fish_2 even if there is no such *mind-independent* fact of the matter.

A general motivation for treating disputes in metaphysics as metalinguistic is that it is easier to make sense of metaphysics as a metalinguistic or conceptual practice within a naturalist framework. In such a framework, there seems to be little methodological room for addressing what-is questions unless they are interpreted as metalinguistic. This is why accounts of metaphysical disputes as verbal treat them as involving either merely metalinguistic disagreement or no disagreement at all. For example, if the truth of generalized identity claims were to depend on essentialist facts, whose epistemology may be regarded as obscure, metaphysics might appear to fall outside the bounds of a respectable

naturalist framework. Therefore, the idea that what-is questions can be answered through metalinguistic inquiry is supposed to clarify the methodology of metaphysics.

However, this seems to rest on an overly simplistic and exceptionalist conception of metaphysics and its methodology. First, several of the examples discussed above show that, in fact, "what-is" questions are not distinctively philosophical; taxonomical questions of this kind arise in many scientific disciplines. More importantly, these questions can be answered by appealing to abductive considerations of simplicity and explanatory power. Determining what a fish is, for instance, can be a matter of identifying which properties would yield the best taxonomical model. In this sense, the idea that metaphysics is not metalinguistic or metaconceptual does not necessarily involve postulating essentialist or otherwise obscure metaphysical facts. Rather, the abductive methodology commonly used in other fields of enquiry is often sufficient to answer these questions.[12]

§4. Disagreement About Identities Without Semantic Consistency

§4.1 Semantic Consistency and Externalism

So far, I have argued that disagreement about identity claims does not entail metalinguistic disagreement, and that, when metalinguistic disagreement does occur, it is often a mere byproduct of disagreement about identity. This means that FIDs are not necessarily $\text{verbal}_{\text{MD}}$. In this section, I will argue that disagreement about identities does not entail semantic consistency—FIDs are not necessarily $\text{verbal}_{\text{SC}}$ either.

[12] For a discussion of the abductive method in philosophy, see Williamson (2021b, section 9.2).

On the Semantic Consistency account, the divergence in how the disputants use the relevant terms leads to a semantic divergence, such that the terms in question express different meanings as used by each party. As a result, their apparently inconsistent assertions are thought to express mutually consistent propositions. This is also assumed to hold at the level of mental content: the disputants are not truly in disagreement because their apparently conflicting beliefs are, in fact, mutually consistent.

However, as briefly discussed in §1.2, the Semantic Consistency account runs into conflict with influential externalist views such as those defended by Burge (1979) and Putnam (1975). Putnam argues that the meaning of a term is not determined solely by a speaker's internal states or representations, but also depends on features of the external environment. Putnam asks the reader to imagine a planet—Twin Earth—that is exactly like Earth in every respect except one: the liquid that fills the lakes, falls from the sky, and is called "water" by its inhabitants is not H_2O, but a different chemical substance with the same appearance and behavior. Suppose the year is 1750, before anyone on Earth or Twin Earth knows about chemical composition. An inhabitant of Earth, Oscar, and his exact molecular duplicate on Twin Earth, Twin-Oscar, both say "Water is wet". Although the utterances are identical in sound and grammar, Putnam argues that they differ in meaning because the word "water" refers to different substances in the two environments: H_2O on Earth and some other substance—call it "t-water"—on Twin Earth. Since the reference and truth-conditions differ, the meaning of the utterances differs too. This supports the idea that meaning is not just "in the head", but depends partly on the speaker's environment.

Burge extended this idea to mental content. In his "arthritis" case, Larry falsely believes he has arthritis in his thigh. This belief is false, because in English "arthritis" refers to a condition affecting the joints. Still, Burge argues that Larry's belief is a belief about arthritis, because Larry inhabits a linguistic community made up of people who use the term

"arthritis" in a certain way, and would defer to experts in that community, if corrected. Thus, when a doctor tells him, "No, that's not arthritis you have. Arthritis is an affliction of the joints", Larry would be inclined to accept this, and revise his belief accordingly. In a counterfactual community where "arthritis" covers a broader range of ailments, the subject's internally identical belief would be true, but it would not be about arthritis. Burge's point is that the contents of Larry's beliefs differ across the two cases, even though everything internal to Larry is the same. The difference in content arises solely from differences in external linguistic practice.

If the meaning of the relevant terms is fixed by factors external to the speakers (as Putnam argued) or by their joint practice (as Burge argued), then these terms will express a certain meaning *regardless* of the idiosyncrasies of individual uses. For example, if the meaning of "fish" is determined by external factors, both disputants will express that same meaning when they use the term, regardless of any differences in their usage. This entails that when they utter the disputed sentence S and its negation not-S (respectively), they will express genuinely inconsistent propositions. Furthermore, if, as Burge contends, mental content is also shaped by such external factors, the disputants' beliefs will likewise be mutually inconsistent, and the speakers will genuinely disagree. Thus, if externalism is true, very few disputes qualify as verbal$_{SC}$—presumably, only those involving polysemous words, homophonic words in different languages, or indexical/demonstrative terms.

For similar reasons, Semantic Consistency is also in tension with views like reference magnetism—the idea that some candidate meanings are intrinsically more eligible than others (Lewis 1983, Sider 2011). According to Sider (2011), for example, there is a privileged concept of existence, which serves as an intrinsically more eligible candidate meaning for "exist", such that "exist" must express that concept regardless of how the term is used. This is why, according to Sider, paradigmatic examples of allegedly verbal disputes,

such as mereological disputes, cannot satisfy Semantic Consistency: no matter how the disputants use the relevant terms, their utterances will express genuinely conflicting propositions, as the terms involved will express the same meaning in both speakers' mouths. Indeed, Sider claims, only those disputes in which the relevant terms have no candidate meaning which is intrinsically more eligible than others may qualify as verbal$_{SC}$.

§4.2 Semantic Consistency and Conceptual Role Semantics

A proponent of Semantic Consistency could, of course, simply reject externalism. However, Semantic Consistency also seems to conflict with certain non-externalist views, such as conceptual-role semantics. In a conceptual-role semantic framework, the meaning of a term is determined by so-called "meaning-constitutive" sentences, which are typically understood to be analytically or a priori true (Block 1996, Boghossian 1996). For example, the meaning of "bachelor" is taken to be determined by the meaning-constitutive sentence "All bachelors are unmarried men". Similarly, the meaning of "water" is purportedly determined by a sentence like "Water is the liquid transparent substance that fills lakes and seas".[13] This framework is meant to explain the intuitive sameness of content in Twin Earth-style cases: although after Putnam (1975) it is generally accepted that Oscar's "water" and Twin-Oscar's "water" differ in meaning, the term plays the same conceptual role in the two cases. Both Oscar and Twin-Oscar, for example, are disposed to assent to the statement "Water is the liquid transparent substance that fills lakes and seas", and to infer "This is water" from "This is the liquid transparent substance that fills lakes and seas". The idea that *some* layer of content is tied to conceptual role helps vindicate the intuition that, despite being about different substances, Oscar's and Twin-Oscar's mental states have something in common—their "narrow" content. The disposition to assent to these meaning-constitutive sentences is

[13] For present purposes, I will set aside potential worries concerning the a priori/a posteriori distinction.

also taken to determine linguistic competence with the relevant terms (Boghossian 1996). For example, one who fails to assent to "All bachelors are unmarried men", or rejects the inference from "John is an unmarried man" to "John is a bachelor", is regarded as incompetent with the term "bachelor".

In the most common version of this framework, it is the conceptual/inferential role that a term plays for the individual which determines its meaning. Thus, a divergence in two speakers' attitudes towards the relevant meaning-constitutive sentences is supposed to determine a difference in the meaning expressed by the term in question as used by each speaker. In other words, if I use the term "bachelor" according to the inference rule above but you do not, we will express different things by "bachelor", as the term plays different conceptual/inferential roles for us.

In the FIDs under consideration, the parties, by hypothesis, diverge in their attitudes towards statements that, like the meaning-constitutive sentences above, appear to express "definitional" claims—e.g. "Fish are cold-blooded creatures with gills", or "Water is H_2O". If such divergence were sufficient to generate a change in meaning, then factual disagreement about identity would entail semantic consistency, and FIDs would thereby qualify as verbal$_{SC}$. To illustrate, if divergence in attitudes towards the claim "To be a fish is to be a cold-blooded creature with gills" were enough to determine a difference in the meaning of the term "fish", then no two speakers could genuinely disagree about whether to be a fish is to be a cold-blooded creature with gills, or about whether whales are fish, without talking past each other.

However, as noted above, conceptual-role frameworks typically restrict the meaning-constitutive sentences to supposedly *a priori* sentences such as "All bachelors are unmarried men", "Water is the liquid transparent substance that fills lakes and seas", and "Vixens are female foxes". Yet, in many cases, two speakers may share the disposition to assent to the

relevant *a priori* meaning-constitutive sentences while diverging in their attitudes toward *a posteriori* and non-analytic identity claims. For instance, two speakers might both assent to "Water is the liquid transparent substance that fills lakes and seas" yet differ in their attitudes towards "Water is H_2O". The latter is the standard example of an *a posteriori* and *non-analytic* identity statement. Accordingly, if meaning-constitutive sentences are required to be analytically or a priori true, "Water is H_2O" should not count as meaning-constitutive for "water". Analogous examples can be drawn from a wide range of familiar identity claims: two speakers might both assent to "Gold is the yellow precious metal" yet diverge on "Gold is the element with atomic number 79". The same applies to "Heat is what causes sensations of warmth" and "Heat is molecular motion", or "Pain is whatever causes sensations of pain" and "Pain is the firing of C-fibers". These pairs are paradigms of supposedly a priori versus supposedly a posteriori claims, central to philosophical debates at least since Kripke (1980). To illustrate, consider the following two sentences:

W': Water is the clear liquid substance that fills lakes and seas
W'': Water is H_2O

What I have aimed to highlight in this section is that if you and I align in our dispositions toward W', we thereby count as meaning the same by "water" by conceptual-role standards and we can thus *genuinely* disagree on whether water is H_2O. This runs against the idea that merely by disagreeing about whether water is H_2O—or, more precisely, by failing to align in our dispositions toward W''—we *thereby* mean different things by "water". The point is thus that we can disagree about an identity claim (and thus engage in a FID) without our dispute thereby counting as verbal in the sense of semantic consistency.

That said, of course, it is also possible that we align in our dispositions toward W'' but not toward W'. In that case, we would indeed count as meaning different thing by "water" within a conceptual-role (or two-dimensional) framework. The divergence in our dispositions toward a meaning-constitutive sentence like W' could itself be taken to stem from an underlying factual disagreement about what counts as water. Yet, a proponent of conceptual-role semantics could maintain that the very fact that we diverge in our attitudes toward W' entails that we do not mean the same by "water", and thus that any apparent disagreement we might have about what water is—including about whether water is the liquid clear substance that fills lakes—whether or not it's the source of our divergence in dispositions, does not amount to a genuine disagreement.

In any case, the point emphasized in this section is, again, that disagreement over identity claims does not necessarily determine a change in meaning. Even in a conceptual-role framework, a term may retain the same meaning for both parties—provided they align in their attitudes towards the relevant meaning-constitutive sentences—and the disputed sentence may express the same proposition as uttered by both speakers. Thus, such disputes will not necessarily qualify as verbal$_{SC}$. Moreover, even on accounts of linguistic competence that tie it to assent (or disposition to assent) to the relevant analytic sentences, both disputants may count as competent users of the terms involved: despite disagreeing on "To be F is to be G" claims, they may still be disposed to assent to the relevant meaning-constitutive analytic sentences.

§4.3 Holism

For the cases considered above to qualify as verbal$_{SC}$, one would need to assume that meaning-constitutive sentences are not restricted to a priori or analytic claims. If even paradigmatically a posteriori statements—such as "Water is H_2O"—are considered

meaning-constitutive, then any disagreement about identity leads to a change in meaning. However, allowing not only supposedly a priori or analytic statements to count as meaning-constitutive introduces a serious risk of semantic holism—the very problem that the restriction to a priori/analytic sentences was designed to avoid. This move could lead to the highly implausible conclusion that any divergence in attitude toward a sentence containing the term prevents genuine disagreement between speakers. Moreover, any disposition to assent to a false sentence involving the term would, on this view, render a speaker less than fully competent with its use (see Fodor & Lepore 1991).

An opponent might reply that, although not *every* sentence containing the term in question is meaning-constitutive, all "To be F is to be G" statements are, regardless of whether they are analytic. While this avoids the commitment to full-blown semantic holism, it still yields highly undesirable consequences. Any disagreement about a posteriori identities would become impossible: the meaning of "water", for instance, would shift simply due to the disputants' divergent attitudes towards "Water is H_2O"—even if the speakers agree that water is the clear liquid that fills lakes and seas.

Moreover, if linguistic competence is defined in terms of assent to meaning-constitutive sentences, and a posteriori identity statements are included among them, then any speaker who holds a false belief about an *a posteriori* identity is classified as incompetent—or at least less than fully competent—with the relevant term. For example, suppose I am not disposed to assent to the statement "Gold is the element with atomic number 79" because I mistakenly believe that gold has atomic number 80. On the view in question, I would count as not competent with the term "gold", despite being otherwise fully integrated in the communal practice of using it, and despite my disposition to assent to statements such as "Gold is a yellow metal", "Gold is rare and expensive", "Gold is used to make jewels", etc. This conclusion is clearly very questionable.

Importantly, some disputes that are typically labeled as verbal do qualify as verbal$_{SC}$ (or verbal$_{MD}$) even by externalist standards. A clear example involves disputes centered on demonstrative or indexical terms. Consider the following scenario. I am in London, speaking on the phone with my friend James, who normally lives in the UK but—unbeknownst to me—is spending the holidays in Italy. James mistakenly believes that I too am in Italy for the holidays. As a result, I believe that both of us are in the UK, while James believes that both of us are in Italy. Suppose that, frustrated with the English weather, I say "This country is too rainy" (call this sentence 'S'). Surprised, James replies "No, this country is not too rainy". Since I assert S and James responds by asserting not-S, we are clearly engaging in a dispute. However, we are expressing consistent propositions: "this country" as used by me refers to the UK; as used by James it refers to Italy. This follows from the standard semantics for indexicals and demonstratives, without relying on any of the semantic assumptions discussed earlier. Thus, the dispute is plausibly verbal$_{SC}$. Assuming that James and I also hold the relevant metalinguistic beliefs, we presumably also disagree about what "this country" refers to, and about what proposition S expresses as uttered by each of us. Therefore, the dispute also qualifies as verbal$_{MD}$.

Other cases of genuinely verbal disputes may arise when the disputed sentence is genuinely ambiguous. Consider, for instance, this example from Vermeulen (2018):

Muriel: "I've seen a thief with our telescope".
John: "No you haven't. The telescope is upstairs right where it belongs".

Here, Muriel means that she saw a thief *through* the telescope, while John understands her to mean that she saw a thief *carrying* the telescope, which he denies since the telescope is

still upstairs. Similarly, one can construct verbal disputes involving polysemous terms like "bank", where one speaker refers to a financial institution and the other to a riverbank:

Jack: "I saw a goose by the bank the other day"
Jill: "No you didn't. There are no geese by the bank".

Again, these cases arguably qualify as verbal by both Semantic Consistency and Metalinguistic Disagreement accounts. Because of the genuine ambiguity of the terms and sentences in question, the disputants can be taken to express consistent propositions, and—assuming they have the relevant metalinguistic beliefs—disagree about the meaning of key expressions. In such cases, the dispute arguably arises from linguistic misunderstandings, rather than from underlying factual disagreements.

§5 Identity Disputes as Frege Puzzles

§5.1 The Objection from Complete Agreement

So far, we have characterized paradigmatic FIDs as disputes in which the parties agree that some object o is G (and, potentially, H, J, etc.), and disagree about whether o is F in virtue of disagreeing about whether to be F is to be G. We have noted that, without further assumptions, FIDs do not necessarily qualify as verbal_{SC} or verbal_{MD}.

However, one may object that, even if these disputes are not verbal, they are still somehow defective. If it is indeed the case that to be F is to be G, and the disputants agree that o is G, then don't they, in a sense, already agree on all the relevant information? And if so, isn't their dispute somehow idle, or insubstantial? The idea that the disputants agree on

all the relevant non-metalinguistic information is, indeed, central to many accounts of verbalness. For example, Jenkins (2014) writes:

> The moral of this, I take it, is that we need to incorporate into our criteria for a dispute's being merely verbal some way of ensuring that, when we discount any disputes arising from (and/or identical to) language-related differences, there is no relevant residual dispute or disagreement between the parties.[…] Merely verbal disputes, then, are ones in which the dispute arises only in virtue of the parties' divergent uses of language. In effect, this moral combines the first theme of this section: that there is no substantive, relevant disagreement between the parties to a merely verbal dispute, with the second theme: that a merely verbal dispute is one that arises in virtue of differences concerning language.
>
> (Jenkins 2014, pp. 19-20)

Similarly, Sidelle treats agreement on the relevant non-metalinguistic information as the "tell-tale sign" of verbalness:

> Since we agree 'on the facts', and the choice of words is what creates the illusory appearance of disagreement, I call it a verbal dispute. […] in all of these cases, we have our above-mentioned tell-tale sign of verbal disputes: the parties agree on all the relevant facts.
>
> (Sidelle 2007, pp. 90-92)

A related idea can be found in Chalmers (2011a). In addition to his metalinguistic definition of verbalness, Chalmers proposes an alternative test for identifying verbal disputes, which he calls the "method of elimination". The core idea is this: if the problematic term at the

heart of a dispute is eliminated and the disputed sentence is paraphrased without it, then the presence or absence of a *residual* dispute helps determine whether the original dispute was substantive or merely verbal. If no residual dispute remains after paraphrasing, the dispute is classified as verbal; if one does, the dispute is substantive. One exception, according to Chalmers, concerns "bedrock" disputes—cases in which the term to be eliminated cannot be analyzed in more basic terms. In such cases, once the term in question is removed, we may be unable to formulate a residual disagreement—not because the original dispute was merely verbal, but because we have simply exhausted the relevant vocabulary.[14]

Of course, what counts as a legitimate paraphrase of the disputed sentence is itself highly controversial. Still, it seems reasonable to assume that (1) "Whales are aquatic vertebrates" and (1b) "Whales are cold-blooded creatures with gills" qualify as acceptable paraphrases of the disputed sentence (2) "Whales are fish". By hypothesis, (1) and (1b) are undisputed in the case in question: the disputants agree that (1) is true and (1b) is false. According to the method of elimination, the original dispute over (2) is therefore verbal, since the parties agree on all the relevant information once the problematic term "fish" is removed. The central thought here is precisely that agreement on (1) and (1b) amounts to complete agreement; thus, any remaining dispute over (2) is idle and insubstantial.

As established in previous sections, the dispute over (2) is not necessarily verbal$_{SC}$ or verbal$_{MD}$. Still, one might press the idea that such complete agreement renders the dispute

[14] A case in point, according to Chalmers, involves disputes about consciousness: "Suppose we disagree over 'Mice are conscious'. If we bar the term 'conscious', are there residual disagreements? We might disagree over 'Mice are phenomenally conscious', but this is at best clarification. If we bar 'phenomenally conscious', it appears that we are left with cognate disagreements over sentences such as: 'Mice have experiences' and 'There is something it is like to be a mouse'. Again, this is not huge progress. Once enough phenomenal terms are barred, it may be that no disagreement can be stated. We might agree on all the nonphenomenal properties of a mouse but disagree on whether it is phenomenally conscious. Again, however, it would be hasty to conclude that the original dispute is verbal. Instead, we have simply exhausted the relevant vocabulary. Intuitively, once we reach a certain point ('phenomenally conscious', say), we have reached bedrock" (2011a, p. 544).

defective, even if not verbal. I will refer to this objection as the "Objection from Complete Agreement" (OCA). In what follows, I argue that the disputants' agreement on (1) and (1b) does not render the dispute over (2) defective.

§5.2 Identity Disputes and Frege Puzzles

Consider again the following sentences:[15]

(1) Whales are aquatic vertebrates.

(2) Whales are fish.

Suppose that, as a matter of fact, being a fish *just is* being an aquatic vertebrate. In that case, "fish" and "aquatic vertebrate" are co-intensional: they have the same extension in any possible world. To assess whether OCA succeeds, we must answer two questions. First, given that the two expressions are co-intensional, do (1) and (2) express the same proposition? Second, if they do, does agreement on (1) entail agreement on (2)? These questions are closely related to those arising in typical instances of Frege's puzzle (1892). Consider Frege's well-known Hesperus/Phosphorus example:

(1*) Hesperus is bright

(2*) Phosphorus is bright

The puzzle arises from the fact that "Hesperus" and "Phosphorus" are just names for the same object, Venus. By the principle of compositionality, we should be able to substitute

[15] For simplicity, I will focus on (1), but a parallel argument can be made for (1b).

terms with the same semantic value *salva veritate*: if "Hesperus" and "Phosphorus" have the same semantic value, then substituting one for the other should preserve the truth-value of any sentence in which they occur.[16] However, substitution seems to fail when the relevant terms occur within the scope of a propositional attitude verb, such as "know", "believe" or, indeed, "agree". Even though "Hesperus" and "Phosphorus" co-refer, it seems possible for someone to believe that Hesperus is bright without believing that Phosphorus is bright. For instance:

(3*) Jane believes that Hesperus is bright

can be true, even if

(4*) Jane believes that Phosphorus is bright

is false. After all, Jane might simply not know that Hesperus is Phosphorus. The core questions arising in this paradigmatic Frege puzzle—whether (1*) and (2*) express the same proposition and whether believing (1*) entails believing (2*)—mirror those arising in the whale case above.

§5.3 Fregeanism and Anti-Fregeanism

There are several standard strategies for responding to Frege puzzles, which are directly relevant to the cases discussed in this chapter (see Nelson 2024). On one side, we find "Fregean" strategies, which rely on a fine-grained semantic framework. Frege's original view (1892) held that terms like "Hesperus" and "Phosphorus", though co-referential, differ

[16] This excludes quotational contexts.

in *sense*, and that in opaque contexts—such as propositional attitudes ascriptions—a term refers not to its ordinary referent but to its usual sense. This, Frege argued, accounts for the failure of substitution: "Hesperus" and "Phosphorus" do not share the same semantic value in these contexts. Contemporary versions of the Fregean view (e.g. Forbes 1990, Chalmers 2011b) share the idea that co-referential terms—like "Hesperus" and "Phosphorus"—may pick out their referent via different *modes of presentation*, which contribute semantically to the proposition expressed. On this view, reference and intension do not exhaust semantic content: two expressions can be co-referential and co-intensional yet differ in meaning because they involve distinct modes of presentation. Consequently, "Hesperus is bright" and "Phosphorus is bright" express different propositions, say p and q, respectively—thereby allowing (3*) and (4*) to differ in truth-value: there is no contradiction in Jane's believing a proposition p without believing a *different* proposition q. The same applies, *mutatis mutandis,* to the pair "fish" and "aquatic vertebrate": even if being a fish just is being an aquatic vertebrate, the Fregean will hold that (1) and (2) do not express the same proposition. It follows that believing (1) does not entail believing (2), which means that agreement on (1) does not entail agreement on (2). Therefore, OCA poses no threat to the Fregean account.

On the other side, we find anti-Fregean coarse-grained views, such as Russellianism and intensionalism. Russellianism holds that propositions are structured entities composed of the objects and properties referred to by the terms in the relevant sentences. Since "Hesperus" and "Phosphorus" co-refer, the sentences "Hesperus is bright" and "Phosphorus is bright" express the same proposition—one composed of Venus and the property of being bright. Intensionalism, by contrast, identifies propositions with sets of possible worlds. Given that Hesperus and Phosphorus are identical, the set of worlds in which Hesperus is bright is the same as the set of worlds in which Phosphorus is bright. Hence, on either account, "Hesperus is bright" and "Phosphorus is bright" express the same proposition.

Analogous considerations apply to our whales case. If, as assumed, to be a fish just is to be an aquatic vertebrate, sentences (1) and (2) will express the same proposition within an intensional framework. Similarly,

(5) To be an aquatic vertebrate is to be an aquatic vertebrate.

and

(6) To be a fish is to be an aquatic vertebrate.

also express the same proposition on such a view. Indeed, on anti-Fregean accounts, even "Hesperus is Hesperus" and "Hesperus is Phosphorus" express the same proposition. This, of course, gives rise to the familiar challenge of explaining how the latter can appear informative and only knowable a posteriori, while the former seems trivial and a priori. Matters become more complicated on Russellian accounts of propositions. Russellian propositions are often taken to mirror the compositional structure of the sentences that express them. Since there is compositional structure in "aquatic vertebrate" but not in "fish", the Russellian propositions expressed by (1) and (2), and by (5) and (6), should differ. Because this issue depends on the specific features of the example, I will set aside these complications for present purposes. Still, it is worth noting that this may introduce an additional difficulty for proponents of OCA, which would then apply only to cases involving terms with the same compositional structure and same denotations for simple constituents within Russellian frameworks.

The question is now whether, given that (1) and (2), and (5) and (6), express the same propositions on anti-Fregean accounts, the disputants' agreement on (1) entails agreement on (2) and their agreement on (5) entails agreement on (6). To answer this question, we need to distinguish between two versions of anti-Fregeanism.

According to *contextualist* anti-Fregeanism (e.g. Crimmins & Perry 1989, Richard 1990, Schiffer 1992), propositional attitude ascriptions are sensitive to contextually relevant modes of presentation of the proposition in question. On this view, "S believes that p" is true in a context c if and only if S believes the proposition p under some mode of presentation relevant in c.[17] For example, "Jane believes that Hesperus is Hesperus" is true in c because Jane believes the proposition p expressed by "Hesperus is Hesperus" under the homophonic mode of presentation—or *sentential guise*—"Hesperus is Hesperus". Even though "Hesperus is Hesperus" and "Hesperus is Phosphorus" both express p, Jane may fail to believe p under the non-homophonic guise "Hesperus is Phosphorus". In this case, Jane cannot be correctly ascribed the belief that Hesperus is Phosphorus (in those words), despite believing that Hesperus is Hesperus.

By the same reasoning, agreement on (1) and (5) does *not* entail agreement on (2) and (6). For two speakers A and B to agree on (2), it is not sufficient that they believe the proposition expressed by (2): they must also believe it under the specific guise "Whales are fish". But since, by hypothesis, one party rejects that sentence, the two cannot be said to agree on (2), even if they agree on (1). The same applies to (5) and (6): agreement on the trivial identity claim "To be an aquatic vertebrate is to be an aquatic vertebrate" does not entail agreement on "To be a fish is to be an aquatic vertebrate", even though these express the same proposition. Thus, the contextualist version of anti-Fregeanism also blocks the Objection from Complete Agreement.

