5-2022

Indeterminacy, Disagreement, and Reasonable Reference Magnetism

Jaocb (Hengyun) Yang
William & Mary

Indeterminacy, Disagreement, and Reasonable Reference Magnetism

A thesis submitted in partial fulfillment of the requirement
for the degree of Bachelor of Arts / Science in Department from
William & Mary

by

Hengyun (Jacob) Yang

Accepted for Honors
(Honors)

Joshua Gert
Dr. Joshua Gert, Director

Paul Davies
Dt. Paul Davies

Talbot Taylor
Dr. Talbot Taylor

Williamsburg, VA
May 11, 2021

Indeterminacy, Disagreement, and Reasonable Reference Magnetism

Hengyun (Jacob) Yang

Acknowledgment

First, I thank my advisor Professor Joshua Gert for the valuable feedback on the many drafts of this thesis and for always pointing me in the right direction when I am stuck. I also thank Professor Paul Sheldon Davies and Professor Taylor Talbot for reading the thesis and providing their insightful comments. All three committee members have deeply influenced the way I think, and I can only hope that those influences shine through the thesis. I am also grateful for all of my instructors in the past four years, especially Professor Philip Swenson and Professor Aaron Griffith whose classes convinced me to be a philosophy major.

Next, I am thankful for my friends who have tolerated me and my obsession with the most "boring" aspect of philosophy. I thank, especially, Simon Yue and Ken Yang since it is with them on Spring break that I wrote the most important chapter of this thesis.

Last but certainly not least, I thank my parents for supporting me financially through college. Sorry I did not go to business school.

Table of Contents

Chapter 1: Introduction

It is a peculiar fact that we can make noises that carry rich meaning. When someone utters "is it cold in this room?", we understand that person to be asking a question. Moreover, we can infer from this question that the person asking is likely feeling cold, and, depending on the context, we may understand the question to be a request to turn down the AC. This thesis investigates this peculiar phenomenon and takes steps to answer the question: what makes such noises meaningful?

The word "meaning" is ambiguous. Many things compete for the meaning of "meaning". Take the word "why"; its meaning can include its connotation (curiosity), its pragmatic force (it is the start of a question), or its definition (the phrase "for what reason"). In this thesis, I focus on one specific aspect of meaning: reference. For example, the referent of "Joe Biden" is the person designated by the term, Joe Biden. In this case, we say that "Joe Biden" refers to Joe Biden. Similarly, the word "president" refers to the property of being president which is given by the set of all the objects that instantiate the property of being president. This set is called the extension of "president". In the following, I will refer to types of noises/scribbles like "Joe Biden", which refers to an individual object, as "names"; and I will refer to types of noises/scribbles like "president", which refers to a property or, what is the same