Finally, we have *anti-contextualist* anti-Fregean frameworks (Braun 1998, Salmon 1986, Williamson 2021a). Proponents of this view adopt a "coarse-grained" anti-Fregean

[17] One way to spell out this idea is to say that "S believes that p" is true in a context just in case S believes p *and* is disposed to assent to some contextually relevant sentence expressing p. This is what Williamson (2021a) calls a "language-sensitive" account of belief-ascription.

semantics according to which (1) and (2) express the same proposition and (5) and (6) do as well. This is combined with an anti-contextualist account of propositional attitude ascription, on which such ascriptions are not sensitive to the mode of presentation under which the proposition is believed. Therefore, on this view, believing (1) entails believing (2), and believing (5) entails believing (6). Given that both disputants believe (1), thereby agreeing on (1), they also both believe (2), thereby agreeing on (2); and given that they agree on (5), they also agree on (6).

This conclusion appears to clash with the claims made in previous sections—namely that, contrary to Semantic Consistency, FIDs involve genuine disagreement. In the whales case, the disagreement concerns both whether to be a fish is to be an aquatic vertebrate (6) and whether whales are fish (2). But if, in this framework, the disputants agree on all the relevant propositions, what exactly are they disagreeing about? The answer is that, within an anti-contextualist anti-Fregean framework, the disputants disagree on the very same propositions that they also agree on. This is because, on the one hand, proponents of this view retain a standard account of belief ascription, whereby we ascribe beliefs to speakers based on the sentences they are disposed to assent to. On the other hand, however, anti-contextualists hold that a speaker's *not* assenting to a sentence does not entail that they fail to believe the proposition it expresses. In other words, while assent to a sentence S is typically taken to entail belief in the proposition p expressed by S, the lack of assent to S does not, on this view, entail lack of belief in p. Recall that the whales dispute arises from the fact that disputant A asserts (2) and a disputant B denies (2), despite their agreement on (1). Since B assents to the negation of (2), the standard account licenses ascribing to B a belief in the negation of (2). Yet, as we have seen, on anti-contextualist anti-Fregeanism, B is also ascribed a belief in (2), in virtue of her belief in (1). The same reasoning applies to (5) and (6).

The apparently counterintuitive result, then, is that B believes both (2) and (6) and their negations. But this is not unique to the whales dispute: under anti-contextualist anti-Fregeanism, any Frege puzzle yields similar outcomes. For example, suppose Jane believes that Hesperus is Hesperus and also that Hesperus is not Phosphorus. On this view, Jane believes both the proposition that Hesperus is Hesperus and its negation. Similarly, in the whales case, A and B agree that whales are fish under the guise "Whales are aquatic vertebrates" but disagree about whether whales are fish under the guise "Whales are fish". Likewise, they agree that to be a fish is to be an aquatic vertebrate under the homophonic guise "To be an aquatic vertebrate is to be an aquatic vertebrate" but not under the guise "To be a fish is to be an aquatic vertebrate". This may seem to pose a serious difficulty for anti-contextualist anti-Fregeanism, since attributions of inconsistent beliefs are typically taken to amount to attributions of irrationality. However, proponents of this view argue that believing both a proposition and its negation does not suffice for irrationality unless both are believed under the same guise, or mode of presentation.

§5.4 Anti-Contextualist Anti-Fregeanism and Complete Agreement

Leaving these complications aside, we have established that anti-contextualist anti-Fregeanism entails that FIDs involve complete agreement between the disputants: their agreement on (1) and (5) entails agreement on (2) and (6). Accordingly, this view is subject to the objection that, although FIDs are not merely verbal, they are still defective in some sense.[18]

[18] Note that, even given scepticism about "To be F is to be G" claims (e.g. assuming that there is no fact of the matter as to whether fish are aquatic vertebrates), OCA still goes through: even leaving aside (5) and (6), so that only (1) and (2) are relevant, agreement on (1) still entails agreement on (2)—hence, complete agreement persists.

In response to OCA, we should note that under anti-contextualist anti-Fregeanism, many perfectly legitimate and intuitively non-defective disputes involve complete agreement. Indeed, under such assumptions, all instances of Frege's puzzle involve complete agreement. For example, suppose that Jack and Jill see someone in the distance and dispute whether that person is their friend John. Jack says "That's John!" and Jill replies "No, that's not John". They obviously agree that John is John. Yet—given anti-Fregeanism and the standard assumptions that demonstratives and proper names are directly referential and that true identity claims involving them are necessarily true—"That's John" and "John is John" express the same proposition, if the referent of the demonstrative is indeed John. Thus, anti-contextualism entails that Jack and Jill agree that the person in the distance is John under the guise "John is John", despite their explicit disagreement.

The same point applies to other classic puzzles. There seems to be nothing defective about a dispute over whether Mark Twain was Samuel Clemens even if both parties agree that Samuel Clemens was Samuel Clemens. Nor is there anything defective about a dispute over whether water is H_2O, even if both parties agree that water is water. If agreement on all the relevant (coarse-grained) propositions involved were sufficient to render a dispute defective, then we would have to treat all of these disputes as defective.

Furthermore, given anti-contextualism, the issue of complete agreement also arises for the metalinguistic account. On that view, the whales dispute arises because the disputants A and B disagree on whether

(7) "Fish" means aquatic vertebrate.

However, on anti-Fregean assumptions, (7) expresses the same proposition as

(8) “Fish” means fish.

Since A and B plausibly agree on (8), anti-contextualism entails that they also agree on (7). Likewise, A and B may disagree about

(9) “Whales are fish” expresses the proposition that whales are aquatic vertebrates.

But, again, if (9) expresses the same proposition as

(10) “Whales are fish” expresses the proposition that whales are fish,

then A and B’s agreement on (10) entails agreement on (9). Hence, even the metalinguistic account inherits the problem of complete agreement: under the “disquotational” guises (8) and (10), A and B agree on the meaning of “fish” and on the proposition expressed by “Whales are fish”, respectively. In both the factual and the metalinguistic accounts, agreement and disagreement will need to be relativized to guises, or modes of presentation, of the relevant propositions.[19]

In sum, we have three replies to OCA. First, the objection only threatens anti-contextualist anti-Fregean views. Second, we have shown that even ordinary, intuitively

[19] In some cases, the appeal to *sentential* guises does not seem to resolve the problem. Consider a variant of Kripke’s Paderewski case (1979): A and B are disputing whether Paderewski can play Mozart’s Sonata in C major, but A thinks of Paderewski as the politician, while B thinks of Paderewski as the pianist. On the factual interpretation, the dispute stems from disagreement about whether Paderewski is Paderewski. But A and B presumably also agree that Paderewski is Paderewski. Relativizing to sentential guises will not help: A and B both agree and disagree about whether Paderewski is Paderewski under the very same sentential guise. But again, the metalinguistic account does not seem to offer a better explanation: on the metalinguistic interpretation, A and B disagree about the reference of “Paderewski”—whether “Paderewski” refers to Paderewski. Yet, A and B surely also agree that “Paderewski” refers to Paderewski.

non-defective disputes involve complete agreement given these assumptions. Finally, the problem of complete agreement extends to the metalinguistic account as well.[20]

§ Conclusion

The central claim has been that paradigmatic examples of allegedly verbal disputes are better understood as factual identity disputes (FIDs)—disputes that are neither verbal nor otherwise defective. FIDs qualify as $\text{verbal}_{\text{MD}}$ only if we assume that identity claims are metalinguistic—a controversial assumption that should not be treated as the default in theorizing about verbal disputes. Interpreting FIDs as $\text{verbal}_{\text{SC}}$ likewise involves specific and controversial semantic commitments. A further objection considered was that, even if FIDs are not verbal, they may still be defective, insofar as the disputants seem to agree on all the relevant non-metalingusitic information. However, this concern—which also arises for metalinguistic accounts—ultimately reduces to a familiar phenomenon: under anti-contextualist anti-Fregean assumptions, any Frege puzzle involves complete agreement. The case of FIDs is no different. Of course, like verbal disputes, FIDs can raise dialectical difficulties. For instance, the dispute over whether whales are fish cannot be resolved without first settling what it is to be a fish. However, this does not mean that such disputes are trivial

[20] It is worth noting that I have here framed the issue of complete agreement in *semantic* terms, treating the objects of agreement and disagreement as propositions. This is in line with the common assumption that questions—the natural candidates for the objects of agreement and disagreement—are sets of propositions. Still, the issue of complete agreement is sometimes formulated in terms of facts rather than propositions. On this formulation, the worry is that, if being F *just is* being G, then o's being F just is the same fact as o's being G (see, e.g., Rayo 2013). Thus, if the disputants agree that o is G, they may be said to agree on all the facts. Addressing this version of the problem would require us to consider competing accounts of facts, rather than the semantic frameworks discussed here.

or pointless. On the contrary, FIDs often reflect disagreement about "heavyweight" and theoretically significant issues—issues that cannot be resolved by consulting linguistic usage alone, but require substantive theoretical work.

CHAPTER 2

Is the Knowledge Argument a Frege Puzzle?

Frank Jackson's Knowledge Argument (KA) claims that Mary—a neuroscientist who knows all the physical facts about colour perception but has never seen colour—learns something new when she sees red, posing a challenge to physicalism. While physicalists deny that Mary acquires knowledge of new facts, they must still explain her apparent epistemic progress. I argue that the intuition that Mary gains new knowledge upon seeing red stems from the alleged opacity of propositional attitude ascriptions—the same phenomenon underlying Frege puzzles. First, I show how standard responses to KA parallel familiar solutions to Frege puzzles. Second, drawing on standard assumptions about the semantics of "know" followed by wh-clauses, I argue that Mary's apparent new knowledge of what it's like to see red introduces a Frege puzzle. Third, I address the objection that, unlike KA, Frege puzzles involve new knowledge of contingent propositions, arguing that both cases ultimately turn on principles of epistemic closure and the semantics of attitude ascriptions. Finally, I respond to the worry that framing KA as a Frege puzzle either introduces non-physical properties or fails to account for the "substantiveness" of Mary's new knowledge. I argue that it is unclear whether this challenge can be formulated without begging the question against physicalism and that, under a sufficiently broad understanding of phenomenal modes of presentation, physicalists can adequately respond.

§ Introduction

In Frank Jackson's well-known thought experiment (1986), Mary is a brilliant neuroscientist, who knows every physical fact about colour perception—for example, facts about the functioning of the receptors and neurons involved in colour vision, as well as the whole network of causal relations between processes underlying colour vision, external stimuli

and behavior—but has spent her entire life confined to a black-and-white room. According to Jackson's Knowledge Argument (KA), when Mary finally leaves the room and sees red for the first time, she learns something new about colour perception. This seems to entail that there are facts about colour perception that she did not know before. Yet, by hypothesis, Mary already knew all the *physical* facts about colour perception. Thus, Jackson concludes, there must be *non-physical* facts about colour perception that Mary comes to know upon leaving the room—hence, physicalism is false.

Physicalists reject the idea that Mary acquires knowledge of a new fact upon leaving the room, but they acknowledge that KA presents an epistemic challenge: it highlights the need for physicalism to account for Mary's apparent epistemic progress within a physicalist framework.

One prominent physicalist response is the Phenomenal Concept Strategy (PCS), which holds that, upon leaving the room, Mary acquires new *phenomenal* concepts for properties she already knew under physical concepts (Horgan 1984, Loar 1997, Perry 2001, Levine 2007, Papineau 2007). On this view, when Mary sees something red for the first time, she comes to know an old fact—one that she already knew in the room—under new concepts. This account aligns with the view I will defend—namely that, from a physicalist standpoint, the Knowledge Argument is an instance of Frege's puzzle (1892). However, for reasons that will become clearer in the final section, I will frame the view in terms of modes of presentation, rather than concepts. I will defend this account by examining the structural symmetry between KA and a paradigmatic Frege case. I will argue that the intuition that Mary learns something new upon leaving her room does not stem from anything unique to phenomenal consciousness, but from the same phenomenon underlying Frege puzzles: the apparent opacity of propositional attitude ascriptions. That is, within a physicalist

framework, whether Mary gains new knowledge ultimately turns on the semantics of attitude ascriptions.

The structure of the chapter is as follows. Following Nida-Rümelin (1995) and Stalnaker (2008), I distinguish three stages in Mary's (putative) epistemic progress: t1, when Mary is in the black-and-white room; t2, when she first sees red; and t3, when she is in a position to describe her experience *as* an experience of red (§1). I argue that the intuition that Mary gains new knowledge at t3, like the familiar intuition that one gains new knowledge upon learning that Hesperus is Phosphorus under an informative guise, is due to the apparent opacity of knowledge ascriptions. The available accounts of Mary's knowledge at t3 parallel those offered in standard Frege puzzles.

It may be objected, however, that Mary's real epistemic progress takes place at t2, when she comes to know *what it's like* to see red. In §2, I examine several accounts of knowing what it's like—including Lewis' ability hypothesis (1990) and acquaintance accounts (Conee 1994). I argue that the semantics of "know" followed by wh-clauses supports the view that knowing what it's like involves propositional knowledge. Once again, whether Mary gains new knowledge at t2 depends on which semantic framework and account of propositional attitude ascriptions are correct. In other words, Mary's knowledge of what it's like to see red introduces a new Frege puzzle.

In §3, I then address the objection that, in paradigmatic Frege puzzles, one comes to know a further contingent proposition when grasping the relevant identity claim under an informative guise. For instance, only when one comes to know the informative claim "Hesperus is Phosphorus" does one learn that there is something which is both the brightest planet in the evening and the brightest planet in the morning. According to the objection, it is unclear whether anything analogous applies to KA. In response, I argue that *in both cases*,

whether new knowledge of contingent propositions is gained depends on which semantic framework is correct and whether certain principles of epistemic closure hold.

Finally, in §4, I discuss a challenge to physicalist accounts of KA as a Frege puzzle, namely that either phenomenal modes of presentation introduce new *non-physical* properties or Mary's new phenomenal knowledge seems insufficiently "substantial". I argue that it is unclear whether this argument can be formulated without begging the question against physicalism and that, in any case, under a sufficiently broad understanding of phenomenal modes of presentation, a posteriori physicalism can respond to the challenge.

In short, what needs to be established is whether there is anything epistemically exceptional about consciousness which gives rise to the intuition that Mary learns something new, or whether this is simply the familiar intuition—present in any Frege puzzle for knowledge ascriptions—that new modes of presentation yield new knowledge.

Frege puzzles typically arise when two co-referential terms pick out the same entity via distinct modes of presentation. To sustain the analogy between KA and Frege puzzles, then, it is natural to focus on *reductive* physicalist views—according to which phenomenal properties are *identical* with physical properties, and the relevant phenomenal and physical terms are therefore co-referential. Reductive physicalism was, after all, the original target of KA, whereas non-reductive physicalism is often regarded as too close in spirit to anti-physicalism to qualify as a genuinely physicalist theory (Kim 1989, Melnyk 2008, Nida-Rümelin & O'Conaill forthcoming).

§1 The Knowledge Argument and Frege Puzzles

§1.1 Mary and Jane

Following Stalnaker (2008) and Nida-Rümelin (1995), Mary's apparent epistemic progress can be conceptualized across three distinct stages:[21]

(t1) At an initial time t1, Mary is inside her black and white room.

At this stage, although she has no direct acquaintance with it, Mary knows that there is a colour—red—that corresponds to a specific wavelength, is more similar to orange than to green, and so on. Given her extensive knowledge of colour perception, Mary also knows that a certain physical property—which she labels "ph-red"—is the phenomenal character of visual experiences of red, their "felt quality" as perceived by a normal visual system under normal lighting conditions.[22] Of course, Mary has never experienced this quality herself, but she possesses theoretical knowledge of it.

(t2) At a later time t2, Mary leaves her room and, for the first time, sees a red object.

However, Mary is not told that the object is red and, let us assume, the object is not a paradigmatically red one (such as a tomato or a fire extinguisher). Thus, Mary is not in a

[21] In Stalnaker's version, at t1 Mary is inside her room and is told that she will be shown either a red object or a green object. She creates labels for the phenomenal character of experiences of red ("phen-red") and green ("phen-green"). At t2 she is then shown a red object and creates a label for the phenomenal character of her experience ("wow"). According to Stalnaker, Mary cannot, at t2, rule out the possibility that she has been shown a green object—that is, that wow is phen-green, rather than phen-red—until she is told, at t3, that the object was red.

[22] I will remain neutral on the nature of the property ph-red—so long as it is something that Mary can have knowledge of from within her room—that is, so long as it can be described in physical terms (e.g. using the vocabulary of neurophysiology). For all that is assumed here, ph-red may, for instance, be a representational property of visual experiences of red (the property of such experiences of representing something as red). This account is thus compatible with strong representationalism—the view that the phenomenal character of experience is reducible to its representational properties.

position, at t2, to apply the term "red" to the object in question. Yet, upon seeing the red object, Mary arguably comes to know *what it's like* to see red. As Stalnaker (2008) points out, Mary can refer demonstratively to the phenomenal character of her experience, but her knowledge is not *essentially* demonstrative. Following Stalnaker, let us suppose that Mary coins the term "wow" for the phenomenal character of the experience she has at t2.

(t3) At t3, Mary is told, in these words, that the experience she had at t2 was an experience of red, and that wow *is* ph-red.

Reductive physicalists hold that wow and ph-red are the same physical property; hence, Mary's learning at t3 does not involve the discovery of any new non-physical property. The question is whether Mary's knowledge at t1—that the phenomenal character of experiences of red is ph-red—and her knowledge at t3—that the phenomenal character of experiences of red is wow—constitute the same piece of knowledge. Does Mary learn anything new at t3? As we shall see, the answer to this question depends entirely on which semantic framework for propositional attitude ascriptions is correct. The intuition that Mary learns something new at t3 stems from the (apparent) opacity of propositional attitude ascriptions—the phenomenon at the heart of Frege puzzles for attitude ascriptions.

Paradigmatic Frege puzzles (Frege 1892) for attitude ascriptions involve co-referential rigid designators—such as "Hesperus" and "Phosphorus" (now both known to refer to Venus)—and attitude reports, such as:

(1) Jane knows that Hesperus is the brightest planet in the evening.

To preserve compositionality, it must be possible to substitute terms with the same semantic value *salva veritate*—that is, preserving the truth value of sentences in which they occur in any non-quotational context. Thus, if "Hesperus" and "Phosphorus" have the same semantic value, they should make the same semantic contribution to the relevant sentences. This means that sentences (1) and

(2) Jane knows that Phosphorus is the brightest planet in the evening

cannot differ in truth-value. The puzzle arises from the fact that, intuitively, Jane may know that Hesperus is the brightest planet in the evening without knowing that Phosphorus is, since she might not know that Hesperus is Phosphorus. The difficulty in fact extends even to the identity claims themselves: by the same principles, it should be impossible for

(3) Jane knows that Hesperus is Hesperus

and

(4) Jane knows that Hesperus is Phosphorus

to differ in truth-value. Yet, while Jane clearly knows that Hesperus is Hesperus, intuitively she may well fail to know that Hesperus is Phosphorus. Indeed, propositional attitude ascriptions—such as knowledge ascriptions—are often taken to generate allegedly "opaque" contexts, namely linguistic contexts where substituting co-referential expressions appears to alter the truth-value of the relevant sentences. We can compare the above three-step description of Mary's epistemic situation to Jane's:

(t1) At t1, Jane knows that Hesperus is the brightest planet in the evening.

(t2) At a later time t2, Jane learns that Phosphorus is the brightest planet in the morning.

(t3) At t3, Jane is told, in these words, that Hesperus is Phosphorus, and thus learns that Phosphorus is the brightest planet in the evening.

Does Jane learn anything new at t3? In both Mary's and Jane's cases, the objects or properties known at t1 under a certain mode of presentation are known at t3 under a different mode of presentation. In Mary's case, this parallels the PCS claim that she comes to know "old" physical properties—properties that she already knew from within the black-and-white room—under new concepts. However, the view that Mary acquires new concepts or modes of presentation does not, by itself, establish whether she acquires new knowledge. As with any Frege puzzle, this depends on which semantic framework and account of propositional attitude ascription turn out to be correct.

§1.2 Three Strategies

In examining whether Mary acquires new knowledge at t3, one immediately sees that the available physicalist accounts of her epistemic situation closely parallel familiar approaches to Frege puzzles. Many physicalists find it hard to deny that Mary gains new factual knowledge upon release and are thus drawn to what is known as the "New Knowledge/Old Fact View".[23]

[23] Versions of this view are defended by Horgan (1984), Loar (1997), Tye (2000), Perry (2001), Papineau (2007).

New Knowledge/Old Facts (Fregean version)

On one version of this view, although Mary does not come to know any new facts at t3, she acquires knowledge of new propositions. Facts are typically understood either as sets of possible worlds or as "structured" entities individuated by the properties, relations, and objects they are about. On either conception, if ph-red and wow are the same property, then the fact that ph-red is wow is identical to the fact that ph-red is ph-red, and the fact that ph-red is the phenomenal character of experiences of red is identical to the fact that wow is the phenomenal character of experiences of red. Thus, under this view, Mary does not acquire knowledge of any new facts at t3.

What she does acquire, however, is knowledge of new propositions—namely, that wow is ph-red, and that wow is the phenomenal character of experiences of red. Propositions, on this view, are thus more *fine-grained* than facts: although the sentences

(5) ph-red is the phenomenal character of experiences of red

and

(6) wow is the phenomenal character of experiences of red

describe the same fact, they express different propositions. This is because the co-referential terms "ph-red" and "wow" pick out the phenomenal character of experiences of red via distinct *modes of presentation*. On this account, modes of presentation play a semantic role such that two co-referential terms that pick out their referent via distinct modes of presentation can make different semantic contributions to the relevant sentence and thus

cannot always be substituted *salva veritate.* Consequently, one can coherently hold a propositional attitude (e.g. knowledge) towards (5) without holding the same attitude towards (6), as these express distinct fine-grained propositions.[24] Mary thus gains knowledge of a new fine-grained proposition at t3, although she does not discover any genuinely new possibility that "narrows down" the set of worlds consistent with what she knows.[25]

This approach is essentially the classical Fregean solution to Frege puzzles, on which "Hesperus" and "Phosphorus" are co-referential but differ in *sense* (Frege 1892). Accordingly,

(7) Hesperus is the brightest planet in the evening

and

(8) Phosphorus is the brightest planet in the evening

do not express the same proposition, despite being true at exactly the same possible worlds and being about the same objects and properties—though of course Frege himself did not think in terms of possible worlds.[26] Thus, when Jane learns at t3 that Phosphorus is the

[24] A version of this view can be developed in a two-dimensional framework which distinguishes between a coarse-grained "broad content" of a sentence (e.g. a set of possible worlds), and a more fine-grained "narrow content". This is typically combined with the idea that propositional attitude ascription is sensitive to narrow content (e.g. Chalmers 2011b).

[25] Tye (2010) calls this the "non-modal" conception of knowledge, which merely involves "coming to think new thoughts", as opposed to eliminating genuine possibilities.

[26] On Frege's original view, propositional attitude verbs create opaque contexts in which terms refer to their usual senses, rather than to their usual referents, so that the terms in question do not even count as co-referential in these contexts.

brightest planet in the evening, she acquires knowledge of a new fine-grained proposition (see Forbes 1990, Chalmers 2011b for contemporary Fregean accounts).

New Knowledge/Old Facts (contextualist version)

Like the Fregean version, this variant of the New Knowledge/Old Facts view holds that Mary does not know at t1 that wow is the phenomenal character of experiences of red, though she does know that ph-red is. However, on this view, this is not because "Wow is the phenomenal character of experiences of red" and "Ph-red is the phenomenal character of experiences of red" express different (fine-grained) propositions.

Rather, on this view propositions are coarse-grained—either Russellian structured propositions whose constituents are the referents of the terms involved or sets of possible worlds—but ascriptions of propositional attitudes are sensitive to contextually relevant modes of presentation. In other words, a difference in mode of presentation does not entail a difference in proposition: if "ph-red" and "wow" are co-referential, (5) and (6) express the same proposition. This makes the view *anti-Fregean*. Nonetheless, according to this view, it is possible for someone to know that ph-red is the phenomenal character of experiences of red without knowing that wow is. This is due to the context-sensitivity of knowledge ascriptions: for Mary to be ascribed knowledge that wow is the phenomenal character of experiences of red, she must entertain that proposition under a contextually relevant mode of presentation—in this case, her newly acquired "wow" mode—and she can only do so at t3.

Thus, although Mary does not gain knowledge of a new proposition at t3, she still acquires new propositional knowledge, which could not be attributed to her at t1. This view can be seen as an alternative version of the New Knowledge/Old Facts view precisely because it entails that Mary gains new knowledge without discovering any genuinely new

possibility (or fact) at t3. Many proposals within the New Knowledge/Old Facts cluster fail to distinguish between the Fregean and contextualist versions of the view, as they do not clarify whether Mary gains new knowledge because she acquires knowledge of a *new* (fine-grained) proposition (but no new possibility), or because she gains *new knowledge* of an old (coarse-grained) proposition.

The contextualist strategy (e.g. Crimmins & Perry 1989, Richard 1990, Schiffer 1992) is a popular approach to Frege puzzles among proponents of anti-Fregean semantic frameworks, such as Russellianism—on which propositions are structured entities composed of the very objects and relations to which the terms involved refer—and intensionalism—on which propositions are sets of possible worlds. On these views, (7) and (8) express the same proposition. In the Russellian framework, they express the same proposition because the semantic values of the terms in both sentences are the same objects and properties—namely, Venus, and the property of being the brightest planet in the evening. In an intensional framework, they express the same proposition because the set of worlds where Hesperus is the brightest planet in the evening *just is* the set of worlds where Phosphorus is the brightest planet in the evening, given that Hesperus is Phosphorus in every possible world. However, since knowledge ascriptions are context-sensitive, at t1 we can ascribe to Jane knowledge that Hesperus is the brightest planet in the evening, but not that Phosphorus is the brightest planet in the evening. Only at t3, when she finally entertains the proposition under the contextually relevant "Phosphorus" mode, can Jane be ascribed knowledge that Phosphorus is the brightest planet in the evening.

Thus, on this account, no new proposition is learned in either case—the same coarse-grained proposition is involved at t1 and t3—but both Mary and Jane gain new propositional knowledge at t3.

No New Propositional Knowledge

On this view, Mary does not possess any propositional knowledge at t3 that she did not already possess at t1. This view combines a coarse-grained account of propositions with an anti-contextualist approach to knowledge ascription. Since "ph-red" and "wow" are co-referential, (5) and (6) express the same proposition. Moreover, since knowledge ascriptions are *not* sensitive to contextually relevant modes of presentation, Mary's knowing that ph-red is the phenomenal character of experiences of red entails that she knows that wow is the phenomenal character of experiences of red. Thus, Mary gains neither new knowledge of an old (coarse-grained) proposition nor knowledge of a new (fine-grained) proposition at t3: she already knew, at t1, that wow is the phenomenal character of experiences of red.

Anti-contextualist anti-Fregean solutions to Frege puzzles are put forward by, among others, Salmon (1986) and Williamson (2021a): if at t1 Jane knows that Hesperus is the brightest planet in the evening, and the proposition that Hesperus is the brightest planet in the evening *just is* the proposition that Phosphorus is the brightest planet in the evening, then at t1Jane also knows that Phosphorus is the brightest planet in the evening. Similarly, insofar as she can be ascribed knowledge that Hesperus is Hesperus, Jane can also be ascribed knowledge that Hesperus is Phosphorus, since these propositions are in fact identical. Thus, this account entails that in neither Jane's nor Mary's case is new propositional knowledge acquired at t3.

So far, Mary's and Jane's cases appear symmetrical: whether new knowledge is gained at t3 seems to depend entirely on the semantics of knowledge ascriptions. And, in both cases, the intuition that new knowledge is gained seems to be generated by the (alleged) opacity of propositional attitude ascriptions. Therefore, so far there appears to be no reason to believe that the intuition that Mary gains new knowledge at t3 is due to anything unique to phenomenal consciousness.

§2 Knowing What it's Like

It may be objected that Mary's actual progress takes place at t2, when she comes to know *what it's like* to see red—even though she is still not in a position to describe her experience in those terms. One thing to note is that, unlike her progress at t3, Mary's progress at t2 does not appear to involve language at all. At t3, Mary is in a position to assent to statements like "Experiences of red are wow" and "Ph-red is wow", and can thus be uncontroversially credited with knowledge of the propositions that experiences of red are wow and that ph-red is wow. By contrast, her progress at t2 arguably has nothing to do with what sentences she would assent to, or which terms she would use to refer to the properties of her experience. The linguistic issue might arise when we *describe* the situation using demonstrative or non-demonstrative terms (e.g. "wow"), co-referential with "ph-red", to refer to the phenomenal character of Mary's experience. But the point of KA seems to be that Mary learns something new when she *sees* red—something that is, prima facie, unrelated to anything linguistic. We should thus focus on what, if anything, Mary learns at t2.

Accounts of knowing what it's like have been formulated in terms of knowledge-how (e.g. Lewis 1990, Nemirow 1990) and acquaintance knowledge (Conee 1994). The following sections will examine these proposals.

§2.1 Knowing How and the Ability Hypothesis

I will start by focusing on a widely discussed account of knowing what it's like, proposed by David Lewis (1990), before drawing more general conclusions based on considerations about the semantics of knowledge ascriptions involving wh-clauses. According to Lewis'

so-called "ability hypothesis", Mary gains no propositional knowledge upon seeing red for the first time, but only knowledge-how. At t2, she comes to know, for instance, how to imagine, remember, and recognize experiences of red—though she need not be in a position to say to herself "It's an experience of red that I'm now able to imagine". In Lewis' words:

> After you taste Vegemite, and you learn what it's like, you can afterward remember the experience you had. By remembering how it once was, you can afterward imagine such an experience. […] Further, you gain the ability to recognize the same experience if it comes again. If you taste Vegemite on another day, you will probably know that you have met the taste once before.
>
> Lewis (1990, p. 98)

Similarly, Nemirow claims that "knowing what an experience is like is the same as knowing how to imagine having the experience" (1990, 495). Against the ability hypothesis, Jason Stanley & Timothy Williamson (2001)—henceforth S&W—argue that knowing-how is a form of knowing-that. On their view, knowing how to imagine experiences of red amounts to knowing a proposition of the form "w is a way for me to imagine an experience of red", entertained under a guise involving a practical mode of presentation of a way.[27] This is, first of all, because ascriptions of knowledge-how—like ascriptions of knowledge-why, knowledge-what, and so on—contain embedded questions, which are standardly interpreted as indicating propositional knowledge (more on this in §2.2).

[27] See Pavese (2020) for an account of practical modes of presentation, on which representing an aspect of the world under a practical mode of presentation means representing it in a way that is a function of our practical abilities, in a sense to be made precise.

Second, S&W note, cross-linguistic evidence suggests that "know" is ambiguous between two readings in English, which are captured by distinct verbs in languages such as German or Italian. For example, the Italian verb "sapere" and the German verb "wissen" express propositional knowledge, whereas the verbs "conoscere" and "kennen" express non-propositional (e.g. objectual) knowledge. Crucially, "knowing how" is translated in these languages using the verb for propositional knowledge, supporting an interpretation of knowledge-how as a species of knowledge-that. Thus, according to S&W, Lewis' claim that Mary only gains new knowledge-how but no knowledge-that at t2 is inconsistent: by gaining knowledge-how, they claim, Mary gains new propositional knowledge. Yet, the issue of whether, by acquiring know-how, Mary gains *new* propositional knowledge at t2 might, in fact, just amount to a new Frege puzzle. As I argue below, whether Mary can be said to gain new propositional knowledge at t2 depends, once again, on the semantics of propositional attitude ascriptions.

S&W consider a possible response on Lewis' behalf: that Mary already knows how to imagine experiences of red at t1, but only acquires the ability to employ that knowledge at t2. However, they reject this idea:

> Our problem with this response is straightforward. It seems absurd to countenance the truth of:
>
> (54) Mary knows how to imagine an experience of red.
>
> with respect to the situation in which Mary is in her black and white room. If she knows how to imagine an experience of red, why is she unable to imagine such an experience? Evidence for the robustness of the intuition that Mary does not know how to imagine an experience of red is the fact that, throughout the

> literature, the falsity of (54) with respect to the envisaged situation is assumed. Therefore, the ability account of Jackson's knowledge argument fails to show that Mary does not acquire propositional knowledge that she did not previously possess upon leaving her black and white room. Indeed, assuming that Mary did not possess the requisite knowledge-how already, the ability account in fact entails that Mary acquires propositional knowledge upon leaving her black and white room.
>
> (Stanley & Williamson 2001, p. 443)

As S&W note, it seems absurd to claim that Mary knows how to imagine experiences of red at t1, while she is still in her black-and-white room—if she really knows how, then why is she unable to do so? In general, knowing how to φ does not entail being able to φ: a pianist who loses her arms may still know how to play the piano, even if she is no longer able to. But in Mary's specific case, it is unclear what could prevent her from having the relevant ability, if she really has the knowledge-how. As S&W suggest, the best explanation for her being unable to imagine experiences of red is that she does not, in fact, know how to do so. Thus, it is plausible to think that Mary does not know how to imagine experiences of red at t1, and only learns how to do so at t2.

Still, even granting S&W's intellectualist account of knowledge-how, it is not clear that Mary's acquisition of knowledge-how at t2 entails that she acquires *new* propositional knowledge at that time (Cath 2009). On the intellectualist account, knowing how to imagine an experience of red entails knowing, for some w, that w is a way for someone to imagine an experience of red, where w is grasped under a *practical* mode of presentation. Of course, at t1 Mary does not know of any w that it is a way to imagine an experience of red *under a practical mode of presentation*—which is why she does not *know how* to imagine experiences of red. This also explains why she is not *able to* imagine experiences of red:

although there is no general entailment from being able to φ to knowing how to φ, S&W concede that such an entailment holds in the case of intentional actions, such as imagining that something is the case. In such cases, not knowing how to φ plausibly entails not being able to φ. Nonetheless, in virtue of her extensive knowledge of neurophysiology, Mary knows precisely, at t1, which parts of the brains and which physical mechanisms are involved in imagining an experience of red. Thus, arguably, she already knows at t1, of some w, that w is a way to imagine an experience of red under a *non-practical* mode of presentation. This will plausibly involve some description "P" of whatever is occurring at the neurophysiological and cognitive level when someone imagines an experience of red. Mary's complete physical knowledge surely enables her to know, at t1, that being in the P-state is a way for someone to imagine an experience of red. At t2, she merely comes to know that being in the P-state is a way for someone to imagine an experience of red under a *practical* mode of presentation, thereby learning how to imagine such an experience, and gaining the corresponding ability.

In sum, at t1 Mary lacks the ability to imagine experiences of red and does not know *how* to imagine experiences of red, but might nonetheless already know the relevant proposition—namely that being in the P-state is a way (for someone) to imagine such experiences. If S&W's intellectualist account is correct, then at t2 Mary comes to know that very proposition under a practical mode of presentation, and thereby acquires the knowledge of how to imagine experiences of red. However, this does not, by itself, entail that Mary acquires knowledge of any *new* proposition. Of course, a fine-grained semantics—on which a difference in modes of presentation corresponds to a difference in propositions—does entail that Mary comes to know a new proposition at t2. But this is, once again, simply the standard Fregean response to Frege puzzles. Given intellectualism about knowledge-how, whether Mary can be said to gain new propositional knowledge at t2 depends entirely on

which view of semantics and propositional attitude ascription—i.e. which of the three strategies considered in §1.2—is ultimately correct.

§2.2 The Semantics of "Knowing What"

Leaving aside intellectualism about knowledge-how, general considerations about the semantics of knowledge ascriptions involving wh-clauses support the idea that "knowing what it's like" expresses propositional knowledge. These considerations closely parallel those advanced by S&W in defense of the view that "knowing how" expresses propositional knowledge. To begin with, just like the sentence "Mary knows how to imagine an experience of red", the sentence "Mary knows what it's like to see red" contains an embedded question, which is standardly taken to indicate propositional knowledge. As Lycan (1996) observes, indirect-question clauses following attitude ascriptions are closely related to that-clauses, both in meaning and grammatically. The schema "S knows wh-" is related to "S knows that…". For instance, "S knows *where* X φs" is true in virtue of S's knowing *that* X φs at p, where "p" names some place; "S knows *when* X φs" is true in virtue of S's knowing *that* X φs at t, where "t" names some time; and so on. By analogy, "S knows what it's like to see red" means, roughly, "S knows that it is (like) Q to see red", where "Q" names a relevant property. As mentioned, at t2, Mary is not in a position to describe her experience as an experience of red. The mode of presentation associated with "see red" in the proposition that it is like Q to see red—the proposition she entertains at t2—might, for instance, be a demonstrative one ("to see *this* colour"). The same applies to her knowledge, at t2, that w is a way for someone to imagine an experience of red.

Stoljar (2016) notes that "what it's like" questions are closely related to "how" questions: "How does it feel to be one of the beautiful people?" is a close variant on "What is it like to be one of the beautiful people?". Most "what is it like" sentences can naturally

be recast as "how" questions. Stoljar argues that, just as "know where" quantifies over places, and "know when" quantifies over times, "know how" quantifies over ways, where a way is either a way a thing is or a way to do something (Stoljar 2016). Drawing on S&W's analysis of sentences like "Carla knows how to ride a bike" in terms of "There is some way such that Carla knows that that way is a way for her to ride a bike", Stoljar claims that sentences like "Dennis knows how Stalin was to his generals" are plausibly analysed as "There is some way such that Dennis knows that that way is the way Stalin was to his generals". This, he suggests, supports an analogous treatment of "knowing what it is like to φ" as knowing, of some way, that that is the way it is like to φ. He writes: "A sentence like 'There is something it is like to have a toothache' has schematically the form 'There is a way x's c-ing is to y'. […] On this treatment, 'John knows what it is like to have a toothache' is plausibly analysed (again, to a first approximation) as 'There is some way such that John knows that it is that way to have a toothache'" (2016, pp. 1165-1171).

Stoljar claims that his account is superior to the property account sketched earlier—on which "There is something it is like to φ" is true if and only if there is some property F such that φ-ing is F. According to Stoljar, this account fails because its right-to-left direction is too weak: an event of φ-ing may have a property without there being something it is like to φ. For example, the event of going to the dentist might have the property of occurring next Tuesday, but there need be nothing it is like to go to the dentist—and one does not count as knowing what it is like merely by knowing that it will happen next Tuesday. Thus, the property account plausibly requires a restriction of the relevant properties to *experiential* properties—such as the phenomenal character of the experience in question. However, it is worth noting that Stoljar's own account arguably requires a similar restriction of the relevant ways to experiential ones, those that can affect the subject experientially (as he himself acknowledges, p. 1176).

Returning to the semantics of "knowing what it's like", the ambiguity of the English verb "know", highlighted by S&W in defense of intellectualism, further supports a propositional reading of such expressions. Like "knowing how", "knowing what it's like" is naturally translated using the verb that expresses propositional knowledge in languages that employ distinct verbs for propositional and non-propositional knowledge. For example, "Mary knows what it's like to see red" is rendered in Italian as "Mary *sa* com'è vedere il rosso", and in German as "Mary *weiß* wie es sich anfühlt rot zu sehen".

Of course, these types of linguistic arguments raise methodological concerns. Even assuming that these semantics considerations are correct, one might worry that the languages in question are misleading as to what actually makes "knowing what" sentences true. Perhaps in a "metaphysically perspicuous" language, one would express things differently. Still, the fact that the languages in question consistently work this way, I believe, should at least make the propositional reading the default interpretation. If wh-clauses in English are indeed ambiguous between a propositional and a non-propositional reading, consulting languages that make an explicit lexical distinction between the two types of knowledge seems a legitimate way to adjudicate the matter—at least in the absence of defeating considerations. It is up to the opponent, then, to explain why this seemingly plausible interpretation should be rejected.

Analogous considerations apply to acquaintance or objectual accounts of knowing what it's like (e.g. Conee 1994). On these accounts, Mary does not acquire any new propositional knowledge at t2, but merely acquaintance, or objectual, knowledge of the phenomenal character of experiences of red (ph-red), which she previously knew only "by description".[28] Yet, once again, this seems to presuppose an *ad hoc* reading of "knowing

[28] The standard view is that propositional knowledge is not knowledge-of. However, the relation between the two is not completely uncontroversial—e.g. see Moss (2025).

what it's like", one that deviates from the standard interpretation of "know" followed by wh-clauses as expressing propositional knowledge.

Stoljar (2016, 2017) discusses one such "non-interrogative" reading of wh-constructs, on which these do not express propositional knowledge. On the interrogative reading, "knowing what it's like to see red" expresses propositional knowledge—i.e. knowledge *that* seeing red is like such and such. On what Stoljar calls the "free relative" reading, by contrast, knowing what it's like to see red amounts to knowing the denotation of "what it's like to see red", just as "Jane loves where the conference is" means that Jane loves the place denoted by "where the conference is". On this reading, "knowing what it's like" expresses non-propositional, objectual knowledge. As Stoljar notes, this reading is not available for all verbs. For example, "Jane wonders where the conference is" forces the interrogative reading. However, the verb "knows" permits both sorts of interpretations. For instance, a sentence like "Paul does not know who Sebastian loves" can be considered as ambiguous between the two readings. Suppose Paul knows Ann, and Ann is the person Sebastian loves. On the interrogative reading, Paul may not know who Sebastian loves, even if he knows Ann, if he does not know *that* Sebastian loves Ann. On the free relative reading, by contrast, if Paul knows Ann, then he does know who Sebastian loves—though this reading may sound unnatural, in part because "who" must be understood as elliptical for "the person who[m]".[29]

While this may show that knowledge ascriptions involving wh-clauses do not *always* ascribe propositional knowledge, the view that "knowing what it's like" expresses objectual knowledge remains vulnerable to the objection concerning the ambiguity of "know" in English. Consider again the example above. The objectual (free relative) reading—on which, if Paul knows Ann, then he also knows who Sebastian loves—would be rendered in

[29] This example is taken from Tye (2010).

languages such as Italian and German using the verbs that express objectual knowledge ("conoscere" and "kennen"). By contrast, the propositional (interrogative) reading—on which Paul does not know who Sebastian loves unless he knows *that* Sebastian loves Ann—would be rendered using the verbs that express propositional knowledge ("sapere" and "wissen"). In short, when a wh-clause following "know" behaves like a noun denoting a place, person, or thing, the relevant occurrence of "know" is naturally translated with the verb for objectual knowledge in languages that make the relevant distinction. Yet, as previously noted, "knowing what it's like" is consistently translated with the verb for propositional knowledge in such languages.

Even if the objectual reading were correct, what Mary acquires at t2 would be knowledge of the denotation of "what it is like to see red"—namely the phenomenal character of visual experiences of red. But Mary already knew this property under a physical description at t1. Thus, she would not gain knowledge of a new property, but just new knowledge, by acquaintance, of an "old" property. Yet, coming to know by acquaintance something one already knew by description amounts to coming to know it under a new (experiential, or phenomenal), mode of presentation. In much the same way, I could be plausibly said to know the Bodleian library after years of studying maps and photographs of it; when I finally see it, I simply come to know it under a new mode of presentation.

Note that the claim that knowing what it's like to see red involves propositional knowledge is entirely consistent with the—very plausible—view that at t2 Mary *also* gains acquaintance knowledge of the phenomenal character of experiences of red. In fact, the propositional interpretation of "knowing what it's like" plausibly *involves* acquaintance insofar as it involves phenomenal modes of presentation in its account of Mary's knowledge at t2 (more on that in §4.2).

In either case, it is plausible to account for Mary's epistemic situation in terms of a Frege puzzle. If what she gains at t2 is merely acquaintance knowledge of the phenomenal character of experiences of red, she can be said to acquire knowledge of something she already knew, though under a new mode of presentation. Whether this amounts to real epistemic progress depends on whether a difference in modes of presentation suffices to yield new knowledge. If, on the other hand, knowledge of what it's like to see red amounts to knowledge of a proposition of the form "It is Q to see red", or "w is the way it's like (for someone) to see red", under reductive physicalism "Q" and "w" must refer to physical properties, states, or processes, which Mary already associates with experiences of red at t1.[30] Once again, Mary's knowledge of what it's like to see red at t2 can be construed in one of three ways: as "old" knowledge of an "old" coarse-grained proposition that she already knew at t1; as new knowledge of an old coarse-grained proposition—new in virtue of the difference in mode of presentation; as new knowledge of a new fine-grained proposition. This depends entirely on which semantic framework for attitude ascriptions is correct. Thus, even if the core of Mary's epistemic progress indeed takes place at t2, KA remains structurally analogous to a Frege puzzle.

§2.3 A Note on Anti-Contextualist Anti-Fregeanism

Contextualist anti-Fregeans can accommodate the intuition that, at t1, Mary does not know what it is like to see red because, on their view, knowing what it's like to see red involves knowing the relevant proposition—that it is Q to see red, or that w is the way it is like to see red—under a *phenomenal* mode of presentation. On this account, Mary cannot be said to

[30] To be clear, my claim is that knowing what it's like to see red involves propositional knowledge *about* one's experience and its phenomenal character. To say that, by having the experience of red, Mary comes to know the proposition that w is the way it is like to see red is not to say that that proposition is the *content* of her experience.

know what it's like to see red at t1 because she does not know the relevant proposition under a phenomenal mode of presentation. This approach parallels S&W's intellectualist account of knowledge-how discussed earlier, according to which knowing how to φ is a matter of knowing, of some w, that w is a way for someone to φ under a *practical* mode of presentation of w (Stanley & Williamson 2001). On S&W's view, someone who knows that w is a way to φ under a non-practical mode of presentation of w cannot be said to know how to φ.

By contrast, anti-contextualist anti-Fregeanism entails that Mary *does* know what it's like to see red already at t1, while still confined to her black-and-white room. On this view, knowing the relevant coarse-grained proposition—that it is Q to see red, or that w is the way it is like to see red— under *any* mode of presentation suffices for ascribing to Mary knowledge of what it's like to see red. This, of course, will strike many as deeply counterintuitive. There are, however, some considerations that can be offered in defense of anti-contextualist anti-Fregeanism, which also serve to further develop the analogy between KA and Frege puzzles.

To begin with, anti-Fregean views have similarly counterintuitive consequences in standard Frege cases. For example, intensionalism entails that "The actual president of the US is the actual president of the US" and "Donald Trump is the actual president of the US" express the same proposition. This is due to the presence of the rigidifying operator "actual", which renders "the actual president of the US" a rigid designator, thereby picking out the actual president of the US (Trump) at every possible world. Combined with anti-contextualism about knowledge ascriptions, this entails that anyone who knows that the actual president of the US is the actual president of the US can be said to know that the actual president of the US is Donald Trump. This, in turn, may be taken to entail that whoever knows that the actual president of the US is the actual president of the US knows *who* the actual president of the US is.

Yet, knowledge-who, *in general*, seems to be highly context-sensitive (Boër and Lycan 1975). In most contexts we would not say that someone who merely knows that the actual president of the US is the actual president of the US knows who the actual president of the US is. The same, in fact, applies to all "know" + wh-clause phrases: even if you know that the place where I am currently located is the place where I am currently located, there remains an important sense in which you do not know where I am currently located. In this sense, the counterintuitive result that Mary knows what it's like to see red at t1 does not seem to pose any *special* epistemic problem unique to phenomenal consciousness: knowledge ascriptions involving coarse-grained propositions under non-informative guises often produce similar counterintuitive consequences.

In defense of anti-contextualist anti-Fregeans, it is worth noting that in contexts where the contrast between experience and theoretical knowledge is not salient, it might in fact be natural to say that at t1 Mary knows what it's like to see red. In such contexts, the question of whether Mary knows what it is like to see red may be plausibly interpreted as asking whether she has acquired the relevant information in physical terms (e.g. that P is the way it is like to see red, or ph-red is the phenomenal character of experiences of red), as opposed to some other piece of physically-described information (e.g. that some P* is the way it is like to see green, or that a property ph-green is the phenomenal character of experiences of green). For example, imagine Mary working alongside a team of scientists in the black-and-white room. During a discussion of her ongoing research, one of her colleagues asks whether she has figured out what it's like to see green. If Mary has just completed her study on the physical basis of red experiences—learning, for example, that ph-red is the phenomenal character of experiences of red—but has not yet begun investigating green, it seems natural for her to reply: "Not yet—I've only learned what it's like to see red so far".

Similar considerations apply to many classic Frege puzzles. Often, when we ask whether someone knows that *p*, we are not specifically concerned with whether they know that *p* under a particular mode of presentation. For example, suppose Alice and Bob both know their neighbour as "John", and are aware that he is a talented gardener. Their friend Carol, who has recently moved into the neighbourhood, also knows this person and his gardening skills—but only under the name "Mr. Smith". If Alice says to Bob "Carol knows that John is a talented gardener", her claim is intuitively true even though Carol does not know John under the name "John". In some contexts, such a statement might falsely implicate that Carol knows the proposition under the guise "John is a talented gardener". But in this context the ascription seems perfectly legitimate, regardless of whether Carol herself would use the name "John" to refer to Mr. Smith.

§3 The Asymmetry Objection

§3.1 Ruling Out Possibilities

In this section, I consider an objection to the claim that KA is symmetrical with paradigmatic Frege puzzles. Let us call this the "Asymmetry Objection" (AO). As discussed in previous sections, whether Jane and Mary can be said to learn something at t3 depends on the semantics of propositional attitude ascription. However, the objection goes, there is a genuine discovery that Jane makes at t3: she learns the contingent proposition that something is both the brightest planet in the morning and the brightest planet in the evening. That is, only at t3 is Jane able to rule out genuine possibilities (metaphysically possible worlds) in which the brightest planet in the evening and the brightest planet in the morning are distinct. The same structure, AO continues, holds for any Frege puzzle—for instance, only when Lois

Lane comes to know that Clark Kent is Superman under the informative guise "Clark Kent is Superman" does she learn the contingent proposition that someone is both a superhero and a journalist at the Daily Planet.

By contrast, according to AO, no similar contingent proposition is learned by Mary at t3. To see why, we should note that it is often assumed that phenomenal properties such as pain and colour sensations have their phenomenology necessarily—anything that does not feel like pain, for example, simply cannot be pain (Kripke 1980, Chalmers 1996, 2009. See Grahek 2011 for skepticism about this claim).[31] Similarly, anything that does not have a red phenomenology cannot count as an experience of red (relative to standard lighting conditions and perceptual systems). This means that having a red phenomenology is not a contingent feature of experiences of red: experiences of red have their phenomenal character necessarily.

According to AO, when Mary comes to know, at t3, the informative identity claim "Wow is ph-red", the relevant proposition in the vicinity is the proposition, *s,* that the wow-phenomenology is the phenomenal character of experiences of red (ph-red). But since the wow-phenomenology *is* the red phenomenology—and the red phenomenology is necessarily co-instantiated with experiences of red in the relevant conditions—*s* is not contingent. Unlike the proposition that the brightest planet in the morning is the brightest planet in the evening, which Jane learns at t3, s is necessarily true. Therefore, according to AO, there is no possibility that Mary is able to rule out at t3 which she was not able to rule out before, because there is no possible world where an experience of red lacks a red phenomenology altogether. Variants of AO underlie arguments against the idea that the relation between phenomenal properties and their physical correlates could be one of a posteriori identity (e.g.

[31] Some even argue that it is a *conceptual* truth about pain, for example, that it has the pain-phenomenology.

Kripke 1980; Chalmers 1996, 2009). A version of the Asymmetry Objection is summarized by Tye (2008) as follows:

> It is clear that the physicalist cannot allow that Mary, in coming to know something new when she sees red for the first time, is making a discovery of just the same sort as the discovery that Hesperus is Phosphorus or that Clark Kent is Superman. In these cases, in making the relevant discoveries, one also discovers that properties previously associated (some would say a priori associated) with the object or individual under one concept are associated with it or him under the other. Thus, in discovering that Hesperus is Phosphorus one comes to realize, among other things, that Phosphorus has the property of being the evening star. Before the discovery, one associated that property only with the heavenly body conceived of as Hesperus. Similarly, in discovering that Clark Kent is Superman, one finds out (among other things) that Superman has the property of being a mild-mannered newspaper reporter. Before the discovery, one associated that property only with the person conceived of as Clark Kent. The reason that Mary's discovery cannot be of this sort, at least if physicalism is true, is that it would then be necessary for Mary to associate with color experiences properties she did not associate with them before she left her room. But there are no such properties if physicalism is true, for Mary in her room knows all the physical facts pertaining to color experiences.
>
> Tye (2008, p.124-125)

Stalnaker (2008) appears to overlook this point in his discussion of the Knowledge Argument. In his version of Jackson's thought experiment, Mary at t1 introduces the names "phen-red" and "phen-green" to refer to the phenomenal characters of the experiences that a

normal observer would have upon seeing a red and a green object, respectively. She is then informed that she will be shown either a red or a green object at t2, depending on the outcome of a coin flip. Following the coin flip—whose result she does not know—Mary is shown a red object at t2 and labels the phenomenal character of her experience "wow". Stalnaker maintains that Mary still does not know, at t2, that phen-red is wow: all she knows at t2 is that either phen-red or phen-green is wow. On a standard account of coming to know that p as being able to rule out as counterfactual or "eliminate" possible worlds inconsistent with p—which Stalnaker himself endorses—this means that Mary is still unable, at t2, to rule out possible worlds where she has been shown a green object (i.e. where wow is phen-green). That is, such possible worlds are, according to Stalnaker, still compatible with her knowledge at t2. In other words, even after seeing the red object, according to Stalnaker, Mary is still unable to rule out as counterfactual the possibility that the object she was shown was green.

As Magidor (2010) points out, this is rather surprising, given that Stalnaker himself takes the connection between an experience and its phenomenology to be necessary. That is, in any world where Mary is shown a green object, she must have an experience with a green phenomenology—all worlds where Mary is shown a green object are worlds where she has an experience with a green phenomenology at t2. Stalnaker's view thus entails that, even after seeing the red object, Mary cannot rule out that her experience had a green phenomenology—an implication that appears deeply implausible. Even externalist arguments to the effect that we are not always in a position to know precisely which phenomenal experience we are having—such as Williamson's (2000) anti-luminosity argument—involve cases where the relevant phenomenal experiences are very similar. By contrast, Stalnaker's claim entails that Mary does not even know which of two *radically different* phenomenal experiences she had (Magidor 2010).

§3.2 A Reply to the Asymmetry Objection

AO claims that at t3 Jane learns that something is both the brightest planet in the evening and the brightest planet in the morning. However, it is worth specifying that whether this is correct depends on the semantic framework in play, as well as on the validity of certain epistemic closure principles. According to our account of Jane's case (§1.1), we have:

(A) At t2, Jane knows <Hesperus is the brightest planet in the evening and Phosphorus is the brightest planet in the morning>. Let us call this proposition '*p*'.

In an anti-contextualist anti-Fregean framework, (A) entails:

(B) At t2, Jane knows <Hesperus is the brightest planet in the evening and Hesperus is the brightest planet in the morning>. On an anti-Fregean view, this proposition is also *p*.

The question is whether (B) entails:

(C) At t2, Jane knows <There is an x such that x is the brightest planet in the evening and x is the brightest planet in the morning>. Let us call this proposition '*q*'.

Now, *q* is entailed by *p* but not identical to it—while every world in which Hesperus is the brightest planet in the evening and the brightest planet in the morning are worlds in which something is both the brightest planet in the evening and the brightest planet in the morning, the converse does not hold. Thus, the inference from (B) to (C) relies on the validity of

certain principles of epistemic closure. In particular, consider the following principle of closure under entailment (Luper 2020):[32]

(KE) If p entails q, S knows p, and S competently deduces q from p,[33] then S knows q.

If KE is satisfied, then it may be argued that (B) entails (C). In that case, Jane already knows *at t2* that something is both the brightest planet in the evening and the brightest planet in the morning, and so cannot be said to learn this proposition at t3, contrary to AO. A difficulty here is whether Jane can be taken to satisfy the competent deduction constraint. Given that she knows p under the guise "Hesperus is the brightest planet in the evening and Phosphorus is the brightest planet in the morning", one might argue that this prevents her from deducing q. At the same time, however, the competent deduction constraint is typically invoked in non-formal epistemology, whereas in simple models for formal epistemology, there is no requirement of competent deduction—the mere fact that p entails q suffices.

Crucially, though, we do not need to settle whether KE should include the competent deduction constraint—and, if so, whether this is satisfied in this case—in order to maintain the symmetry between KA and Frege puzzles. As I will explain, a parallel issue arguably arises in Mary's case. This brings us to a possible reply to AO—namely that, contrary to the objection's assumption, it is possible to construct analogous contingent propositions that Mary can be said to learn at t3. One of these may be, for instance, the proposition *r* that the phenomenal character of an experience of a colour called "red" is the phenomenal character

[32] The inference from (B) to (C) can be formulated in terms of other epistemic closure principles—such as the principle of closure under simplification, whereby K(*p*&*q*) entails K*p* and Kq (possibly, with an added competent deduction constraint). This is because, if p entails q, then p is intensionally equivalent to p&q. The question is then whether S's knowing p&q entails that S knows q.

[33] Formulated as "If p entails q, S knows p, and S knows that p entails q, then S knows q", this constraint threatens to generate an infinite regress of the kind discussed by Carroll (1895).

of the kind of experience Mary has at t2—that is, that there is something which is both the phenomenal character of an experience of a colour called "red" and the phenomenal character of the kind of experience Mary has at t2. Clearly, r is contingent: first, it is not necessarily the case that Mary has an experience of red at t2—there are possible worlds where she is shown a green object at t2; second, there are surely worlds where "red" refers to a different colour, or no colour at all.

The idea underlying AO seems to be that the relevant proposition in the vicinity of an informative identity claim is determined by the modes of presentation associated with the terms involved. But this, one might argue, is not sufficient to establish whether the modes of presentation associated with "wow" and "ph-red" yield r or s. Both "the phenomenal character of experiences of the colour called 'red'" and "the phenomenal character of the kind of experience I had at t2" may be apt verbalizations of what Mary cognitively associates with "wow" and "ph-red". We are, of course, assuming that modes of presentation do not, without rigidification, fix the reference of the terms in question across possible worlds. After all, "the brightest planet in the evening" and "the brightest planet in the morning" do not fix the reference of "Hesperus" and "Phosphorus" across possible worlds either—otherwise such terms would function as non-rigid designators, contrary to what is standardly assumed after Kripke (1980). Modes of presentation do, however, account for the cognitive significance of the terms in question. And both "the phenomenal character of experiences of a colour called 'red'" and "the phenomenal character of the experience Mary has at t2" are possible candidates for this role.

Therefore, in coming to know the informative identity claim "Wow is ph-red" Mary might come to know *r*. Since *r* is contingently true, there might be genuinely possible worlds that Mary is unable to rule out before t3—for instance, a world w where blue is called "red" and where Mary is shown a green object at t2. Just as in Jane's case, whether Mary is indeed

unable to rule out this possibility before t3 depends on which semantic and epistemological principles hold:

(A*) At t2, Mary knows <Ph-red is the phenomenal character of an experience of a colour called "red" and wow is the phenomenal character of the kind of experience Mary has at t2>

Given anti-contextualist anti-Fregeanism, (A*) entails:

(B*) At t2, Mary knows <Ph-red is the phenomenal character of an experience of a colour called "red" and ph-red is the phenomenal character of the kind of experience Mary has at t2>

If KE is satisfied, (B*) entails:

(C*) At t2, Mary knows <There is an x such that x is the phenomenal character of an experience of a colour called "red" and x is the phenomenal character of the kind of experience Mary has at t2>

Thus, just as in Jane's case, given anti-contextualist anti-Fregeanism and the relevant closure principle, Mary can, in fact, be said to know *r* (and thus to rule out w) already at t2.

At this point, however, we should pause to consider a further objection to this response to AO. Suppose, for instance, that Mary is *metalinguistically stunted*—that is, she is unable to form metalinguistic thoughts of the kind involved in entertaining r. Intuitively, we would still want to say that Mary learns something new at t3, even though she does not come to know any metalinguistic proposition. Of course, one might respond by appealing to

other, non-metalinguistic, propositions that Mary may be taken to learn at t3—for example, that ph-red is the phenomenal character of the experience she had at t2.

A more general objection, however, is that metalinguistic, time-relative, or indexical/demonstrative propositions are too "trivial" and fail to capture the apparent "robustness" of what Mary seems to learn. One might argue, then, that the Asymmetry Objection challenges us not simply to explain how Mary could learn some new proposition, but how she could learn a "substantial" or "significant" one. The contingent proposition Jane learns, the objection continues, is a substantive one; the contingent propositions we might attribute to Mary are not.

In response, however, it is worth noting that paradigmatic Frege puzzles do not necessarily involve the acquisition of new "substantial" or "robust" knowledge—whatever precisely those terms are taken to mean (more on that later). Indeed, the relevant modes of presentation in such cases can be rather "thin", and even merely metalinguistic. Suppose, for instance, that someone is looking at two objects in the distance and, after a while, realizes that they are in fact parts of one and the same object. At that point, she utters the informative identity claim "That is that". Here, the speaker might associate the two token demonstratives with rather "insubstantial" contents, such as "the thing I am pointing to", or "the thing I am looking at". This applies especially to cases in which the speaker is not sure what the referent of the demonstrative is, and is thus unable to associate it with any specific sortal. Another illustrative case: suppose I overhear someone mention a person named "James" but am not paying close attention. The cognitive content I associate with "James" might amount to no more than "the person called 'James'", or "the individual that guy just mentioned". Now suppose I later overhear the same speaker refer to someone named "Jimmy". Eventually, I might discover that James is Jimmy. What I thereby come to know in that scenario is hardly substantial: I may simply acquire metalinguistic knowledge that the person called "James"

is the person called "Jimmy", or that the individual mentioned earlier is the same one mentioned just now.

In this sense, the supposed lack of "substantial" (e.g. non-metalinguistic) knowledge in Mary's case should not be taken to make her case significantly different from standard Frege cases. Moreover, someone who thinks of Hesperus as *the actual evening star* and of Phosphorus as *the actual morning star* will not discover any substantial contingent proposition upon coming to know that Hesperus is Phosphorus. This is because the rigidifying operator "actual" renders the proposition that *the actual morning star is the actual evening star* a necessary one—and indeed one that, on certain coarse-grained accounts, is the very same proposition expressed by "Hesperus is Phosphorus". Arguably, this does not make the case any less a genuine instance of Frege's puzzle.

In sum, in both Mary's case and standard Frege cases, subjects may acquire merely metalinguistic or demonstrative knowledge by coming to know the relevant informative identity claims. Moreover, in both cases, the propositions thereby known can be necessary or contingent. In sum, the claim that the terms involved in standard Frege puzzles pick out their referent via contingent and substantial modes of presentation, whereas those in Mary's case do not, appears to be mistaken.

§4 The New Challenge

§4.1 Substantial Knowledge

It may be objected that the point concerning the substantiveness or robustness of Mary's new knowledge is not merely about the asymmetry with Frege puzzles, but has independent plausibility. In other words, regardless of the analogy between Mary's case and standard

Frege puzzles, it is independently implausible to claim that phenomenal modes of presentation are "thin"—for instance, demonstrative or metalinguistic.

Indeed, a "New Challenge" (Schroer 2010, Veillet 2015) has been raised against physicalist accounts of the Knowledge Argument that appeal to differences in concepts or modes of presentation. The challenge for physicalism is no longer just to explain how Mary could gain new knowledge when she leaves the room. Rather, the challenge is now to explain how Mary can come to acquire new knowledge that is "substantial", "rich", and "robust". Levine writes: "The first-person access we have to the properties of experience seems quite rich; we are afforded a very substantive and determinate conception of a reddish experience merely by having it" (2007, 163). Similarly, Levin claims that the goal is to explain "why the knowledge that Mary acquires when she leaves her black-and-white room seems so substantive", or what "seems to be the rich and robust knowledge of experience Mary gains when she leaves her black-and-white room" (2007, pp. 90-93). This, it is claimed, is what makes it implausible to suggest that Mary's new phenomenal mode of presentation of ph-red merely conveys metalinguistic information or functions as a bare demonstrative "pointer," devoid of further content.

The issue raised by the New Challenge against physicalist theories that appeal to new concepts or modes of presentation can be framed as follows: either a new phenomenal concept or mode of presentation—whether demonstrative or non-demonstrative—"brings with it" a new property (as part of its definition, reference-fixing material, or sortal) or it does not. If phenomenal concepts do bring with them new properties, it becomes unclear whether the account still qualifies as physicalist, as this arguably entails that there are properties that Mary did not associate with colour experience while confined to her room. But Mary, by stipulation, already associated all the relevant *physical* properties with colour experience. On the other hand, if phenomenal concepts do not introduce any new property,

then the resulting phenomenal knowledge seems insufficiently substantial, or robust (Veillet 2015). Variants of this argument appear in Block (2007), Levine (2007), Tye (2008), Schroer (2010), Levin (2019).

To see this more clearly, consider the options available to the physicalist. One option is to construe phenomenal concepts as employing phenomenal descriptions—whether as definitions, reference-fixing material, or sortals (e.g. "an experience like this"). According to Tye (2008), this approach is problematic because it reintroduces irreducibly phenomenal properties within the mode of presentation itself:

> This sets off a vicious regress and so gives us no satisfactory account of how phenomenal concepts operate. The same is true if we say that phenomenal concepts are primitive rigid concepts whose reference is fixed by a phenomenal description, for how do the concepts expressed in the phenomenal description refer? Given that phenomenal concepts have their reference fixed by a phenomenal description, the answer must be "by further associated phenomenal descriptions", and so on without end.
>
> (Tye 2008, p.44)

Block (2007) observes that this amounts to an instance of the so-called "Property Dualism Argument": positing phenomenal modes of presentation effectively reintroduces phenomenal properties as the means through which those modes of presentation pick out their referents.[34]

Another option is to construe phenomenal concepts as employing *physical* descriptions—again, either as definitions, reference-fixing material, or sortals. But here the

[34] See Díaz-León (2016) for discussion.

problem is that if phenomenal concepts fix their reference by means of physical descriptions, then it is unclear why Mary could not have acquired them while still inside her room. Further, if that were the case, phenomenal truths would arguably be deducible from physical truths.

A third option is to maintain that phenomenal concepts refer *directly*, without sortals or reference-fixers (see Loar 1997, Tye 2000, Díaz-León 2016). To a first approximation, a concept C can be taken to refer directly to a quality Q if and only if, under normal cognitive conditions, C is tokened in an act of thought just in case Q is tokened and because Q is tokened (Tye 2000). However, Tye later dismisses this proposal:

> [On this proposal] *what* Mary thinks is not new when she leaves her room. What is new is the *way* she is thinking what she is thinking. That isn't enough. What Mary knows before time t (the time of her release) is exactly the same as what she knows after time t. But if what she knows before and after her release is the same, she does not make a discovery in any really robust sense. This is counterintuitive. Surely if anyone ever made a significant discovery, Mary does here. The proposal, in the end, is not convincing. Thus, the phenomenal-concept strategy is in deep trouble. No one has yet managed to produce a plausible account of phenomenal concepts that gives them the features they must have in order to do the work needed to defend physicalism.
>
> Tye (2008, pp. 55-56)

According to a posteriori physicalists, Mary's new knowledge can be explained in terms of her acquiring new modes of presentation of the same old properties and facts—and, in a Fregean framework, thereby entertaining new *fine-grained* propositions. So, there is a sense in which it is indeed true that she does not make a "substantive" discovery: she does not come to know any new facts or entertain any new coarse-grained proposition, if we set aside

the "insubstantial" (e.g. metalinguistic) propositions mentioned in the previous section. Thus, in a sense it is true that what is new is not *what* Mary is thinking, but rather the *way* she is thinking about it. This is precisely what proponents of anti-Fregeanism—who posit coarse-grained propositions—claim. In other words, as Díaz-León (2016) notes, to say that Mary "merely" comes to know the same fact in a new way (under a new mode of presentation), or merely comes to acquire knowledge of new *fine-grained* propositions is, in effect, just another way of stating the a posteriori physicalist's position.

At this point, it is worth pausing to ask how clear a grasp we have on the explanandum of this New Challenge to physicalism. It is far from obvious what, exactly, physicalists are being asked to explain. Arguably, this "substantiveness" of Mary's new experience cannot be reduced to a *mere* difference in cognitive significance relative to her prior knowledge, since that is precisely what the introduction of new modes of presentation is meant to explain. According to the Fregean criterion, two sentences or thoughts differ in cognitive significance if and only if the same rational agent can simultaneously believe one to be true and the other to be false. Clearly, Mary's new thoughts do differ in this sense from her previous thoughts—that difference in cognitive significance is built into the very structure of the thought experiment. But the introduction of new modes of presentation *does* suffice to explain differences in cognitive significance. As the examples from the previous section show, there can be differences in cognitive significance even when "insubstantial" (e.g. merely demonstrative or metalinguistic) modes of presentation are involved. Therefore, arguably, *substantiveness*—what physicalists are being asked to explain—cannot reduce to mere cognitive significance.

Alternatively, talk of substantial knowledge could mean knowledge of new facts—or new coarse-grained propositions—of the sort that would result if the new mode of presentation introduced new properties. But, it may be argued, asking physicalists to explain

why new phenomenal knowledge is substantial in that sense would amount to asking them to explain why it grants thinkers access to new (non-physical) properties. As Veillet (2015) points out, it may *turn out* that the best account of the significance of new phenomenal knowledge involves an appeal to new properties, but significance itself cannot be initially spelt out in terms of the grasping of new properties, or else the challenge ends up begging the question. A similar problem arises if substantial knowledge is defined as knowledge that provides new information: if "information" is intended in the fine-grained sense, then the physicalist can account for it through a mere difference in mode of presentation. But if "information" is intended coarsely, the challenge amounts to asking the physicalist how Mary's new knowledge can consist in knowledge of new facts, or possibilities—which is just to ask the physicalist to explain the falsity of physicalism. The burden is on those who believe that the challenge has been misconstrued to provide a clearer account of what "substantial", or "robust" knowledge is supposed to be.[35]

§4.2 Acquaintance

Even setting aside concerns about whether the New Challenge can be formulated without begging the question, and assuming that our intuitive grasp of the notions of "richness" or "substantiveness" involved in its formulation is good enough, I believe the a posteriori physicalist can still meet the challenge by appealing to modes of presentation. Tye's (2008)

[35] According to Veillet, what might constitute the explanandum of the challenge is our *judgment* that Mary acquires substantive new knowledge, and the prospects for the physicalist to provide a satisfactory account of this phenomenon seem promising. As Veillet notes, *Zombie Mary*—who lacks phenomenal consciousness—would likewise judge, upon seeing red for the first time, that she has just acquired new and significant information. Therefore, whether or not this judgement is true, there must be a purely physicalist explanation of her coming to token that first-person judgment. And if so, there must likewise be a purely physical explanation of *Mary's* coming to token the parallel judgment that she has just acquired new and significant knowledge. Arguably, the same account also applies to *our* judgment that Mary has acquired significant knowledge, since that judgment is formed by imaginatively putting ourselves in Mary's situation.

own account offers a potential solution. Although Tye's view is not a version of the Phenomenal Concept Strategy, the claim that Mary's case is analogous to a Frege puzzle—in that her new knowledge involves new modes of presentation of the relevant properties—is not necessarily a version of the PCS either. This is because the notion of a phenomenal *mode of presentation* is broader and more flexible than that of a phenomenal concept.

On Tye's (2008) account, Mary gains new acquaintance knowledge of phenomenal character. What she lacks in the black-and-white room, despite knowing all the relevant facts (or propositions) about experiences of red, is knowledge by acquaintance. Acquaintance knowledge, in this sense, is knowledge of things directly encountered in experience. By contrast, we may have knowledge of things we have not encountered in experience, but this kind of knowledge essentially involves knowing truths about the things in question, whereas acquaintance does not. One can be acquainted with something without knowing any truths about it. As Tye puts it:

> Mary, in knowing what it is like to experience red, stands in the knowing-that relation to the fine-grained proposition that this is what it is like to experience red and is in a position to entertain this proposition in a phenomenal way via her acquaintance with the color red. Mary's consciousness of red gives her objectual knowledge by acquaintance of red, and (partly) via that knowledge she knows a certain proposition. On this view, we can say that after she leaves her room Mary knows a certain fact (partly) by knowing a certain entity she did not know in her room (namely red or the phenomenal character of the experience of red) and that this combined knowledge is what is needed to know what it is like to experience red.
>
> (Tye 2008, p. 133)

The reason why this account is compatible with—and indeed well aligned to—the proposal that Mary's case is analogous to a Frege puzzle is that, on a broad enough understanding of modes of presentation, entertaining a proposition "in a phenomenal way", "via acquaintance" (in Tye's own words) with a property just amounts to knowing the relevant proposition under a certain, phenomenal, mode of presentation. Of course, Mary gains acquaintance knowledge upon seeing red for the first time. But this view is consistent with the claim that knowing what it's like to see red amounts to propositional knowledge: just as, according to intellectualism, knowing how to ride a bike requires knowing *that* w is a way for someone to ride a bike under a *practical* mode of presentation, knowing what it's like to see red might require knowing *that* seeing red is like Q under a *phenomenal* mode of presentation—where this, in turn, requires knowing Q by acquaintance.

As Salmon (1986) observes, the mode of apprehension, or the way in which one entertains a proposition (i.e. its mode of presentation) depends on the mode of apprehension of its constituents:

> What is important is to recognize that, whatever mode of acquaintance with an object is involved in a particular case of someone's entertaining a singular proposition about that object, that mode of acquaintance is part of the means by which one apprehends the singular proposition, for it is the means by which one is familiar with one the main ingredients of the proposition. This generates something analogous to an "appearance" or a "guise" for singular propositions. If an individual has a certain appearance, either objective or subjective, and through perceiving the individual one comes to have some thought directly about that individual—say, a thought that would be verbalized as "Gee, is he tall"—then there is a sense in which the cognitive content of the thought may be said

to have a certain appearance for the thinker since one of its major components does.

(Salmon 1986, p.109)

The proposal is thus that phenomenal modes of presentation involve acquaintance with the relevant properties. Acquaintance does not introduce new properties. By definition, it does not present a property or entity via a *description* that would itself introduce further properties. Rather, we are presented with phenomenal properties not via descriptive mediation, but directly.

For present purposes, I will not commit to a specific account of acquaintance. However, we arguably want to be able to say that Mary knows what it's like to see red—or is able to entertain the relevant propositions—under her newly acquired phenomenal mode of presentation even when she is not directly experiencing red. Several moves are available here. One might say that standing belief states are dispositions to be in occurrent belief states, and that an occurrent belief state involving the phenomenal character of experiences of red, entertained under a phenomenal mode of presentation, requires at least visually imagining an experience of red, and thus being acquainted with it. On this view, Mary counts as knowing what it's like to see red via acquaintance because she is disposed to form beliefs that involve imagining experiences of red, and thereby being acquainted with their phenomenal character. Another option, drawing on the mental file metaphor, is that a mental file is created whose referent is fixed through acquaintance when Mary first has the experience of red. The file then persists over time, retaining the same referent, and can be later deployed in thought or imagination. Since the referent of this file is determined by the

acquaintance relation, the file—the mode of presentation—remains, in this sense, "acquaintance-based".[36]

The idea that modes of presentation should be broad enough to encompass acquaintance is not *ad hoc*, given the variety of things that are independently required to serve as modes of presentation. In this case, the mode of presentation is simply experience, the means through which we come into contact with the relevant entities—namely, phenomenal properties. Modes of presentation are not limited to concepts. Practical modes of presentation are, arguably, non-conceptual. Linguistic expressions (such as terms or sentences) can also sometimes serve as "guises" or modes of presentation of their referents, particularly in cases where there is no semantic difference between the expressions themselves (e.g. in the case of synonyms like "furze" and "gorse"). Purely linguistic differences can give rise to differences in cognitive significance: for any two sentences of the form "a=a" and "a=b", where "a" and "b" are co-referential, there will always be room for rational doubt as to whether one is true and the other is false, because the difference in signs alone can generate (rational) doubts about identity and co-reference. Plausibly, this phenomenon is, at least in part, what has led many philosophers to posit very fine-grained and language-sensitive modes of presentation. If expressions themselves can serve as modes of presentation, then, arguably, so can other kinds of non-conceptual representational "vehicles", including experience. This suggests that the notion of a mode of presentation is broader than that of a concept, at least as conceived by proponents of the PCS. Consequently, the account of Mary's case as a Frege puzzle is not necessarily a version of the PCS.

In short, the idea is that Mary comes to know the phenomenal character of experiences of red under an acquaintance mode of presentation—one that is not descriptive and therefore does not introduce any new reference-fixing properties. Knowing phenomenal

[36] I am grateful to Robbie Williams for raising this issue, which deserves further development.

properties under phenomenal modes of presentation requires having been acquainted with them in experience. This explains why Mary could not know the relevant propositions under a phenomenal mode of presentation before leaving her room, even though, as Tye (2008) notes, she may already have possessed deferential phenomenal concepts. This also helps explain why it is difficult to find a contingent proposition—other than metalinguistic or demonstrative ones—that Mary comes to know at t3, when she learns the informative identity claim "Wow is ph-red": there are no reference-fixing properties that could be used to construct contingent propositions from the relevant modes of presentation. Acquaintance does not present the relevant properties via other properties, but directly.

A potential objection is that Mary might already be acquainted with the phenomenal character of experiences of red by observing the relevant property instantiated in someone else's brain from her black-and-white room. Tye's (2008) account sidesteps this worry by holding that phenomenal character is a property of external objects, rather than of experiences. However, the objection may be addressed without appealing to this claim. For example, one can note that there are different ways of being acquainted with the same property. We can be acquainted with a property (e.g. a shape property) through different sensory modalities—e.g. visually or tactually—without thereby employing different concepts or introducing new sortals via distinct token demonstratives (a move that would arguably reintroduce the problem posed by the New Challenge). Similarly, acquaintance with a phenomenal property observed from a third-person perspective may differ significantly in cognitive value from acquaintance with the same property as instantiated in one's own experience.

The broader lesson from Frege puzzles is that the sense of substantiveness we associate with certain discoveries may derive not from acquiring genuinely new information, but from coming to know new fine-grained propositions, or old propositions under new,

cognitively richer modes of presentation. Substantiveness, whatever exactly it amounts to, need not be explained by appeal to genuinely new information, of the kind that opens up new facts or new possible worlds.

As I have noted, the challenge is, first, to provide a precise account of the alleged "substantiveness" or "robustness" of Mary's new knowledge. Second, we must avoid assuming from the outset that any shift in mode of presentation is too insubstantial to account for such knowledge. My view is that the richness of Mary's new knowledge lies not in its involving new content, but in its mode of presentation. In particular, the nature of acquaintance—as direct, immediate, and cognitively rich—may give the impression that Mary has gained access to genuinely new non-physical facts. But, under a sufficiently broad conception of modes of presentation, Mary's case arguably remains a Frege puzzle.

§ Conclusion

To recap, there appear to be no structural differences between Mary's and Jane's epistemic progress at t3. In both cases, what is acquired may be knowledge of new fine-grained propositions, new knowledge of old propositions, or no new knowledge at all—depending entirely on the semantics of propositional attitude ascriptions, particularly on whether differences in modes of presentation affect propositional content or play a role in attitude attribution. The intuition that Mary learns something new at t3 is thus analogous to the intuition that Jane does. Similarly, under plausible assumptions about the semantics of "know" + wh-clauses, Mary's coming to know what it's like to see red at t2 amounts to propositional knowledge under a phenomenal mode of presentation. This gives rise to a new

Frege puzzle: whether Mary gains new knowledge at t2 turns on which account of the semantics of propositional attitude ascription is correct.

The objection that Jane learns a new contingent proposition at t3, while Mary does not, fails to undermine the analogy: with respect to epistemic progress, there is no asymmetry between paradigmatic Frege puzzles and KA. Finally, as for the so-called "New Challenge" to a posteriori physicalism—according to which either phenomenal modes of presentation introduce new non-physical properties or Mary's phenomenal knowledge is insufficiently substantial—it is unclear what, precisely, the challenge demands of the physicalist. In any case, a sufficiently broad understanding of modes of presentation allows the physicalist to account for phenomenal modes of presentation in terms of acquaintance, which presents phenomenal properties in a direct and "rich" manner without introducing further reference-fixing properties.

CHAPTER 3

On Representational Grounding

One key distinction between approaches to grounding, or metaphysical explanation, concerns whether grounding connects distinct facts or distinct representations of a single fact. On "worldly" generative approaches, grounding connects distinct, *more or less fundamental facts*. On "representational" reductive approaches, grounding connects distinct, *more or less perspicuous truths* representing the same fact. In this chapter, I focus on the reductive approach, arguing that the notion of representational grounding needs to satisfy two constraints: Independence and Structure. On the one hand, reductive and generative approaches are mutually inconsistent in specific cases and yield competing views in metaphysical debates. Therefore, I argue, the relation of representational grounding needs to be defined *independently* of worldly grounding (Independence). On the other hand, representational grounding needs to have certain structural features: despite being sometimes labeled as a "reductive" relation, it gives rise to a hierarchy of *distinct* truths, which requires it to be an asymmetric and irreflexive relation (Structure). After considering various candidate relations of representational grounding, I contend that none satisfies both constraints. I conclude by suggesting that the perceived asymmetry in representational fundamentality between the truths in question may be explained away by appealing to pragmatic features of explanation.

§ Introduction

One of the distinctions that can be drawn between approaches to grounding, or metaphysical explanation, concerns whether grounding connects distinct facts or distinct representations

of a single fact.[37] In a recent paper, Rubenstein (2024) refers to these two approaches as "generation" and "reduction", though similar distinctions have been drawn using different terminology—for example, "worldly" vs "representational" approaches (e.g. Jones 2022). In this chapter, I use the expressions "worldly grounding" and "representational grounding" as interchangeable with "generation" and "reduction", respectively.

On the generative approach, the more fundamental "generates", or "gives rise to" the less fundamental. This generative relation is often taken to be structurally analogous to causation. Just as causation connects distinct events, generation connects distinct facts: the grounding facts give rise to the grounded facts. Grounding is typically treated as a strict partial order: that is, as irreflexive (no fact grounds itself), asymmetric (if fact F grounds fact G, then G does not ground F), and transitive (if F grounds G, and G grounds H, then F grounds H).[38] On this view, reality is "layered" in a hierarchy of worldly levels of fundamentality. For ease of presentation, I will frame the generative approach in terms of grounding between facts, though analogous formulations can be developed in terms of properties, entities, or propositions.[39]

On the reductive approach, on the other hand, grounding is a relation between what Rubenstein calls "truths", representing, or being made true by, the same fact. Accounting for a certain case reductively means positing *one* worldly level (hence the label "reduction") and multiple representational levels. One of the main attractions of this view is precisely that it

[37] For present purposes, I will set aside concerns about the nature of the relation between grounding and metaphysical explanation.

[38] The relation of full grounding is often understood as a relation between a collection of facts Γ and a fact G, such that the facts in Γ together fully ground G. Each member of Γ is then typically said to partially ground G. The relation of partial grounding can itself be characterized as irreflexive, asymmetric, and transitive. This framework also supports the attribution of non-monotonicity to grounding: even if a collection of facts Γ grounds a fact G, it does not follow that every larger collection of facts including Γ also grounds G, since the addition of facts irrelevant to the obtaining of G may undermine the grounding relation.

[39] The entity-based understanding of grounding is more controversial (see List 2017, p. 12; Kim 2002, p. 11).

avoids positing unparsimonious worldly layers, and purports to account for the apparent worldly asymmetry in terms of a representational asymmetry.

The distinction can be applied to various metaphysical debates, where the two approaches correspond to competing views (Rubenstein 2024). For example, the nature and existence of a composite object is often taken to be metaphysically explained by the existence and nature of its parts: a table exists and has a certain mass at least partially because the atoms which make it up exist and have certain masses. On the generative approach, the fact that the atoms exist and have certain masses generates, or gives rise to, the fact that the table exists and has a certain mass. On the reductive approach, by contrast, talk of atoms arranged "table-wise" and talk of tables represent the same fact, the same underlying portion of reality. However, proponents of reduction claim that truths about atoms are somehow more "perspicuous" than truths about tables, in the sense that they better represent the fact in question.

Similarly, an object's determinable properties are naturally explained in terms of its determinate properties: the apple is red because it is scarlet. On the generative approach, the facts involving the former are generated by the facts involving the latter: the fact that the apple is scarlet *grounds* the fact that it is red. By contrast, on the reductive approach there is a single fact about the colour of the apple, which is better represented by truths about scarlet than by truths about red: saying that the apple is scarlet yields a more perspicuous representation of facts concerning the colour of the apple than saying that the apple is red.

Thus, the two approaches posit competing hierarchies. One is a representational hierarchy of truths, from the more perspicuous to the less perspicuous, the other is a worldly hierarchy of facts, from the more fundamental to the less fundamental. In general, on Rubenstein's and my usage, whether a hierarchy is "worldly" or "representational" depends entirely on whether its relata are worldly items—properties, facts, objects, etc.—ordered

according to some feature like fundamentality (or joint-carvingness, or naturalness), as opposed to representational items—predicates, sentences, fine-grained propositions, etc.—ordered according to their perspicuity.

Many have expressed scepticism about generation, or worldly grounding (e.g. Hofweber 2009, Daly 2012, Wilson 2014). Some complain that it is unclear what grounding is, and how it should be defined in the first place, since the notion is usually explicated via examples. Grounding theorists themselves often say that the notion of grounding is primitive, or not definable in more basic terms, but claim that it is clear enough for their purposes (Fine 2001, Schaffer 2009). Another frequent complaint is that worldly grounding is too general to be useful, and we should stick with more specific relations of metaphysical dependence, such as functional realization, classical mereological parthood, the set membership relation, the proper subset relation, the determinable/determinate relation, and so on (Wilson 2014). Furthermore, far from solving all metaphysical problems, grounding seems to create new problems of its own: for instance, concerning what grounds the grounding facts (Litland 2017, Clark 2018, Sider 2020), or whether there is a metaphysically bottom level (Schaffer 2010, Bliss 2013, Raven 2016). While I am sympathetic with these complaints, in this chapter I am going to argue against reduction, or *representational* grounding.

Several authors have discussed concepts in the ballpark of reduction and generation as formulated above, but there are relevant differences between the various available views. For example, while Rubenstein (2024) conceives of representational grounding as a relation between representational items (truths), in Jones' view (2022), only the *generated* phenomenon must be a representation for the relevant relation to count as representational. Also, on Rosen's (2010) account, reduction *entails* generation (worldly grounding), rather than being inconsistent with it (as Audi 2012 claims). Also, as Rubenstein notes, a distinction

is sometimes drawn between "representational" (or "conceptual") grounding and "worldly" (or "metaphysical") grounding (e.g. Correia 2010, Correia & Schnieder 2012, Correia & Skiles 2017). On Rubenstein's terminology, however, this distinction remains internal to the generation approach: it concerns differences in the granularity of facts, rather than corresponding to the distinction of interest here. Trying to distill a general view of representational grounding from the various accounts in the literature would likely be both difficult and of limited use, since any such formulation of the view would probably be more controversial than any critique of it. I shall thus focus on a recent version of the view, by Rubenstein (2024), though many of my points apply more broadly.

The rest of the chapter is structured as follows. In §1, I will identify two constraints that the relevant notion of representational grounding should satisfy: Independence and Structure. I will also discuss the relata of representational grounding, arguing that they can only be characterized as sentences or *fine-grained* propositions. In §2, I will consider various candidate notions of perspicuity—in terms of which representational grounding is defined—and argue that they do not satisfy Independence. In §3, I will consider one candidate notion of representational grounding which satisfies Independence—a familiar notion of entailment—but argue that it does not satisfy Structure. Moreover, I will show how attempts to characterize representational grounding in terms of a restricted entailment relation with the right structural properties, are bound to fail. I will conclude by suggesting that the perceived asymmetry in representational fundamentality between the truths in question may not reflect any *objective* difference between such truths. Instead, our judgments about differences in relative perspicuity may be explained away as the byproduct of a cluster of linguistic, cognitive, and pragmatic factors, including pragmatic features of explanation. What I mean by "objective" perspicuity here can be illustrated through the following example from Rubenstein. The number 476 can be represented as built up from powers of

ten, as $(4 \times 10^2) + (7 \times 10^1) + (6 \times 10^0)$. This construction is, arguably, "perspectivally" but not objectively privileged: it is highlighted by our decimal system of representation, yet nothing in the numbers themselves distinguishes it. In this sense, this representation is not *objectively* perspicuous. Alternatively, the same number can be represented as built up from its prime factors: 476 is $2^2 \times 7 \times 17$. As Rubenstein observes, this construction is objectively but not perspectivally privileged: it reflects the deep nature of the numbers themselves, but appears strangely random from our own perspective (2024, p. 1126).

For present purposes, I will rely on a distinction between facts and propositions, but nothing significant hinges on this assumption. A proponent of the identity between facts and true propositions may simply define generation as a relation between facts/propositions and reduction as a relation between sentences expressing the same proposition. Following Rubenstein, I will also treat grounding claims as expressed by a relational predicate, rather than a sentential operator,[40] and adopt the following notation: '<p>' denotes the truth that p—as we will see, '<p>' may denote either a sentence or a proposition—while '[p]' denotes the fact that p. '[p]' denotes whichever fact, if any, <p> represents.

§1 Representational Grounding

§1.1 Independence and Structure

What is the relation between representational and worldly grounding? One might think that, as Rubenstein argues, metaphysical inquiry as a whole requires both tools: some metaphysical issues may be best treated in terms of reduction, others in terms of generation.

[40] As Correia (2010) notes, much of the issues discussed in this chapter, such as factual and propositional identity, arise for the operational view as well.

For example, cases involving facts that are not even intensionally equivalent may be best treated in terms of generation. These include multiply realizable facts and their realizers, or logically simple and logically complex facts, such as [A ∨ B] and [A]. The fact [there is a cat], for instance, may be realized by the obtaining of various distinct cat-like configurations of particles C1, C2, C3, etc. Thus, [there is a cat] and [C1 obtains] do not obtain at exactly the same possible worlds: in some worlds, [there is a cat] obtains in virtue of C2 obtaining instead. In this sense, [there is a cat] is, arguably, not identical to [C1 obtains], and thus reduction seems ill-suited to account for this case. Since genuinely distinct facts are involved, only *generation* can account for the relation between them. Similarly, assuming that B does not entail A, [A ∨ B] is not intensionally equivalent to [A], as in worlds where B but not A is the case, A ∨ B is the case. Thus, [A ∨ B] and [A] are not identical: they are distinct facts. But then, again, reduction seems ill-suited to account for this case, because it posits just *one* fact and multiple, more or less perspicuous, representations of it.[41]

On the other hand, consider [John is a bachelor] and [John is an unmarried man]. Most theorists would agree that this case just involves *one* fact, which can be represented by distinct truths, such as <John is a bachelor> and <John is an unmarried man>. In this case, then, generation will not apply, as there are no distinct facts for worldly grounding to connect. A reductive view seems best-suited: a relation of representational grounding obtains between <John is an unmarried man> and <John is a bachelor>, but both truths represent the same underlying portion of reality. In this sense, metaphysics might require both reduction and generation: neither a purely reductionist nor a purely generative approach to metaphysics *as a whole* might be tenable.

[41] Of course, some theorists deny that there are any such things as disjunctive facts. I discuss this view in §2.2.

However, reduction and generation are mutually exclusive in *specific* cases.[42] As explained, the distinction between reduction and generation can be used to make sense of opposing views in several metaphysical debates, including debates about mereology, determinable properties, or phenomenal states. For example, a proponent of reduction might claim that, although any truth about pain reduces to a distinct truth about C-fiber firing—perhaps on the grounds that one might believe that one is in pain without believing that one has C-fiber firing—only one property, and thus one fact, is involved. According to the reductionist, the fact [John is in pain] *just is* the fact [John's C-fibers are firing], but <John's C-fibers are firing> is a more perspicuous representation of this fact than <John is in pain>. The reductive approach in this case will thus yield a reductive form of physicalism. By contrast, a generator will say that any fact about C-fiber firing gives rise to a *distinct* fact about pain: [John's C-fibers are firing] gives rise to [John is in pain]. This will thus amount to a non-reductive form of physicalism. In this sense, reduction is inconsistent with generation, just as reductive physicalism is inconsistent with non-reductive physicalism. The same applies to the debates about mereology and determinate/determinable properties described earlier. Where generation posits distinct facts, reduction posits one fact and distinct truths.

Of course, other accounts of reduction may not be inconsistent with generation—for example, Rosen (2010) argues that reduction *entails* generation. Rubenstein specifies: "Since Rosen posits a worldly distinction between the facts in question, reduction in his sense is not reduction in my sense" (2024, p. 1111). To clarify the disagreement between Rubenstein and Rosen, we should look more closely at Rosen's account. Rosen posits a

[42]Although Rubenstein claims that, in some particular cases, the metaphysical explanation of a certain phenomenon will involve both reductive and generative *steps* (for example, cases involving "imprecise truths" or context-relative facts).

Grounding-Reduction link, according to which, if <p> is true and <p> ⇐ <q>, then [p] ← [q]; that is, if <p> reduces to <q>, then [q] grounds [p]. For example, if to be a square is to be an equilateral rectangle, then if ABCD is a square, it is a square *in virtue* of being an equilateral rectangle.

However, as Rosen acknowledges, the link generates a puzzle: if to be a square *just is* to be an equilateral rectangle, then the fact that ABCD is a square and the fact that ABCD is an equilateral rectangle are the very same fact. But then, the Grounding-Reduction link must be mistaken, since every instance of it would violate the irreflexivity of grounding. Rosen's solution is to adopt an extremely fine-grained account of facts on which "the operation of replacing a worldly item in a fact with its real definition never yields the same fact again" (Rosen 2010, p.124). Yet, most metaphysicians will find this response unpersuasive: while it is not uncommon to posit fine-grained propositions, facts are typically treated as more coarse-grained. Indeed, Audi (2012) responds to Rosen arguing precisely that reduction, far from entailing generation, is inconsistent with it. In short, in order to maintain the Grounding-Reduction link, Rosen must assume an extremely fine-grained view of facts. On his account, even if being a bachelor *just is* being an unmarried man, the fact that John is a bachelor and the fact that John is an unmarried man are distinct (and the latter grounds the former). In what follows, I will follow Rubenstein in maintaining that reduction does not entail generation and is, in fact, inconsistent with it. Similarly, Jones (2022) conceives of analogous views, which he calls "Representational Levels" and "Worldly Levels", as mutually inconsistent.[43] A reasonable requirement for a candidate notion of reduction in our sense is therefore that it be defined independently of worldly grounding:

[43] In a later paper, Jones (forthcoming) proposes an account that combines worldly fundamentality and representational fundamentality roughly as follows: 'a is fundamental' is true just in case 'a' is a structurally fundamental (or perspicuous) representation of its denotation. This account entails that fundamentality is

Independence**:** Reduction needs to be understood independently of generation, in the sense that the representational hierarchy should be constructed without positing a corresponding worldly hierarchy.

There are two main motivations for Independence.[44] The first is what we may call the *identity motivation*. Proponents of reduction claim that some truth <p> reduces to some other truth <q>. As explained, if <p> reduces to <q>, then [p]=[q], even though <p> ≠ <q>. Therefore, reduction cannot be defined in a way that entails worldly grounding between the relevant facts, since worldly grounding is supposed to be asymmetric, and thus any such definition would entail that [p] ≠[q]. Consider, for example, the toy case above involving the two truths <John is in pain> and <John's C-fibers are firing>. The reductionist holds that the latter is more perspicuous than the former. Independence tells us that our account of this asymmetry in perspicuity cannot entail that [John's C-fibers are firing] grounds [John is in pain].

The second motivation for Independence is what we could label the *flat reality* motivation. The idea here is that reduction offers an alternative to worldly grounding, allowing views that reject hierarchical structure in reality to nonetheless posit hierarchical structure in our *representation* of reality. In other words, Independence is required by any account of reduction that posits a *flat* ontology—one that rejects metaphysical hierarchies among facts, entities, or properties—while still preserving the idea that reality can be represented in objectively better or worse ways.

---

opaque: on this view, it might be true that a is fundamental while b is not even though 'a' and 'b' are co-referential.

[44] I am grateful to Nick Jones for insightful discussion on this point.

In many cases, explaining the asymmetry in perspicuity between two truths by appealing to grounding relations between the constituents of the relevant facts will end up imposing a worldly hierarchy among the facts themselves, thereby violating the identity motivation. For example, one might attempt to explain the asymmetry in perspicuity between <John is in pain> and <John's C-fibers are firing> by saying that "pain" is less perspicuous than "C-fibers firing", which in turn is explained by positing a grounding relation between the properties of C-fiber firing and pain. However, this worldly asymmetry between the properties would entail an asymmetry at the level of facts: [John's C-fibers are firing] would be distinct from [John is in pain], violating the identity constraint.

That said, this does not hold in all cases—or at least not under certain assumptions. There are cases in which <p> and <q> represent intensionally equivalent facts, even though <p> and <q> involve predicates referring to distinct properties that may be taken to stand in a grounding relation—thereby explaining the difference in perspicuity—without entailing a distinction between the relevant facts. For instance, the two truths <Romeo loves Juliet> and <Juliet is loved by Romeo> are intensionally equivalent and thus represent the same fact in an intensional framework ([Romeo loves Juliet] = [Juliet is loved by Romeo]). Suppose, however, that "being loved by Romeo" and "loving Juliet"—which denote distinct properties even on an intensional view—differ in perspicuity, and that *this* difference is explained by positing a grounding relation, or some other metaphysical asymmetry, between the corresponding properties. In turn, this could account for a difference in perspicuity between the relevant truths. These cases do not technically violate the identity motivation, since the *facts* involved remain identical, even though the account of the asymmetry in perspicuity between the truths in question *does* invoke a worldly hierarchy (between properties). Nonetheless, these cases remain in tension with reduction because they violate the *flat reality*

motivation for Independence, which rules out any worldly hierarchy in our account of the asymmetry in perspicuity—even if no such hierarchy is posited at the level of facts.

There is another constraint that a candidate notion of representational grounding needs to satisfy, which has to do with the *structure* of the relation itself. We said that the representational relation between truths posited by the reductive approach is supposed to account for the perceived asymmetry in fundamentality between apparently distinct facts. Let us consider the pain example again: the reductionist claims that [John is in pain] and [John's C-fibers are firing] are one and the same fact, which entails that one cannot be more fundamental than the other. In order to account for the apparent asymmetry in fundamentality between [John is in pain] and [John's C-fibers are firing], the reductionist posits a corresponding relation at the level of representations: <John's C-fibers are firing> is simply *a more perspicuous representation than* <John is in pain> of the only fact involved ([John is in pain]= [John's C-fibers are firing]).

For any property P, the predicate "more P than" denotes a relation that is asymmetric, irreflexive, and transitive: intuitively, nothing can be more P than itself (irreflexivity), if x is more P than y, then y is not more P than x (asymmetry), and if x is more P than y and y is more P than z, then x is more P than z (transitivity). Asymmetry, irreflexivity and transitivity are, in a sense, built into the very idea of a hierarchy. Thus, being defined in terms of truths being *more perspicuous than* other truths, reduction will need to satisfy the following principle:

Structure**:** Representational grounding needs to be asymmetric, irreflexive, and transitive.

As said, representational grounding is sometimes labeled "reduction" and sometimes conceived as a matter of truths "giving rise to" other truths. The discussion above, however, suggests that the difference is merely terminological, given that all these theorists take the relevant relation to yield a hierarchy of truths, which requires it to satisfy Structure.

Before discussing the *relation* of representational grounding, in the rest of this section I will discuss its *relata*, arguing that the role of truths can only be played by true sentences or true *fine-grained* propositions.

§1.2 The Relata of Representational Grounding

Rubenstein claims he wishes to remain neutral on the exact nature of the relata of reduction: "Truths, as I use the term, are just true truth-bearers: perhaps sentence-tokens, or sentence meanings/propositions" (2024, p. 1109). However, the distinction between sentences and propositions is not irrelevant, as these play very different theoretical roles. On a standard view, propositions, not sentences, are the objects of propositional attitude ascriptions. Also, sentences are language-specific, while propositions are not—if sentences are the relata of representational grounding, the question of their relative representational fundamentality might need to be addressed in different languages, possibly resulting in distinct representational hierarchies for different natural languages. Furthermore, there are various accounts of propositions, differing in how fine-grained they are, what their constituents are, and whether they are structured or not. These distinctions, by contrast, do not apply to sentences. And even in the most fine-grained frameworks, propositions are (supposedly) more coarse-grained than sentences, and correspond to sentences one-to-many. These factors, and especially questions of fineness of grain, matter in trying to construct a representational hierarchy. In particular, as I will explain, the role of truths can only be filled

by true sentences or true *fine-grained* propositions, whereas coarse-grained propositions cannot serve as the relata of representational grounding.

We can start by observing that the tension between reduction and generation resurfaces even *within* the representational hierarchy. On the one hand, the relation between the representational levels is, at least nominally, reductive. As Rubenstein claims, "reducers take ground to be a matter of truths reducing to (consisting in, or collapsing into) others […] intuitive examples are the reduction of truths about water to truths about H_2O and of truths about heat to truths about molecular motion" (2024, p.1112). It is natural, he claims, to hold that the truths in question represent the same fact. For example, the water in the glass "is really just" a collection of H_2O molecules. On the other hand, however, we have seen that the concept itself of a hierarchy is somehow intrinsically non-reductive: in order to prevent the hierarchy from collapsing into a single level, its relata need to be distinct from each other—hence the irreflexivity of representational grounding. This balance between reduction and generation is only tenable under a very fine-grained conception of truths as either sentences or fine-grained propositions.

Suppose that intensionalism—the view that necessarily equivalent propositions are identical—is correct. Since the true propositions <There is water in the ocean> and <There are H_2O molecules in the ocean> are, we may assume, necessarily equivalent, it cannot be the case that the proposition <There are H_2O molecules in the ocean> is more representationally fundamental, or more perspicuous, than <There is water in the ocean>. Yet, as I will explain below, the relata of reduction cannot be intensionally distinct propositions either, as these would correspond to distinct facts, rendering the cases in question unsuitable for a reductive account.

To see why intensions cannot be the relata of reduction, we should note that, on the reductive approach, truths correspond many-to-one to facts: cases of reduction are those

where only one fact is involved, and multiple truths representing this fact form the representational hierarchy. Therefore, if truths are propositions, truths must be *more fine-grained* than facts. Given that the most coarse-grained (reasonable) account of facts defines them in terms of sets of worlds, truths will need to be more fine-grained than sets of worlds on *any* reductive account. If, for example, [2+2=4] and [3+1=4] are the same fact—corresponding to the full set of worlds—the relevant propositions <2+2=4> and <3+1=4> will need to be distinct, despite being co-intensional. This means that, if truths are propositions, they must be *hyperintensional* propositions.

Intensional accounts of facts are not universally accepted, as many theorists conceive of facts as more fine-grained than sets of worlds. In particular, it might be argued that facts are individuated by their constituents and the relations between them: for example, because of their difference in constituents, [2+2=4] and [all cats are cats] may be considered as distinct facts, despite obtaining at the same set of worlds (the set of all worlds). Let us call this the "structured" account of facts. Reduction requires truths to correspond many-to-one to facts. In this framework, truths must therefore be more fine-grained not only than sets of worlds, but also than Russellian structured propositions, which correspond one-to-one to structured facts. To see why, suppose that any object o's instantiating a property P determines a fact [Po] which has o and P as constituents. For any such fact, there will be a Russellian proposition <P,o> whose constituents are the object and the property in question. Vice versa, for any Russellian proposition <P, o>, there will be a fact [Po] with the same constituents. But then, if truths were as coarse-grained as Russellian propositions, there would be at most one truth for every fact, leaving no room for a representational hierarchy. Thus, on the structured account of facts, truths must be not only hyperintensional, but even more fine-grained than Russellian propositions.

Indeed, Rubenstein's own examples suggest that he conceives of truths as extremely fine-grained. He states that truths about water may be distinct from truths about H_2O, just as truths about heat may be distinct from truths about molecular motion. However, since "water" and "H_2O" (just like "heat" and "molecular motion") pick out the same property, neither Russellian nor intensional frameworks can distinguish truths about water from truths about H_2O, or truths about heat from truths about molecular motion. Once again, truths need to be even more fine-grained. For example, truths may be conceived of as Fregean propositions, whose constituents are *representational items* (e.g. concepts) corresponding many-to-one to the relevant properties and objects which constitute the fact in question. Alternatively, truths may simply be sentences, which clearly correspond many-to-one to facts. The distinction between fine-grained propositions and sentences will not matter hugely for our purposes, and most of the considerations in the following sections apply to both. This should not be surprising, considering that fine-grained semantic frameworks seem to be somehow "language-sensitive", in the sense that they seem to project metalinguistic differences onto semantics. In particular, accounts on which the cognitive significance of sentences supervenes on their meaning tend to posit very fine-grained propositions, as differences in cognitive significance can arise from basically any syntactic difference (Williamson 2021a). On these accounts, propositions will thus end up being so fine-grained as to approximate sentences. In what follows, I will use the term "truth" as neutral between sentences and fine-grained propositions.

Of course, different accounts of facts and propositions will produce different verdicts on which cases can be treated reductively. However, a minimum requirement for a case to count as a *candidate* for reduction is that the facts involved come out as at least intensionally equivalent. For example, potential candidates for reduction include:

(1) [There is water in the ocean], [There is H_2O in the ocean]

(2) [John is a bachelor], [John is an unmarried man]

(3) [Snow is white], [The proposition that snow is white is true]

(4) [Socrates is Socrates], [Socrates is the member of {Socrates}]

For each pair in (1)-(4), the two facts involved are necessarily equivalent—i.e., obtain at exactly the same possible worlds—and thus count as the same fact in intensional frameworks. Of course, one may deny, particularly on the basis of examples like (3) and (4), that necessary equivalence suffices for identity between facts. For example, many would follow Fine (1994) in claiming that being Socrates and being the member of {Socrates} cannot be the same property, as one, but not the other is essential to Socrates—meaning that the facts in (4) cannot be identical.

The point, however, is that all these cases are *potential* candidates for reduction, as the apparent difference in fundamentality between the relevant facts can, in principle, be explained in terms of a representational difference between the corresponding truths, while maintaining a single underlying fact. If these facts were not even co-intensional, identifying them as the same would be impossible, thereby ruling out a reductive account of the case from the outset. By contrast, despite being candidates as relata for generation, the following pairs of facts are worse candidates for reduction, because they do not meet the minimum requirement of co-intensionality:

(5) [The rose is scarlet], [The rose is red]

(6) [John exists], [Particles arranged John-wise exist] (assuming that in some possible worlds John has microphysical duplicates distinct from himself).

The point is that, while the facts in (1)-(4) are, on some accounts, identical—and the corresponding truths are thus candidates for reduction—the facts in (5) and (6) come out as distinct even in the most coarse-grained frameworks. To treat (5) and (6) as cases of reduction, one would need to say that (5) and (6) involve just *one* fact not because the facts in (5) and (6) are identical, but because, for each pair, one of the two facts is not a fact at all. For example, someone who believes that only *determinate* properties exist could argue that only [The rose is scarlet] is a genuine fact, while there is no such fact as [The rose is red]. I will discuss this view in §2.2. However, anyone who regards the pairs in (5) and (6) as involving two legitimate facts will be unable to adopt a reductive approach to those cases.

Let us now focus on the *relation* of representational grounding. In the next section, I will discuss various proposals on how to characterize the notion of perspicuity involved in the definition of reduction and argue that none of them satisfies Independence. In §3, I will argue that a familiar notion of entailment would satisfy Independence but fail to satisfy Structure.

§2 The Relation of Representational Grounding: Perspicuity

§2.1 Perspicuity

In the previous section, we noted that, although the relation of representational grounding is sometimes couched in terms of truths "reducing" to other truths, "reduction" obviously cannot mean identity—otherwise the representational hierarchy would collapse onto one representational level. In other words, identity fails to satisfy Structure. It needs to be explained, then, what it means for a truth to "reduce" to another. Rubenstein proposes to

account for reduction in terms of perspicuity, where <q> reduces to <p> (or <p> representationally grounds <q>) just in case:

i) <p> and <q> represent the same fact, and

ii) <p> is perspicuous and <q> is not.

For example, Rubenstein writes, the claim that <The room temperature is y> reduces to <The mean molecular energy of the air in the room is x> means that: i) these truths represent the same fact ([the mean molecular energy of the air in the room is x]=[the room temperature is y]); ii) <The mean molecular energy of the air in the room is x> is perspicuous and <The room temperature is y> is not.

We should note that our judgements about perspicuity seem to admit of *degrees* of perspicuity. For example, <The room temperature is y> seems less perspicuous than <The mean molecular energy of the air in the room is x> but more perspicuous than <The room temperature is y or the room temperature is y and some emerald is grue>, although all these truths are necessarily equivalent—$\alpha \vee (\alpha \wedge \beta)$ has the same truth-conditions as plain α—and may be thus taken to describe the same fact. Therefore, we might want to rephrase the definition of perspicuity as follows: <q> reduces to <p> (or <p> representationally grounds <q>) just in case:

i) <p> and <q> represent the same fact, and

ii) <p> is more perspicuous than <q>.

The plausibility of this definition of reduction (or representational grounding) largely depends on the plausibility of the notion of perspicuity involved. Of course, we should take the Independence and Structure constraints to extend from representational grounding to perspicuity, given that the former is defined in terms of the latter. The rest of this section will discuss various possible accounts of perspicuity.

Rubenstein provides two possible definitions of perspicuity, which I will label "Correspondence" and "Joint-Carvingness".

Correspondence: <p> is perspicuous iff for each representational constituent of <p>, [p] has a corresponding worldly constituent, and for each structuring relation between the constituents of <p>, the constituents of [p] are correspondingly related (Fine 2001).

Joint-Carvingness: <p> is perspicuous just in case each constituent of <p> is joint-carving.

Following Sider, Rubenstein takes the notion of joint-carvingness to extend Lewisian naturalness "beyond the predicate", to names, sentential operators, quantifiers, etc. (Sider 2011). For example, Rubenstein specifies, "is grue" does not denote a perfectly natural property; hence <This emerald is grue> is not a perspicuous truth. On the other hand, perhaps "is negatively charged" does denote a perfectly natural property; hence <Electrons are negatively charged> may be perspicuous.

On the one hand, Correspondence resembles Armstrong's (1978) view that, in order to evaluate whether a certain predicate carves at the joints, we need to look at whether a certain entity exists—a universal corresponding to that predicate. On the other hand, Joint-Carvingness resembles Lewis' (1983) view that, in order to evaluate whether a predicate

carves at the joints, we need to look at the set of the predicate's actual and possible instances and ask whether that entity (the set in question) is natural (Sider 2011). In this sense, Correspondence holds that whether a predicate is perspicuous depends on whether it refers at all, while Joint-Carvingness entails that whether a predicate is perspicuous depends on whether its referent is natural.

I will discuss these notions in turn. As said, I will remain neutral on whether truths are sentences or fine-grained propositions, and whether their constituents are thus things like names and predicates, or non-linguistic items such as concepts. The differences between them will be largely irrelevant for our purposes: non-perspicuous truths are those sentences or fine-grained propositions whose terms or concepts lack a corresponding worldly item (Correspondence) or are not "joint-carving" (Joint-Carvingness). For simplicity, I will mostly refer to sentences and their constituents.

§2.2 Perspicuity as Correspondence

For a truth to satisfy Correspondence, each of its constituents must correspond to a worldly constituent. But, prima facie, that just seems to be the case for every truth: for every true sentence, for instance, the names and predicates in it must, arguably, refer to worldly entities (objects, properties, relations), for the sentence to be true in the first place. The same applies to concepts: the concept *cat*, for example, must be non-empty for <Tibbles is a cat> to be true. And, of course, it must be the case that these worldly items are appropriately related. In order for "Tibbles is a cat" to be true:

(i) "Tibbles" must refer to an object, Tibbles, and "cat" must refer to a property;

(ii) Tibbles and the property of being a cat must stand in the appropriate relation (e.g. instantiation).

But then, how can a *true* sentence be non-perspicuous? Rubenstein does provide an example of how there can be non-perspicuous truths in the sense characterized, using the sentence "The average family has 2.2 children" (2024, p. 1112). This sentence is true, but there is no such object as the average family. However, Rubenstein himself notes in a footnote that when it comes to his example, "the structure of the corresponding sentence-meaning may not match that of the sentence" (2024, p. 1112). In these kinds of cases, where the structure of the sentence-meaning does not match that of the sentence, a sentence may well be true without perspicuously representing. However, paradigmatic examples of (allegedly) non-perspicuous truths are those involving "gerrymandered" predicates such as "grue"—a predicate stipulated to refer to things that are either green and observed before a time t or blue and observed after t. It is implausible that the structure of the meaning of "This emerald is grue" is different from the structure of the meaning of "This emerald is green"—i.e. that "This emerald is grue" is non-perspicuous because of a mismatch between the sentence's structure and the structure of its meaning.

Let us focus on examples of this latter kind. Consider the sentence S "This emerald is grue". Arguably, S is non-perspicuous because "grue" is non-perspicuous. By Correspondence, this means that there is no worldly item corresponding to "grue". The idea is that, in a sparse ontology where only universals—perfectly natural properties such as mass and charge—exist, nothing worldly corresponds to a logically complex, time-relative predicate such as "grue". Still, this does not answer the question of how S can be true, if "grue" fails to refer to any property. This combination of claims—that S is true but "grue" has no referent—is only tenable under specific semantic assumptions. In particular, we need

to assume that "This emerald is grue" is not made true by the fact that a certain emerald is grue—there is no such fact, given that there is no such property as being grue—but by some fact whose constituents are simpler, more "respectable" sparse properties.

An example of this strategy is Cameron (2008)'s truthmaker semantics. According to Cameron, one of the benefits of truthmaker theory is precisely that it allows for sentences about some item x to be made true by something other than x. This framework, he claims, enables us to endorse a sparse ontology—for example a mereological nihilist's ontology countenancing only subatomic particles—while still allowing for the truth of sentences about "abundant" entities, like macroscopic objects. For example, "The book is on the table" is true even in the absence of books or tables, because it is made true by underlying facts about subatomic particles. Similarly, on this view, "This emerald is grue" is made true by facts that do not involve the property of being grue among their constituents: facts like [this emerald is green] and [this emerald is observed before t] suffice to make true <This emerald is grue>. In this sense, "grue" is non-perspicuous: the world provides no items corresponding to that predicate.

Note, however, that on this account, even statements about the *existence* of the problematic abundant items are made true by something other than these items themselves (or facts about these items).[45] For example, <Tables exist> may be made true by the fact that there are particles arranged table-wise, so it may be true even *in the absence of tables,* and of facts about tables. Yet, prima facie, the idea that it could be literally true that there are tables in the absence of tables (and facts about them) sounds quite absurd: proponents of this view seem to be committed to the inconsistent claim that there are tables and there are no tables. In response to this objection, Cameron (2008) specifies that, on this view, tables *do*

[45] Not all versions of truthmaking allow for this move: Armstrong (2002), for example, says that x is always a truthmaker for <x exists>.

exist but are not among the ontological commitments of the theory, where this is spelled out in terms of a distinction between what *merely* exists and what *really* exists (Fine 2001). In short, a theory is committed only to the truthmakers of the relevant true sentences: since "Tables exist" is true, tables do exist, but only what makes the sentence true—namely, the particles—is what *really* exists.

However, it might be objected that this distinction between what really exist and what merely exist reintroduces a worldly hierarchy, as it looks like a mere variant of the view that some things are more fundamental than others. In other words, while the truthmaking approach is meant to avoid positing the existence of certain entities, it seems to only succeed in "downgrading" them to things that merely exist, rather than really existing. Cameron downplays the metaphysical significance of the distinction between existing and really existing, saying the distinction is merely a "way of talking" (2008, p. 7), and that the merely existing entities constitute an "ontological free lunch" because they do not figure among the ontological commitments of a theory. However, even theorists who explicitly endorse a hierarchical view of reality, with fundamental entities at the bottom and non-fundamental entities above, also sometimes claim that non-fundamental entities are an ontological free lunch (Schaffer 2009, p. 361). Moreover, as Schaffer claims, a natural way of interpreting the claim that certain entities are "an ontological free lunch" is precisely in terms of (worldly) fundamentality:

> In Quinean terms, whatever supervenes is an addition to being in the only available sense—it is an additional entry on the list of beings. But in Aristotelian terms, there is a straightforward way to understand Armstrong['s claim that what supervenes is no addition to being]: whatever is dependent is not fundamental, and thus no addition to the sparse basis.

Schaffer (2009, p. 353)

Therefore, saying that things that merely exist are an ontological free lunch does not really clarify the difference between Cameron's view and a generative approach.

Williams (2010) offers an account similar to Cameron's, on which many sentences we assert about macroscopic objects and abstract objects—abundant entities more generally—are true, without this entailing any ontological commitment to such entities. For example, in this framework, "Billy sits" is true just in case the actual world "compound-represents" the existence of Billy and that Billy is one of the set of z such that the actual world compound-represents it as sitting—where "w compound-represents that q" is true just in case there is some p such that the actual world represents p as being the case, and p, combined with mereology entails q. In this case, Williams claims, it suffices that there be some simples arranged Billy-wise—for then, given mereology, it follows that Billy exists—and that these simples are arranged sitting-wise—for then, given mereology, it will follow that he sits. In this way, he argues, set-theoretic and compound-object propositions can be true without any requirement that the world contain compound objects or abstracta (2010, pp. 123-24). Like Cameron, Williams acknowledges the need for some kind of fundamentality (or "in reality") operator:

> Sets and compound objects exist then—one expresses a true proposition by saying so—but the whole point of the enterprise just sketched was to reach this point without being ontologically committed to anything more than simple *concreta*. One wants to give voice to this by saying: sets and compound objects do not exist. But to say *this* would be to fall into contradiction. So, we need to do something else to give voice to our minimal metaphysics. What we need,

> therefore, is an expressive device that will allow us to articulate the situation. […] The basic idea is to introduce an operator 'Fundamentally ϕ' which will be true at a world *w*, just in case *w* […] *really* represents that [the relevant conditions] are met, in the sense that the conditions hold at *w*. Thus, a world containing only microphysical simples might compound-represent the existence of tables, and this might be all that is required for the truth of 'there are tables'. But, the idea will be, 'Fundamentally, there are tables' should require that *w* truly *represent* that there are tables, not merely compound-represent this. And so, as desired, this will be false at a world where only the simples exist.
>
> Williams (2010, pp. 124-25)

Williams claims that this does not amount to positing a metaphysical distinction, in the sense of a metaphysical hierarchy. However, it is unclear whether we make sense of the claim that it is literally true that there are tables and literally true that fundamentally, there are no tables, if not in terms of the layered picture. In other words, it is not immediately clear how Williams' framework differs from a layered picture in which tables exist but are not fundamental, or the fact that tables exist obtains but is not a fundamental fact. In a later paper, Williams (2012) addresses a similar objection:

> There's a way of reporting the views that I've just been advocating that makes it sound close to the views of Schaffer, Fine and other friends of stratified metaphysics. For on this view, a certain image of what there is is projected from total theory. 'There are numbers', 'there are macroscopic objects' and the like will be true according to view developed. To put it less coyly and without qualification: numbers and macroscopic objects exist. What could be more

> natural than to call the totality of what exists our 'ontology'? Within the ontology, there are some entities that not only exist, but are such that they form the 'requirement-base' for the rest—that is, such that what is "required" of reality, in order that the truths be true, never invokes anything outside of this base. We could call this 'fundamental ontology', and call any part of ontology that isn't part of fundamental ontology 'merely derivative'. While I earlier suggested that the existential component of reality requirements be called 'ontological commitments'; why not call it instead 'fundamental ontological commitments', and allow a standard understanding of 'ontological commitments' simpliciter, in terms of what must feature as the values of our variables for a sentence or theory to be true? Insofar as the existence of a is part of what's required for 'b exists' to be true, we might choose to say that b is grounded in a. And so forth.
>
> Williams (2012, pp. 182-83)

In response to this line of argument, Williams notes that Quineans identify ontological commitments with the values of the variables in the total theory *once it is properly paraphrased*. But one might instead take the ontological commitments to be the values of the variables in the total theory, and regard the values of the variables in the paraphrased version of the theory as the *fundamental* ontological commitments. In this sense, even the Quinean approach—typically taken as the paradigm of a flat ontology—could be seen as implicitly introducing levels of reality. Yet, even conceding that this interpretation of the Quinean view is viable, it still leaves room for genuinely flat ontologies that do not rely on paraphrase. A mathematical realist, for instance, need not posit any paraphrase of number talk, and thus the objection does not apply to her view. Not any purportedly flat ontology

covertly introduces a hierarchy of worldly levels. One might object that even a committed realist may occasionally need paraphrase—for example, in accounting for the truth of sentences like "She did it for your sake" while avoiding a commitment to sakes. However, linguistically, this case is quite different from numerical or composite-object discourse. Dictionaries typically treat "sake" as part of an idiomatic construction, defining only the phrase "for x's sake". Indeed, the term has very limited possibilities of occurrence: expressions like "two sakes", "part but not all of my sake" are ill-formed. Thus, formalization that is not concerned with a particular philosophical agenda will plausibly not treat "sake" as an independent unit, but will so treat "number", or "table".

Williams also maintains that his view is unlikely to end up being analogous to traditional grounding frameworks like Schaffer's or Fine's, as he is skeptical that he can have a sense of "grounding" that does not relate entities immediately down to the fundamental, whereas theorists like Schaffer and Fine can posit whole chains of grounding. Still, this would arguably result in a two-level ontology, rather than a flat one. Finally, Williams claims that, whereas Schaffer's and Fine's views start from a primitive relation of grounding, his own primitive is linguistic: the reality-requirements of a sentence. However, it seems plausible that different theoretical paths can converge on similar outcomes. Schaffer and Fine may end up postulating a level (or multiple levels) of nonfundamental worldly items via a non-linguistic, explicitly metaphysical route, but this does not mean that Williams' framework avoids postulating such levels.

The problematic connection between perspicuity and metaphysical fundamentality seems to also emerge in certain passages of Rubenstein's paper, where the predicate "non-perspicuous" is applied not to truths (sentences or propositions), but to the existence of non-fundamental entities. In several passages, Rubenstein states that, on the reductionist approach, nonfundamentalia exist only in a "non-perspicuous sense" (2024, pp. 1120-21).

This terminological confusion underscores the issue noted above: in order to make non-perspicuous sentences true, we end up positing things that do not *really* exist, thereby introducing a worldly asymmetry between things that merely exist and things that really exist.

Another potential issue for Correspondence is that it seems ill-suited to account for differences in *degrees* of perspicuity. Suppose the sentence S "Something is grue" ($\exists xGx$) is true but not perspicuous, because "grue" lacks a corresponding worldly item—that is, there is no such property as grue-ness. The sentence S' "Something is grue or bleen" ($\exists x(Gx \vee Bx)$) is, arguably, even less perspicuous, as the predicate "grue or bleen" seems even more "gerrymandered" than the predicate "grue". Yet, if perspicuity is accounted for in terms of the constituents of a truth not "matching" any worldly constituents, the predicate "grue" and the predicate "grue or bleen" are equally non-perspicuous, as neither of them refers to any property in a sparse ontology. Thus, Correspondence fails to account for the intuitive difference in perspicuity between S and S'.

§2.3 Perspicuity as Joint-Carvingness

Joint-Carvingness defines perspicuity as a matter of the constituents of a truth being joint-carving. Following Sider (2011), Rubenstein specifies that whether (and to what extent) a term is joint-carving depends on whether (and to what extent) its referent is natural:

> […] 'is grue' and 'is taller than' do not denote perfectly natural properties/relations; hence <this emerald is grue> and <Trump is taller than Obama> are not perspicuous truths. On the other hand, perhaps 'is negatively

> charged' does denote a perfectly natural property; hence <Sparky is negatively charged> may be perspicuous (insofar as the name 'Sparky' is itself structural!).
>
> (Rubenstein 2024, p. 1113)

However, this account of perspicuity rules out the "flat", one-level ontology required by representational grounding, as it commits us to at least two levels of reality: one comprising the perfectly natural worldly items and the other comprising everything else. Moreover, it is plausible to allow for a notion of relative naturalness—Sider himself acknowledges the need for a comparative notion that accommodates varying degrees of naturalness (2011, p. 129), and Lewis' notion of naturalness is likewise comparative. If perspicuity also admits of degrees, we are thus once again faced with a full-fledged worldly hierarchy, with perfectly natural items at the foundation and multiple levels of progressively less natural items above.

Proponents of naturalness often treat "fundamental" as interchangeable with "perfectly natural" (e.g. Lewis 2009). As Dorr and Hawthorne put it, "it is far from clear what point there would be in distinguishing the question whether the property of being F is perfectly natural from the question whether F-ness is fundamental, or whether it is (or could be) true in reality that things are F" (Dorr & Hawthorne 2013, p. 72). Still, the relationship between grounding, fundamentality, and naturalness is not straightforward. For example, a widely accepted thesis in the grounding literature holds that existential generalizations are partially grounded by each of their instances (Fine 2012, Rosen 2010). Suppose a piece of clay x has a highly specific, complex shape-property S. On this view, the fact that x *has some shape or other* is partially grounded by the fact that it has S. Yet, the idea that S is more natural than the property of having some shape or other may sit uneasily with other roles associated with naturalness—such as its connection to similarity (Dorr 2024). Thus, if we

understand fundamentality in terms of grounding—e.g. F is more fundamental than G if F grounds G—then naturalness and fundamentality may come apart.

That said, the present proposal does not rely on a perfect alignment between naturalness and fundamentality. Even if the roles of naturalness and fundamentality do not fully overlap, the distinction between natural and non-natural items still introduces a metaphysical hierarchy. Like fundamentality, naturalness is a property of worldly items—properties, entities, and (derivatively) facts—and generates a ranking among them. But the very aim of reduction is to avoid positing such worldly hierarchies, allowing only representational ones. A commitment to naturalness thus risks reintroducing exactly what Independence is meant to rule out: naturalness groups things together on the basis of similarity, but the idea is precisely that some groupings are objectively privileged. We can group things together by *grue* or by *green*—but the claim that one grouping is objectively superior introduces a hierarchy among the corresponding properties (in this case, grue and green) based on their degree of naturalness. Natural properties, even if not fully coextensive with fundamental ones, remain metaphysically privileged—and this kind of privilege reintroduces the very sort of hierarchy that Independence seeks to avoid. Indeed, Tahko (2023) acknowledges that "naturalness is no doubt a close cousin of fundamentality", and Schaffer (2009, p. 353) explicitly includes Lewis's view among hierarchical metaphysical frameworks.

In addition to the two previously discussed notions of perspicuity, Rubenstein briefly mentions a related but distinct alternative: the idea that, in order for a truth to be perspicuous, its constituents must be "metaphysically primitive," meaning they lack real definitions. Notably, real definitions are *worldly* definitions—things, not words (or concepts), have real definitions. This proposal thus entails, first, that the constituents of a truth are worldly items, which can only hold under a Russellian account of propositions. As explained, however,

Russellian propositions may be too coarse-grained to serve as the relata of representational grounding. Secondly, the asymmetry in representational fundamentality—or perspicuity—between the relevant truths would clearly stem from a metaphysical asymmetry between the corresponding worldly items. Real definitions differ from identities precisely because the right-hand side and left-hand side are taken to be asymmetrical in a metaphysical, non-merely epistemic sense. This account, therefore, would also violate *Independence*.

§2.4 Perspicuity as Structure-Matching, Simplicity, or Semantic Priority

On a related proposal, an expression's perspicuity depends on how well the expression matches, or reflects, the structure of its referent. This idea could help explain why, for example, an expression like "configuration of particles arranged table-wise" may be more perspicuous than the simple term "table", as it reveals the composite structure of its referent. Likewise, "C-fibers firing" may be regarded as more perspicuous than "pain" because the constituents of the expression "C-fibers firing" pick out the constituents of the relevant property. This general idea is captured by the following definition of perspicuity:

Structure-Matching**:** <p> is more perspicuous than <q> iff <p>'s structure reflects the structure of its referent better than <q>'s.

Jones (forthcoming) proposes a related view. He observes that some entities admit of multiple decompositions into constituents. For example, I can be decomposed into (a) a top half and a bottom half; (b) a left hand, a right hand, and the rest of me; (c) various biological systems, such as the skeletal, nervous, and circulatory systems. These decompositions, Jones argues, are not all equal. Some are more privileged than others: "My decompositions (a) and

(b) are relatively superficial: they provide little information about my underlying nature and play little role in explaining my behaviour. Decomposition (c) does better: it provides substantive information about what kind of thing I am and how I interact with other things" (Jones forthcoming, pp. 17-18).

Jones extends this idea from concrete individuals to other kinds of entities, including properties and propositions. For example, he argues, a single proposition may have many decompositions into constituents. The proposition that Romeo loves Juliet might decompose into the constituents (a) loves, Romeo, and Juliet; (b) loves Juliet, and Romeo; or (c) Romeo loves, and Juliet. Once again, according to Jones, these decompositions are not all equal: decompositions (b) and (c) may be relatively superficial, in the sense that they may not carve the proposition along its underlying joints. By contrast, decomposition (a) may go deeper, in that—Jones claims—it seems to capture the proposition's underlying relational nature, and thereby explains why the proposition exists and other central facts about it. Jones refers to the "deep structure" of a proposition as a decomposition that is maximally privileged in this way. Different sentences expressing a proposition may capture different decompositions of it. A sentence captures a decomposition by having simple syntactic constituents that denote the constituents according to the decomposition. When a sentence captures a decomposition that is the deep structure of the proposition it expresses, the sentence is a "structurally fundamental"—or, in our terms, perspicuous—representation of that proposition. Other representations of the proposition are structurally superficial.

However, this proposal seems to ultimately reduce to the ones previously discussed and is thus incompatible with Independence. If descriptions of me in terms of my top and bottom half and in terms of my skeletal, nervous and circulatory systems are both true, then expressions like "my top half" and "my bottom half" must refer to something—otherwise, the difficulties raised by Correspondence resurface. If they do refer, then there must be such

things as my top and bottom half. On this view, the reason why a description of me in terms of my top and bottom half, despite being true, is not particularly perspicuous, is that the objects it invokes lack metaphysical privilege. In other words, they are not natural. Similarly, the view entails that a sentence whose constituent predicate is "loves" is more "structurally fundamental", or perspicuous, than a sentence whose constituent predicate is "loves Juliet" because the corresponding propositional constituent—the property of loving—is more natural than the property of loving *Juliet.* Jones himself suggests that deep structure can be defined in terms of joint-carvingness: every decomposition of an entity into joints is a deep structure of that entity. That is to say that every decomposition of an entity into natural components is a deep structure of that entity. But then, if an expression's perspicuity is a matter of its constituents denoting the constituents of the deep structure of its referent, then an expression's perspicuity is a matter of its constituents denoting natural worldly items—as in the Joint-Carvingness account discussed earlier. This brings us back, once again, to a hierarchy of worldly entities—objects, properties, relations—ranked according to their level of metaphysical privilege. Importantly, this objection concerns the compatibility of the Structure-Matching account with reduction as defined above, in particular with the flat reality motivation for Independence. Since flat reality is not part of Jones' own proposal, this objection does not target Jones (forthcoming).

Another problem with the Structure-Matching account is that it seems applicable only to a limited class of cases. For example, it is unclear why, on this view, paradigmatically perspicuous terms such as "mass" or "charge" should count as more perspicuous than others, as it is not clear in what sense these terms reflect, or match, the structure of their referents. Suppose we introduce the term "mass*" by stipulating that whenever an object's mass is n,

its mass* is $\sqrt[17]{n}$, i.e. that mass*=$\sqrt[17]{mass}$.[46] It follows that mass=mass*^{17}. Now we may ask: which representation of *mass* is more perspicuous—"mass" or "mass*^{17}"? After all, if mass* were more natural than mass, a representation of mass in terms of mass* would, arguably, better capture its structure: it would reveal that an object's mass is ultimately a function of its more natural property mass*. This would make "mass*^{17}" a more perspicuous representation of mass than "mass". Thus, to justify the claim that "mass" is more perspicuous than "mass*^{17}" we must first establish that mass* is less natural than mass. But this, again, requires settling the question of relative naturalness *first*. Therefore, Structure-Matching does not satisfy Independence.

Another natural thought is that the perspicuity of a truth may have to do with its simplicity. This idea seems plausible in cases involving logically complex predicates. For example, <The ball is red> seems more perspicuous than <The ball is red or red and round>, even though the two truths have the same truth conditions. Similarly, the simple truth <Socrates is wise> seems more perspicuous than the complex truth <The member of {Socrates} is wise>, just as <Snow is white> seems more perspicuous than <The proposition that snow is white is true>.

However, this idea does not seem to work in other cases. The predicate "cow", though not particularly perspicuous, is nonetheless logically simple. If we measure the simplicity of a truth by the length of its statement in an *arbitrarily chosen* language, any truth whatsoever can be made simple. As Lewis (1983) observed, further restrictions are needed. According to Lewis, we need to measure the simplicity of a claim by how easily stateable it is in a language where all predicates denote *perfectly natural* properties (Lewis 1983, Dorr & Hawthorne 2013). This explains why "grue" is non-perspicuous: its definition in natural

[46] See Dorr (2013).

terms is a logically complex one. Thus, we may define perspicuity in terms of simplicity as follows:

Simplicity**:** <p> is more perspicuous than <q> iff the definition of <q>'s constituents in terms denoting perfectly natural properties is longer (more complex) than the definition of <p>'s constituents.

However, as Lewis acknowledges, this approach requires not only a distinction between more and less natural properties—or at least between natural and non-natural properties—but also that natural properties be individuated *prior* to assessing the simplicity of the claims in question. This means that we cannot define perspicuity in terms of simplicity without first establishing a metaphysical distinction between natural and non-natural properties.

Consider the following example. "Grue" is defined as green and observed before t or blue and observed after t. Thus, "grue" is not perspicuous because its definition in joint-carving terms is logically complex. Similarly, "bleen" is defined as either blue and observed before t or green and observed after t. However, given the definitions of "grue" and "bleen, "green" can be defined as either grue and observed before t or bleen and observed after t (Goodman 1955)—its definition is equally complex. Why, then, is "green" more perspicuous than "grue" or "bleen"? Because not all definitions are the same: only definitions in terms of natural properties are relevant to Simplicity. But then, to determine whether "green" or "grue" is more perspicuous, we need to *first* establish whether green is more natural than grue. Again, Simplicity does not allow us to build a representational hierarchy without first building a worldly hierarchy.

Similar considerations apply to other notions of priority between terms. For instance, Smithson's (2020) account relies on the idea that certain expressions have inferential roles that are constitutive of the meanings of those expressions—e.g. "vixen" is inferentially linked to "female" and "fox", and this inferential connection is constitutive of its meaning. On this account, a term T1 is "semantically prior" to a term T2 just in case the inferential role for T2 involves T1. Smithson endorses a definition of perspicuity along the lines of:

Semantic Priority: <p> is more perspicuous than <q> iff <p>'s constituents are semantically prior to <q>'s constituents.

Leaving aside concerns about the underlying semantic assumptions, there are at least two major problems with this account. First, in many cases we may want to say that one truth is more perspicuous than another even if the terms involved have no inferential connection whatsoever. For example, <Electrons are negatively charged> is, intuitively, more perspicuous than <This emerald is grue> even though "negatively charged" is not semantically prior to "grue", because it is not part of its inferential role.

Secondly, we have seen that it is not clear how we can rule out the idea that "grue" and "bleen" may be semantically prior to "green" without positing a worldly hierarchy. In a linguistic community where people describe things in terms of "grue" and "bleen", "green" may be introduced as either grue and observed before t or bleen and observed after t. In this scenario, "grue" and "bleen" are part of the inferential role of "green" and thus count as semantically prior to it. The core issue is that there seems to be no ad hoc way of ruling out such uses as less legitimate than ours, unless we posit a difference in naturalness between the relevant properties. Thus, Semantic Priority also fails to satisfy Independence.

In sum, all the notions of perspicuity introduced in this section seem to be in tension with Independence. Without Independence, as said, representational grounding loses most of its appeal as a more metaphysically parsimonious alternative to worldly grounding.

§3 The Relation of Representational Grounding: Entailment

§3.1 Structure: Asymmetry and Irreflexivity

A relation between truths that is intelligible independently of any worldly hierarchy—and thus satisfies Independence—*is* available, namely entailment. In this section, I will first argue that, while entailment arguably holds between the relevant truths, it cannot play the role of representational grounding, as it fails to satisfy Structure. I will then show that even a restricted version of entailment which satisfies Structure is ill-suited to account for representational grounding.

Doubts about worldly grounding often focus on whether we really need a relation of grounding in addition to familiar metaphysical relations between facts, such as supervenience or realization. Analogous doubts about representational grounding may concern whether we do need a *new* relation in addition to the familiar relation of entailment between truths. One might thus wonder whether representational grounding could simply be accounted for in terms of entailment. Here, "entailment" is understood as strict implication, meaning that <p> entails <q> just in case in every world where <p> is true, <q> is true as well. On this proposal, relative representational fundamentality may be defined as follows:

Entailment**:** <p> is representationally more fundamental than <q> iff <p> entails <q>.

Entailment satisfies Independence: the claim that <p> entails <q> does not require us to posit worldly grounding between the corresponding facts [p] and [q]. In fact, the claim that <p> entails <q> is often inconsistent with the idea that [p] grounds [q]: <Tibbles is a cat> entails <Tibbles is a cat>, but [Tibbles is a cat] does not "give rise to" [Tibbles is a cat], because worldly grounding is irreflexive.

Moreover, consider the truths <London is north of Paris> and <Paris is south of London>. Arguably, these truths entail each other: <London is north of Paris> entails <Paris is south of London> and vice versa. However, even assuming that [London is north of Paris] and [Paris is south of London] were distinct facts, neither of them would "give rise" to the other, as there would be no difference in metaphysical "priority" between them. Most importantly, it would be impossible for [London is north of Paris] to ground [Paris is south of London] *and* for [Paris is south of London] to ground [London is north of Paris], because grounding is asymmetric. Similarly, <The number of planets is 4+4> entails <The number of planets is 6+2> (again, regardless of whether the relevant propositions are identical), but it just sounds wrong to say that [the number of planets is 4+4] *gives rise to* (or *is more fundamental than*) [the number of planets is 6+2], even if these facts are considered as distinct.

In sum, the claim that <p> entails <q> does not entail that [p] grounds [q]. Entailment is independent of any worldly hierarchy. Entailment, qua relation between truths, might be taken to correspond at the metaphysical level to a relation of supervenience between facts, although there may still be mismatches between entailment and supervenience.[47] Setting this

[47] Suppose supervenience is defined as follows: a fact [F] supervenes on a fact [G] iff any worlds that are G-indiscernible are F-indiscernible. Intuitively, the idea is that there can be no difference in F-related aspects without a difference in G-related aspects. Consider a world *v* where a fact [B] obtains. Because [B] obtains, the

issue aside, Entailment is quickly ruled out by considerations concerning the structure of the relation of representational grounding. Like worldly grounding, representational grounding is irreflexive and asymmetric, while entailment is, as shown by the examples above, reflexive, and thus non-asymmetric. Therefore, Entailment fails to satisfy Structure.

§3.2 Three Arguments Against Representational Grounding as "One-Way" Entailment

Proponents of representational grounding might try to obtain the required asymmetry and irreflexivity by placing restrictions on the relevant entailment relations. In particular, the relation between representational levels may be modeled as "one-way" entailment—i.e. asymmetric entailment between different truths. On this proposal:

One-Way Entailment**:** <p> is representationally more fundamental than <q> iff <p> entails <q> and <q> does not entail <p>.[48]

This proposal may seem effective in some of the usual cases. For example, suppose "C" describes the *specific* microphysical configuration of Tibbles the cat. Arguably, <Tibbles is C> one-way entails <Tibbles is a cat>: given that anything that is C is a cat, <Tibbles is C> entails <Tibbles is a cat>. Moreover, since being a cat is multiply realized by various

disjunctive fact [B ∨ C] obtains at *v*, although [C] does not obtain at *v*. Now consider another world *u* where neither [B] nor [C] obtains. Of course, [B ∨ C] does not obtain at *u*. But then, *u* differs from *v* as to whether [B ∨ C] obtains, but does not differ from *v* in whether [C] obtains. This means that, by the definition above, [B ∨ C] does not supervene on [C]. However, the truth that C is the case, <C>, entails the truth that either B or C is the case, < B ∨ C>. See also McLaughlin & Bennett (2023).

[48] This entails that <p> is not identical to <q>, as all asymmetrical relations are irreflexive.

microphysical configurations, not every cat is C. Intuitively, Tibbles could have been a cat without being C. Thus, <Tibbles is a cat> does not entail <Tibbles is C>. One-Way entailment thus seems to capture the difference in representational fundamentality between these two truths: <Tibbles is C> seems more perspicuous than <Tibbles is a cat>.

Nevertheless, this proposal introduces new difficulties. In the rest of this section, I outline three problems, each of which is sufficient to reject One-Way Entailment as a viable account of representational grounding.

The first problem is that a relation of one-way entailment can only hold between truths representing distinct facts. Indeed, one-way entailment holds between <p> and <q> iff <p> entails <q> but <q> does not entail <p>. Since, by definition, <p> entails <q> just in case in every world where <p> is true, <q> is true as well, if <p> entails <q> but <q> does not entail <p>, then at every world where <p> holds, <q> holds, but not vice versa. Now consider the corresponding facts. Even in the most coarse-grained account of facts, where facts are defined as sets of worlds, if <p> one-way entails <q>, [p] and [q] correspond to distinct sets of worlds, thereby counting as distinct facts. For example, because they individuate distinct sets of worlds, [Tibbles is C] and [Tibbles is a cat] are distinct facts. Thus, even if <Tibbles is C> one-way entails <Tibbles is a cat>, no relation of representational grounding can hold between them, as representational grounding requires the relevant truths to represent the same fact.

The second problem arises from the fact that any truth entails any necessary truth. If <p> one-way entailing <q> is sufficient for <q> to be less representationally fundamental than <p>, we get that paradigmatically non-perspicuous truths, such as <This emerald is grue>, are more representationally fundamental than intuitively quite perspicuous necessary truths, like <All electrons are electrons>. The relation between these two truths *is* one of one-way entailment, but the account seems to get the (alleged) difference in representational

fundamentality wrong. This issue resembles a similar one in the debate on worldly grounding. The difference between grounding and merely modal relations like supervenience is typically illustrated precisely with examples of this kind (see McLaughlin & Bennett 2023). While supervenience allows for "trivial" connections such that necessary facts supervene on any fact, grounding is intended to capture a stronger and non-merely modal kind of metaphysical dependence between facts.

Unless we are willing to accept that basically any contingent truth—no matter how "gruesome"—is more representationally fundamental than any necessary truth, representational grounding cannot be modeled according to One-Way Entailment. Furthermore, if <p> is an intuitively quite perspicuous truth and <q> is an independent non-perspicuous truth, <p and q> one-way entails <p> but is, arguably, less fundamental than it.

The third issue seems equally fatal. Consider two properties F and G, and suppose that anything that is F is also G. Hence, <o is F> entails <o is G>. Suppose, further, that the relation of entailment between these two truths is one-way entailment—i.e. <o is G> does not entail <o is F> (meaning that <o is F> and <o is G> are distinct truths). According to One-Way Entailment, this means that <o is F> is representationally more fundamental than <o is G>. Since <o is F> and <o is G> differ only with respect to the predicate, "being F" should qualify as representationally more fundamental, or more perspicuous, than "being G". In terms of concepts, this would mean that the concept of F is more representationally fundamental than the concept of G. However, $p \rightarrow q$ entails $\neg q \rightarrow \neg p$. Thus, if <o is F> entails <o is G>, then <o is not G> entails <o is not F>: negation reverses the direction of entailment. The trouble is that, intuitively, negation should not reverse the direction of representational fundamentality: if "being F" is more perspicuous than "being G", arguably "being not F" is *still* more perspicuous than "being not G".

An example: <o is an electron> entails <o is an electron or a cow>. Since we are modelling representational grounding in terms of one-way entailment, we can say that, since there is a relation of one-way entailment between <o is an electron> and <o is an electron or a cow>, <o is an electron> representationally grounds <o is an electron or a cow>. In this framework, this means that <o is an electron> is representationally more fundamental than <o is an electron or a cow>. But the two propositions differ only with respect to the predicate, which means that, plausibly, the predicate "being an electron" is more perspicuous than the predicate "being an electron or a cow". So far so good. However, <o is neither an electron nor a cow> entails <o is not an electron>. Moreover, <o is neither an electron nor a cow> *one-way* entails <o is not an electron>, as <o is not an electron> does not entail <o is neither an electron nor a cow>. Thus, <o is neither an electron nor a cow> representationally grounds <o is not an electron>, meaning that it is more fundamental in the hierarchy of representational levels. However, <o is not an electron> and <o is neither an electron nor a cow> only differ with respect to the predicate. But then, "being neither an electron nor a cow" would need to stand on a more basic representational level than the predicate "not being an electron"—i.e. it would need to be more perspicuous. Yet, this is surely very implausible: "being neither an electron nor a cow" looks like a paradigmatic non-perspicuous predicate. Thus, One-Way Entailment is not a good account of representational grounding.

Although entailment satisfies Independence, there seems to be no way to define relative representational fundamentality in terms of entailment without violating Structure. However, definitions of relative representational fundamentality in terms of other notions fail to satisfy Independence. In sum, no available candidate for representational fundamentality seems to satisfy both Structure and Independence.

§ Conclusion

I have argued that no candidate relation of reduction, or representational grounding, satisfies both Independence and Structure, posing a serious challenge to the reductive approach to metaphysical explanation. Perhaps a new account of perspicuity could be developed that meets these constraints—but that task falls to the proponents of reduction. It may also be that there is simply no coherent way to combine a flat ontology with the idea that some ways of representing reality are *objectively* better than others. If that is the case—if no such relation of representational grounding exists—then we should ask ourselves what explains the perceived asymmetry between the relevant truths. Here, the intensionalist approach to apparent cases of hyperintensionality in explanation contexts offers valuable insights. Consider the following truths:

(7) Snow is white

(8) The proposition that snow is white is true

Most people intuitively judge that the truth of (8) is determined by the truth of (7), but not vice versa: the proposition that snow is white is true *because* snow is white, while snow isn't white *because* the proposition that snow is white is true. There seems to be an objective asymmetry between them. The same intuition arises with:

(9) Socrates is wise

(10) The member of {Socrates} is wise

However, within an intensionalist framework, (7) and (8)—as well as (9) and (10)—in fact express the same proposition, since they correspond to the same set of possible worlds. As a result, intensionalists also face the challenge of explaining why there appears to be an asymmetry between these truths—albeit for different reasons. The intensionalist denies any objective asymmetry between (7) and (8) on the grounds that they express the same proposition. However, one may also reject the idea that there is an objective asymmetry between these truths not because the propositions are identical, but because there is no objective relation of representational grounding that could account for this asymmetry. These two claims are consistent, but distinct. Williamson (2021a, 2024) argues that the appearance of an objective asymmetry between the truth of (7) and the truth of (8) is a projection of the pragmatics of explanation. Consider the following two exchanges:

(11) Q: Why is the proposition that snow is white true?

A: Because snow is white

(12) Q: Why is snow white?

A: Because the proposition that snow is white is true.

Everyone agrees that (11) offers a better explanation than (12), even though both rely on the schematic equivalence of 'p' and 'the proposition that p is true'. In (11), the complex, unfamiliar, and non-obvious "The proposition that snow is white is true" is explained in terms of the simpler, more familiar, and obvious "Snow is white". Helpful explanations often move from the unfamiliar to the familiar, while in (12) the explanation moves in the opposite direction. As Williamson (2024) notes, an explanation can be good or bad because of a

mixture of linguistic and non-linguistic factors, including the order in which the information is presented, the use of common or uncommon vocabulary, sentence complexity, and the hearer's familiarity with the subject matter.

In sum, we may tend to place truths like (7) and (8) on different levels of a representational hierarchy because one appears to offer a more explanatory or perspicuous representation of reality than the other. However, our judgements about whether a given representation is more or less perspicuous than another may not reflect any genuine objective feature of the representations themselves, or of their relations to the reality they describe. Instead, these judgements may be best explained away as the products of a combination of cognitive, linguistic, and pragmatic factors, rather than taken as evidence for a relation of representational fundamentality.

Conclusion

This dissertation has examined a range of metaphysical problems through the framework of Frege puzzles. In doing so, it has sought to show that the semantics of propositional attitude ascriptions plays a central role in the relevant debates, and that the intuitions which animate these debates frequently stem from the presence of terms that generate apparently opaque contexts. Moreover, many proposed solutions to the problems in question amount, in effect, to familiar strategies for resolving Frege puzzles. By drawing out these connections, the dissertation has aimed to offer a unifying framework that brings together seemingly unrelated issues, while also shedding light on the scope and significance of Frege puzzles and the conceptual tools involved in their formulation. More broadly, the aim has been to motivate an anti-exceptionalist stance toward the problems discussed.

§1 Contributions by Chapter

In addition to offering a unifying framework, this dissertation aims to make original contributions to each of the individual debates it engages with. Framing these debates in terms of Frege puzzles advances each discussion in distinct ways. What follows is a summary of the contributions made in each chapter, as well as their connections to broader philosophical questions.

Verbalness and Disagreement

The central claim of the first chapter is that paradigmatic examples of allegedly verbal disputes—including many metaphysical disputes—are neither merely metalinguistic nor devoid of genuine disagreement. Such disputes appear merely verbal only under specific contentious semantic and epistemological assumptions. In fact, they often arise from disagreement over theoretically "heavyweight" issues that cannot be resolved by debating linguistic matters, but instead require substantive theoretical inquiry. For example, when two speakers disagree about whether whales are fish in virtue of disagreeing about what counts as a fish, their disagreement cannot be straightforwardly dismissed as a mere divergence about language. Rather, it may reflect deeper concerns, such as which criteria for speciation yield the most adequate taxonomic model. Similar considerations apply to metaphysical disputes, supporting an anti-deflationist approach to metaphysics. Moreover, the central question of whether the disputants fully agree on the non-linguistic facts—like the issue of belief and knowledge in classic Frege puzzles—depends on the semantics of propositional attitude ascriptions.

The Knowledge Argument and the Problem of Consciousness

The second chapter highlights how, within a physicalist framework, the scenario presented in Jackson's Knowledge Argument can be treated as a Frege puzzle. On this view—provided we can formulate a satisfactory account of phenomenal modes of presentation, for example in terms of acquaintance—the Knowledge Argument does not pose a distinctive epistemic challenge to physicalism. Its force wanes once the scenario is viewed as an instance of Frege's cases: the intuitions it elicits stem from the apparent opacity of knowledge ascriptions, rather than from any unique feature of consciousness. This reduction helps to "deflate" the problem of consciousness by tracing it back it to a familiar phenomenon, thereby supporting an anti-exceptionalist stance towards consciousness, even from an

epistemic perspective. Whether Mary can be said to acquire new knowledge upon leaving her black-and-white room ultimately turns, much like in Frege puzzles, on the semantics of propositional attitude ascriptions.

Representational Grounding and Perspicuity

The third chapter raises a further issue concerning the relationship between representation and reality: it is difficult to reconcile a parsimonious "flat" ontology—one that rejects metaphysical hierarchies among objects, facts, or properties—with the idea that some representations of reality are *objectively* more accurate or perspicuous than others. In short, any attempt to establish an objective measure of perspicuity tends to reintroduce hierarchies into reality itself. This observation is not only of general philosophical significance but also poses a specific challenge within the debate on grounding and metaphysical explanation. In particular, it casts doubt on the view that a representational account of grounding—according to which grounding connects more and less perspicuous truths, rather than more and less fundamental facts—provides a viable alternative to "worldly" grounding. Instead, our judgements about relative perspicuity may be better understood as byproducts of cognitive and pragmatic factors, especially the pragmatic dimensions of explanation. In other words, such judgements can be explained away through the kind of error theories that anti-Fregeans provide for our intuitions in apparently opaque explanation contexts.

§2 Issues for Further Research

The discussion in the three main chapters brings to light several issues that invite further investigation—both within the scope of each individual chapter and in the broader

application of the Frege puzzle framework to a range of philosophical debates. While this dissertation has sought to address a number of core questions, it is clear that many avenues remain open for deeper inquiry. Further research could enhance our understanding of these issues and help extend the applicability of the Frege puzzle approach to new philosophical challenges. What follows is an outline of areas where continued exploration may prove especially fruitful, beginning with specific topics from the three chapters, and then moving on to broader themes that arise from generalizing Frege puzzles across philosophical debates.

§2.1 Specific Topics for Further Research

Agreement on the Facts, and the Relation between Metalinguistic Beliefs and Identity Beliefs

In Chapter 1, the issue of complete agreement is framed in semantic terms, on the assumption that the objects of agreement and disagreement are propositions. However, some theorists hold that verbal disputes involve agreement on all the relevant non-metalinguistic *facts*. While there are general reasons to believe that the objects of agreement and disagreement are, in fact, questions—understood, on a standard account, as sets of propositions—one might instead wish to formulate the issue in terms of facts. On this formulation, the problem of complete agreement arises because, if being F *just is* being G, then o's being F is the same *fact* as o's being G (see, e.g., Rayo 2013). Accordingly, if the disputants agree that o is G, they might be said to agree on all the facts. This version of the problem requires us to consider different accounts of facts, rather than the semantic frameworks discussed in Chapter 1. Although this issue lies beyond the scope of this thesis, it remains an interesting topic for future research.

Moreover, the relationship between metalinguistic beliefs and beliefs about identity—particularly the question of whether one can believe an identity claim without

forming any corresponding metalinguistic belief—highlights an interesting connection to research on cognitive aspects related to language use and linguistic competence. These issues also touch on important questions about how language is acquired, and the role that implicit or explicit metalinguistic knowledge might play in language acquisition. While Chapter 1 only briefly touches on these issues, a fuller account of metalinguistic disagreement and disagreement about identity would require a more thorough exploration of these questions.

The Context-Sensitivity of Knowledge-What and Phenomenal Modes of Presentation

Chapter 2 invites further work on several issues, including the context-sensitivity of knowledge-wh attributions. This could help determine whether Mary can be said to know what it's like to see red merely by knowing the relevant proposition under a physical mode of presentation, or whether such knowledge requires knowing this proposition under a *phenomenal* mode of presentation. More broadly, a comparison between the context-sensitivity of knowledge-what and that of other forms of knowledge-wh—such as knowledge-who—may offer further insight. Another central topic in Chapter 2 concerns phenomenal modes of presentation. As noted, there is little consensus regarding the nature of phenomenal concepts or modes of presentation. Moreover, as Levine (2007) has emphasized, physicalist accounts of phenomenal modes of presentation must satisfy what he calls the "materialist constraint": they must be formulated in terms that themselves admit of a physicalist account. The plausibility of treating the Knowledge Argument as a Frege puzzle hinges on the availability of a coherent and suitably constrained physicalist account of phenomenal modes of presentation.

How the Pragmatics of Explanation shapes our Judgements about Metaphysical Dependence

Finally, the topic of Chapter 3 is closely connected to the issue of opacity in explanation contexts. In such contexts, co-intensional expressions do not appear to be substitutable *salva veritate*: for instance, "The member of {Socrates} is wise *because* Socrates is wise" seems true, whereas "Socrates is wise because the member of {Socrates} is wise" seems false—even though "Socrates" and "the member of {Socrates}" are co-intensional. This apparent asymmetry raises important questions about whether claims of metaphysical explanation—which supposedly reflect objective relations between facts—are in fact sensitive to how those facts are represented. Judgments about metaphysical explanation may be shaped by cognitive, linguistic, and pragmatic factors, rather than tracking objective metaphysical relations (Williamson 2024). It is therefore important to investigate how the pragmatics of explanation influences our intuitions about metaphysical explanation. Building on work in cognitive science and the philosophy of science (e.g. Van Fraassen 1980), further research could examine the extent to which context-sensitivity and interest-relativity in explanation contribute to our judgements concerning metaphysical dependence.

§2.2 General Topics for Further Research

Besides the specific issues addressed in each chapter, the broader application of the Frege puzzle framework to philosophical problems opens up a range of more general questions for further research.

Error Theories and Heuristics

As we have seen, anti-Fregeanism—particularly in its anti-contextualist form—sits uneasily with our intuitive judgements about meaning and attitude ascriptions. Proponents of this view typically respond to these tensions not by revising their semantic theories or accounts

of propositional attitude ascriptions to better align with our intuitions, but by offering error theories to explain why our intuitions go wrong.

Some theorists (e.g. Salmon 1986) have appealed to pragmatics, drawing on the Gricean notion of conversational implicature (Grice 1989). According to this view, certain propositional attitude ascriptions are avoided in practice because they give rise to false implicatures. For example, if Jane has never encountered the name "Phosphorus", uttering "Jane believes that Hesperus is Phosphorus" might misleadingly suggest that Jane would assent to the sentence "Hesperus is Phosphorus", even if the belief ascription itself expresses a true proposition. This, it is argued, accounts for our reluctance to endorse such reports.

However, this pragmatic explanation faces a serious difficulty: speakers do not merely refrain from asserting the relevant belief reports (in this case, "Jane believes that Hesperus is Phosphorus")—they regard them as false. But this is not the typical effect of false implicatures. Generally, when an utterance generates a false implicature, we regard it as misleading, not as false. To borrow an example from Williamson (2021a), if saying "The Professor is sober this morning" falsely implicates that the Professor is usually drunk, we do not then assert "The Professor is *not* sober this morning". Thus, further work is needed on error theories that aim to explain our intuitions in Frege cases involving propositional attitudes.

Williamson's recent account (2021a, 2024) in terms of heuristics offers a promising direction. Williamson argues that our intuitions in these cases are shaped by our ordinary heuristics for attitude ascriptions. In particular, in the case of belief ascriptions, we seem to rely on what Williamson calls the "Just Ask" heuristics, which involves asking "Do you believe that P?" as a method for determining whether someone believes the relevant proposition. Because this heuristic is language-sensitive—i.e. sensitive to differences between sentences, even if the propositions expressed are identical—it can lead to systematic

errors in Frege cases. For example, if one fails to assent to a sentence like "Hesperus is Phosphorus" we may be misled into thinking they do not believe the proposition it expresses, even if they do under a different guise.

Heuristics like *Just Ask* are typically reliable enough for everyday purposes, but they are not infallible. As Williamson emphasizes, philosophical inquiry often treats our intuitions about hypothetical cases as central data. However, if those intuitions are the products of imperfectly reliable heuristics, then they may be mistaken, and theories should not be expected to predict their correctness in every instance. In some cases, the role of the theory should shift: rather than accommodating these intuitions at all costs, we should seek to understand their origins and assess whether they reflect genuine insight or cognitive error.

Indeed, according to Williamson, Fregean semantic theories may have fallen into the trap of overfitting, elaborating increasingly complex theories to fit unreliable data. A better approach would aim to understand how our heuristics operate, treating them charitably where possible, but recognizing that they may sometimes lead us astray.

This raises an important methodological question: how can we distinguish between the reliable and unreliable outputs of our heuristics? Determining when our heuristics lead us astray is far from straightforward. In other words, we need a method for distinguishing genuine errors from trustworthy data. In perception, we can often verify whether a particular judgement is mistaken through independent means. For example, our perceptual heuristics can produce visual illusions, such as the Müller-Lyer illusion. Yet, in this case, we can simply measure the lines to confirm that they are, in fact, equal in length. In philosophy, by contrast, it is far less clear whether comparable independent checks are available. While a systematic and generally applicable method of verification may be out of reach, even case-by-case evaluation may prove difficult in some cases. In sum, further research—both in

cognitive science and philosophy—is needed to better understand the role of heuristics in philosophical inquiry.

Methodological Issues and Overfitting

This dissertation remains neutral on which approach to Frege puzzles should ultimately be adopted and, accordingly, refrains from taking a definitive stance on the specific problems discussed across the three chapters—for instance, whether the parties to factual identity disputes agree on all the relevant propositions, or whether Mary gains new knowledge upon leaving her black-and-white room.

Taking a stance on Frege puzzles—and thereby on these questions—would require further discussion, particularly of a methodological nature. Fregean and anti-Fregean approaches to Frege puzzles can be viewed as prioritizing different aspects of a philosophical model of the relevant phenomena. On the one hand, Fregean theories are typically motivated by the idea that our models should vindicate the intuitions prompted by the cases under consideration. Frege's own distinction between sense and reference was driven by the intuition that one can believe that Hesperus is bright without believing that Phosphorus is, or that sentences like "Hesperus is Phosphorus" and "Hesperus is Hesperus" differ in informativeness—and therefore in meaning.

On the other hand, as noted, anti-Fregean theorists tend to be more cautious about relying on intuitions. The issue is not that they reject the use of intuitions as data in philosophical theorizing, but rather that they are more sensitive to the possibility that intuitions may be systematically unreliable in certain cases. Williamson (2024), for instance, argues that complicating a theory—in this case, a semantic framework or an account of propositional attitude ascription—in order to accommodate unreliable intuitions may lead to overfitting. A symptom of overfitting in philosophy is precisely the proliferation of

conditions or distinctions that serve only to capture intuitions elicited by exceptional or marginal cases, thereby increasing a theory's complexity without improving its explanatory power. In the cases at issue, overfitting often involves introducing ad hoc restrictions to otherwise general principles (such as Leibniz's Law), or adding ad hoc clauses—for example, the claim that the reference of a term shifts to its customary sense within attitude contexts, or that propositional attitude verbs are context-sensitive. The notion itself of an additional layer of meaning playing the role of Fregean sense—such as primary intension in two-dimensional Fregean frameworks—can be seen as introducing poorly regimented elements into the theory, effectively adding degrees of freedom.

In this sense, the choice between Fregean and anti-Fregean approaches appears to be fundamentally methodological. Ideally, we would have a simple and well-regimented framework which also validates our intuitions. In the absence of such an ideal theory, however, we face a trade-off—one that can only be resolved through methodological reflection.

Modes of Presentation

Modes of presentation play a central role in Frege puzzles, as they are posited in both Fregean and anti-Fregean frameworks. The key difference between these frameworks lies in whether they assign modes of presentation a semantic role, as opposed to a merely cognitive one.

Unsurprisingly, there is no consensus on what modes of presentation are, or on what their identity conditions might be. Some accounts essentially equate them with concepts, but analogous questions concerning identity conditions and semantic role arise for concepts as well. As noted, Fregean accounts characterize modes of presentation as senses—or in terms

of some closely related semantic notion—understood in either descriptive or non-descriptive terms depending on the specific version of Fregeanism.

Sometimes, modes of presentation are identified with *syntactic* items such as terms or sentences, particularly in anti-Fregean frameworks. The modes of presentation of a given proposition p, for example, may be taken to be the various sentences that express p—hence the label "sentential guise". Similarly, terms are sometimes taken to be modes of presentation of their referents: for instance, "Hesperus" and "Phosphorus" may count as distinct modes of presentation of Venus simply in virtue of being distinct terms that refer to it. This approach draws on the idea that differences in modes of presentation should reflect differences in cognitive significance, and such differences can arise merely from syntactic variation. Insofar as distinctions in cognitive significance are fine-grained enough to track syntactic differences, it is plausible to identify modes of presentation with linguistic expressions.

However, this metalinguistic account of modes of presentation fails in other cases. Kripke's Paderewski example (Kripke 1979), in particular, shows that even distinct tokens of the same term can differ in cognitive significance. If the role of modes of presentation is to account for such differences, this suggests that modes of presentation cannot *always* be identified with linguistic guises. Indeed, Frege puzzles do not seem to be an essentially linguistic phenomenon. As Salmon (1986) observes, they can be understood as cases of failure to recognize objects, properties, or propositions. While one common source of such failures involves adopting a "language-sensitive" approach to content individuation—where syntactic differences give rise to perceived semantic differences (Williamson 2021a)—these failures of recognition need not involve language at all. Non-human animals and young children, for instance, often fail to recognize things, and Frege-type cases can arise for them as well. A dog, for example, may fail to recognize himself in a mirror, forming the belief that he (under a visual demonstrative guise) is a rival, without believing that he (under a self-

relating guise) is a rival. Clearly, this kind of failure of recognition does not involve language or assent to contextually relevant sentences, yet it exhibits the core structure of paradigmatic Frege cases. This suggests that modes of presentation cannot be identified with natural language sentences or their meanings, but should instead be understood in terms of a more general notion—such as mental states or mental representations (see Braun 1998, Soames 1995, Salmon 1986).

Various philosophical problems call for non-linguistic kinds of modes of presentation. One example is that of phenomenal modes of presentation of properties such as pain. While such modes of presentation are not necessarily *sui generis*—they are often reduced to other types of modes of presentation, including demonstrative, indexical, or recognitional ones—they are nonetheless non-linguistic in nature. Similarly, a difference in cognitive significance between two uses of "that", which may unknowingly refer to the same object, can often be explained only in perceptual terms—that is, by appeal to distinct perceptual modes of presentation. Intellectualist accounts of practical knowledge (knowing how), which treat it as a species of propositional knowledge (knowing that), also posit practical modes of presentation of propositions. Finally, as noted in the introduction, distinctive challenges arise in connection with indexical modes of presentation, particularly within Fregean frameworks.

In sum, modes of presentation are a key component in our understanding of Frege puzzles, and further work is certainly needed to develop a general account capable of accommodating the various kinds of modes of presentation mentioned above. Such an account is also likely to impact other aspects of our treatment of Frege cases, including the formulation of error theories.

Hopefully, this discussion has highlighted the complexity and nuance of Frege puzzles, emphasizing the need for a broader and more flexible theoretical framework.

Further research into these issues promises to open up new avenues for addressing long-standing philosophical challenges.

References

Abreu, Pedro (2023) “Metalinguistic Negotiation, Speaker Error, and Charity”, *Topoi* 42 (4):1001-1016.

Armstrong, David (1978) *A Theory of Universals. Universals and Scientific Realism Volume II*, Cambridge University Press.

Armstrong, David (2002) “Truthmakers for Modal Truths”, in Lillehammer & Rodriguez-Pereyra (eds.), *Real Metaphysics: Essays in Honour of D. H. Mellor, With His Replies,* Routledge.

Audi, Paul (2012) “Grounding: Toward a Theory of the In-Virtue-Of Relation”, *Journal of Philosophy* 109 (12):685-711.

Balcerak Jackson, Brendan (2014) “Verbal Disputes and Substantiveness”, *Erkenntnis* 79 (1):31-54.

Bennett, Karen (2009) “Composition, Colocation, and Metaontology”, in Wasserman, Manley & Chalmers (eds.), *Metametaphysics: New Essays on the Foundations of Ontology*, Oxford University Press.

Bliss, Ricki (2013) “Viciousness and the Structure of Reality”, *Philosophical Studies* 166 (2):399-418.

Block, Ned (1996) “Conceptual Role Semantics”, in Craig (ed.), *Routledge Encyclopedia of Philosophy,* Routledge.

Block, Ned (2007) “Max Black’s Objection to Mind-Body Identity”, *Oxford Studies in Metaphysics* 2:3-78.

Boër, Steven & Lycan, William (1975) “Knowing Who”, *Philosophical Studies* 28 (5):299 - 344.

Boghossian, Paul (1996) “Analyticity Reconsidered”, *Noûs* 30 (3):360-391.

Bonardi, Paolo (2020) “Frege’s Puzzle and Cognitive Relationism: An Essay on Mental Files and Coordination”, *Disputatio* 12 (56):1-40.

Braun, David (1998) “Understanding Belief Reports”, *Philosophical Review* 107 (4):555-595.

Burge, Tyler (1979) “Individualism and the Mental”, *Midwest Studies in Philosophy* 4 (1):73-122.

Burgess, Alexis & Plunkett, David (2013) “Conceptual Ethics I”, *Philosophy Compass* 8 (12):1091-1101.

Cameron, Ross (2008) “Truthmakers and Ontological Commitment: or How to Deal with Complex Objects and Mathematical Ontology Without Getting into Trouble”, *Philosophical Studies* 140 (1):1-18.

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

Carroll, Lewis (1895) “What the Tortoise Said to Achilles”, *Mind* 4 (14):278-280.

Cath, Yuri (2009) “The Ability Hypothesis and the New Knowledge-How”, *Noûs* 43 (1):137-156.

Chalmers, David (1996) *The Conscious Mind: In Search of a Fundamental Theory (2nd edition)*, Oxford University Press.

Chalmers, David (2002) “On Sense and Intension”, *Philosophical Perspectives* 16:135-82.

Chalmers, David (2009) “The Two-Dimensional Argument Against Materialism”, in McLaughlin & Walter (eds.), *Oxford Handbook to the Philosophy of Mind*, Oxford University Press.

Chalmers, David (2011a) “Verbal Disputes”, *Philosophical Review* 120 (4):515-566.

Chalmers, David (2011b) “Propositions and Attitude Ascriptions: A Fregean Account”, *Noûs* 45 (4):595-639.

Clark, Michael J. (2018) “What Grounds What Grounds What”, *Philosophical Quarterly* 68 (270):38-59.

Conee, Earl (1994) “Phenomenal Knowledge”, *Australasian Journal of Philosophy* 72 (2): 136-150.

Correia, Fabrice (2010), “Grounding and Truth-Functions”, *Logique Et Analyse* 53 (211):251-279.

Correia, Fabrice (2017) “Real Definitions”, *Philosophical Issues* 27 (1):52-73.

Correia, Fabrice & Schnieder, Benjamin (eds.) (2012) *Metaphysical Grounding: Understanding the Structure of Reality*, Cambridge University Press.

Correia, Fabrice & Skiles, Alexander (2017) “Grounding, Essence, and Identity”, *Philosophy and Phenomenological Research* 98 (3):642-670.

Crimmins, Mark & Perry, John (1989) “The Prince and the Phone Booth: Reporting Puzzling Beliefs”, *Journal of Philosophy* 86 (12):685.

Daly, Chris (2012) “Scepticism about Grounding”, in Correia & Schnieder (eds.), *Metaphysical Grounding: Understanding the Structure of Reality*, Cambridge University Press.

Díaz-León, Esa (2016) “Phenomenal Concepts: Neither Circular nor Opaque”, *Philosophical Psychology* 29 (8):1186-1199.

Dorr, Cian (2013) “Reading ‘Writing the Book of the World’”, *Philosophy and Phenomenological Research* 87 (3):717-724.

Dorr, Cian (2016) “To be F is to be G”, *Philosophical Perspectives* 30 (1):39-134.

Dorr, Cian (2024) “Natural Properties”, in Zalta & Nodelman (eds.), *The Stanford Encyclopedia of Philosophy,* https://plato.stanford.edu/archives/sum2024/entries/natural-properties/.

Dorr, Cian & Hawthorne, John (2013) “Naturalness”, in *Oxford Studies in Metaphysics 8*.

Dummett, Michael (1973) *Frege: Philosophy of Language*, London: Duckworth.

Evans, Gareth (1982) *The Varieties of Reference*, Oxford University Press.

Evans, Gareth (1985) *Collected Papers*, Oxford University Press.

Fine, Kit (1994) “Essence and Modality”, *Philosophical Perspectives* 8:1-16.

Fine, Kit (2001) “The Question of Realism”, *Philosophers' Imprint* 1:1-30.

Fine, Kit (2007) *Semantic Relationism*, Malden, MA: Blackwell.

Fine, Kit (2012) “Guide to Ground”, in Correia & Schnieder (eds.), *Metaphysical Grounding: Understanding the Structure of Reality*, Cambridge University Press.

Fodor, Jerry & Lepore, Ernest (1991) “Why Meaning (Probably) isn't Conceptual Role”, *Mind and Language* 6 (4):328-43.

Forbes, Graeme (1990) “The Indispensability of Sinn”, *Philosophical Review* 99 (4):535-563.

Frege, Gottlob (1892) “Uber Sinn und Bedeutung”, *Zeitschrift für Philosophie Und Philosophische Kritik* 100 (1):25-50.

Glick, Ephraim (2015) “Practical Modes of Presentation”, *Noûs* 49 (3):538-559.

Goodman, Nelson (1955) *Fact, Fiction & Forecast*, Harvard University Press.

Grahek, Nikola (2011) *Feeling Pain and Being in Pain*, MIT press.

Gray, Aidan (2017) “Relational Approaches to Frege's Puzzle”, *Philosophy Compass* 12 (10):e12429.

Grice, Herbert Paul (1989) *Studies in the Way of Words*, Cambridge: Harvard University Press.

Hirsch, Eli (2005) “Physical-Object Ontology, Verbal Disputes, and Common Sense”, *Philosophy and Phenomenological Research* 70 (1):67–97.

Hirsch, Eli (2009) “Ontology and Alternative Languages”, in Wasserman, Manley & Chalmers (eds.), *Metametaphysics: New Essays on the Foundations of Ontology*, Oxford University Press.

Hofweber, Thomas (2009) “Ambitious, Yet Modest, Metaphysics”, in Wasserman, Manley & Chalmers (eds.), *Metametaphysics: New Essays on the Foundations of Ontology*, Oxford University Press.

Horgan, Terence (1984) “Jackson on Physical Information and Qualia”, *Philosophical Quarterly* 34:147-52.

Jackson, Frank (1986) “What Mary Didn't Know”, *Journal of Philosophy* 83 (5):291-295.

Jenkins, C. S. I. (2014) “Merely Verbal Disputes”, *Erkenntnis* 79 (S1):11-30.

Jones, Nicholas (2022) “Against Representational Levels”, *Philosophical Perspectives* 36 (1):140-157.

Jones, Nicholas (forthcoming) “Opacity in the Book of the World?”, *Philosophical Studies*.

Kaplan, David (1968) “Quantifying In”, *Synthese* 19 (1-2):178-214.

Kaplan, David (1989) “Demonstratives: An Essay on the Semantics, Logic, Metaphysics and Epistemology of Demonstratives and other Indexicals”, in Almog, Perry &Wettstein (eds.), *Themes from Kaplan*, Oxford University Press.

Kim, Jaegwon (1989) “The Myth of Non-Reductive Materialism”, *Proceedings and Addresses of the American Philosophical Association* 63 (3):31-47.

Kim, Jaegwon (2002) “The Layered Model: Metaphysical Considerations”, *Philosophical Explorations* 5 (1):2 – 20.

Kripke, Saul (1979) “A Puzzle about Belief”, in Margalit (ed.), *Meaning and Use*, Reidel.

Kripke, Saul (1980) *Naming and Necessity*, Harvard University Press.

Levin, Janet (2007) “What is a Phenomenal Concept?”, in Alter & Walter (eds.), *Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism*, Oxford University Press.

Levin, Janet (2019) “Once More Unto the Breach: Type B Physicalism, Phenomenal Concepts, and the Epistemic Gap”, *Australasian Journal of Philosophy* 97 (1):57-71.

Levine, Joseph (2007) “Phenomenal Concepts and the Materialist Constraint”, in Alter & Walter (eds.), *Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism*, Oxford University Press.

Lewis, David (1983) “New Work for a Theory of Universals”, *Australasian Journal of Philosophy* 61 (4):343-377.

Lewis, David (1990) “What Experience Teaches”, in Lycan (ed.), *Mind and Cognition*, Blackwell.

Lewis, David (2009) “Ramseyan Humility”, in Braddon-Mitchell and Nola (eds.), *Conceptual Analysis and Philosophical Naturalism,* MIT Press.

List, Christian (2017) “Levels: Descriptive, Explanatory, and Ontological”, *Noûs* 53 (4):852-883.

Litland, Jon (2017) “Grounding Grounding”, *Oxford Studies in Metaphysics* 10.

Loar, Brian (1990) “Phenomenal States”, *Philosophical Perspectives* 4:81-108.

Loar, Brian (1997) “Phenomenal States II”, in Block, Flanagan & Guzeldere (eds.), *The Nature of Consciousness: Philosophical Debates*, MIT Press.

Luper, Steven (2020) “Epistemic Closure”, in Zalta (ed.) *The Stanford Encyclopedia of Philosophy,* https://plato.stanford.edu/archives/sum2020/entries/closure-epistemic/.

Lycan, William (1996) *Consciousness and Experience*, MIT Press.

Magidor, Ofra (2010) “Robert Stalnaker, Our Knowledge of the Internal World”, *Philosophical Review* 119 (3):384-391.

Magidor, Ofra (2015) “The Myth of the De Se”, *Philosophical Perspectives* 29 (1):249-283.

McLaughlin, Brian & Bennett, Karen (2023) “Supervenience”, in Zalta & Nodelman (eds.) *The Stanford Encyclopedia of Philosophy,* https://plato.stanford.edu/archives/win2023/entries/supervenience/.

Melnyk, Andrew (2008) “Can Physicalism Be Non-Reductive?”, *Philosophy Compass* 3 (6):1281-1296.

Moore, Richard (2013) “Imitation and Conventional Communication”, *Biology and Philosophy* 28 (3):481-500.

Moss, Jessica (2025) “Knowledge-That is Knowledge-Of”, *Philosophers' Imprint* 25.

Nelson, Michael (2024), “Propositional Attitude Reports”, in Zalta & Nodelman (eds.) *The Stanford Encyclopedia of Philosophy*, <https://plato.stanford.edu/archives/fall2024/entries/prop-attitude-reports/>.

Nemirow, Laurence (1990) “Physicalism and the Cognitive Role of Acquaintance”, in Lycan (ed.), *Mind and Cognition*, Blackwell.

Nida-Rümelin, Martine (1995) “What Mary Couldn't Know: Belief About Phenomenal States”, in Metzinger (ed.), *Conscious Experience*, Ferdinand Schoningh.

Nida-Rümelin, Martine & O'Conaill, Donnchadh (forthcoming) "Is Dualism Compatible with Consciousness Being Grounded in the Physical?", in Rabin (ed.) *Grounding and Consciousness*, Oxford University Press.

Papineau, David (2002) *Thinking About Consciousness*, Oxford University Press.

Papineau, David (2007) "Phenomenal and Perceptual Concepts", in Alter & Walter (eds.), *Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism*, Oxford University Press.

Pavese, Carlotta (2020) "Practical Representation", in Fridland & Pavese (eds.), *The Routledge Handbook of Philosophy of Skill and Expertise*, Routledge.

Perry, John (1979) "The Problem of the Essential Indexical", *Noûs* 13 (1):3-21.

Perry, John (2001) *Knowledge, Possibility, and Consciousness*, MIT Press.

Plunkett, David & Sundell, Tim (2021) "Metalinguistic Negotiation and Speaker Error", *Inquiry: An Interdisciplinary Journal of Philosophy* 64 (1-2):142-167.

Plunkett, David & Sundell, Timothy (2023) "Varieties of Metalinguistic Negotiation", *Topoi* 42 (4):983-999.

Putnam, Hilary (1975) "The Meaning of 'Meaning'", *Minnesota Studies in the Philosophy of Science* 7:131-193.

Quine, Willard Van Orman (1956) "Quantifiers and Propositional Attitudes", *Journal of Philosophy* 53 (5):177-187.

Quine, Willard Van Orman (1970) *Philosophy of Logic*, Englewood Cliffs, N.J.: Prentice-Hall.

Raven, Michael (2016) "Fundamentality Without Foundations", *Philosophy and Phenomenological Research* 93 (3):607-626.

Rayo, Agustín (2013) *The Construction of Logical Space*, Oxford University Press.

Recanati, François (2012) *Mental Files*, Oxford University Press.

Richard, Mark (1990) *Propositional Attitudes: An Essay on Thoughts and How We Ascribe Them*, Cambridge University Press.

Rosen, Gideon (2010) “Metaphysical Dependence: Grounding and Reduction”, in Hale & Hoffmann (eds.), *Modality: Metaphysics, Logic, and Epistemology*, Oxford University Press.

Rosen, Gideon (2015) “Real Definition”, *Analytic Philosophy* 56 (3):189-209.

Rubenstein, Ezra (2024) “Two Approaches to Metaphysical Explanation”, *Noûs* 58 (4):1107-1136.

Russell, Bertrand (1905) “On Denoting”, *Mind* 14 (56):479-493.

Salmon, Nathan (1986) *Frege’s Puzzle (2nd edition)*, Ridgeview Publishing Company.

Schaffer, Jonathan (2009) “On What Grounds What”, in Wasserman, Manley & Chalmers (eds.), *Metametaphysics: New Essays on the Foundations of Ontology*, Oxford University Press.

Schaffer, Jonathan (2010) “Monism: The Priority of the Whole”, *Philosophical Review* 119 (1):31-76.

Schiffer, Stephen (1992) “Belief Ascription”, *Journal of Philosophy* 89 (10):499-521.

Schiffer, Stephen (2006) “Two Perspectives on Knowledge of Language”, *Philosophical Issues* 16 (1):275–287.

Schroer, Robert (2010) “Where's the Beef? Phenomenal Concepts as Both Demonstrative and Substantial”, *Australasian Journal of Philosophy* 88 (3):505-522.

Schwitzgebel, Eric (2024) “Belief”, in Zalta & Nodelman (eds.) *The Stanford Encyclopedia of Philosophy*, <https://plato.stanford.edu/archives/spr2024/entries/belief/>.

Sidelle, Alan (2007) “The Method of Verbal Dispute”, *Philosophical Topics* 35 (1-2):83-113.

Sider, Theodore (2006) “Quantifiers and Temporal Ontology”, *Mind* 115 (457):75-97.

Sider, Theodore (2009) “Ontological Realism”, in Wasserman, Manley & Chalmers (eds.), *Metametaphysics: New Essays on the Foundations of Ontology*, Oxford University Press.

Sider, Theodore (2011) *Writing the Book of the World*, Oxford University Press.

Sider, Theodore (2020) “Ground Grounded”, *Philosophical Studies* 177 (3):747-767.

Smithson, Robert (2020) “Metaphysical and Conceptual Grounding”, *Erkenntnis* 85 (6):1501-1525.

Soames, Scott (1995) “Beyond Singular Propositions?”, *Canadian Journal of Philosophy* 25 (4): 515-549.

Sosa, Ernest (1970) “Propositional Attitudes De Dicto and De Re”, *Journal of Philosophy* 67 (21):883-896.

Stalnaker, Robert (2008) *Our Knowledge of the Internal World*, Oxford University Press.

Stanley, Jason & Williamson, Timothy (2001) “Knowing How”, *Journal of Philosophy* 98 (8): 411-444.

Stoljar, Daniel (2016) “The Semantics of ‘What it’s like’ and the Nature of Consciousness”, *Mind* 125 (500):1161-1198.

Stoljar, Daniel (2017) “The Knowledge Argument and Two Interpretations of 'Knowing What it's Like'”, in Jacquette (ed.), *The Bloomsbury Companion to the Philosophy of Consciousness*, Bloomsbury Academic.

Tahko, Tuomas (2023) “Fundamentality”, in Zalta & Nodelman (eds.), *The Stanford Encyclopedia of Philosophy*, <https://plato.stanford.edu/archives/win2023/entries/fundamentality/>.

Thomasson, Amie (2017) “Metaphysics and Conceptual Negotiation”, *Philosophical Issues* 27 (1):364-382.

Tomasello, Michael (2008) *Origins of Human Communication*, MIT Press.

Torre, Stephan & Weber, Clas (2019) “De Se Puzzles and Frege Puzzles”, *Inquiry: An Interdisciplinary Journal of Philosophy* 65 (1):50-76.

Tye, Michael (1995) *Ten Problems of Consciousness: A Representational Theory of the Phenomenal Mind*, MIT Press.

Tye, Michael (2000) *Consciousness, Color, and Content*, MIT Press.

Tye, Michael (2008) *Consciousness Revisited: Materialism Without Phenomenal Concepts*, MIT Press.

Tye, Michael (2010) “Knowing What it’s Like”, in Bengson & Moffett (eds.), *Knowing How: Essays on Knowledge, Mind, and Action*, Oxford University Press.

Van Fraassen, Bas (1980) *The Scientific Image*, Clarendon Press.

Veillet, Bénédicte (2015) “The Cognitive Significance of Phenomenal Knowledge”, *Philosophical Studies* 172 (11):2955-2974.

Vermeulen, Inga (2018) “Verbal Disputes and the Varieties of Verbalness”, *Erkenntnis* 83 (2):331-348.

Williams, J. Robert G. (2010). Fundamental and Derivative Truths. *Mind* 119 (473):103 - 141.

Williams, J. Robert G. (2012) “Requirements on Reality”, in Correia & Schnieder (eds.), *Metaphysical Grounding: Understanding the Structure of Reality*, Cambridge University Press.

Williamson, Timothy (2000) *Knowledge and its limits*, Oxford University Press.

Williamson, Timothy (2007) *The Philosophy of Philosophy*, Wiley-Blackwell.

Williamson, Timothy (2021a) “Epistemological Consequences of Frege Puzzles”, *Philosophical Topics* 49 (2):287-319.

Williamson Timothy (2021b), *The Philosophy of Philosophy (*2nd edition), Wiley-Blackwell.

Williamson, Timothy (2024) *Overfitting and Heuristics in Philosophy*, Oxford University Press.

Wilson, Jessica (2014) “No Work for a Theory of Grounding”, *Inquiry: An Interdisciplinary Journal of Philosophy* 57 (5-6):535-579.

Zalta, Edward (2024) “Gottlob Frege”, in Zalta & Nodelman (eds.), *The Stanford Encyclopedia of Philosophy,* <https://plato.stanford.edu/archives/fall2024/entries/frege/>.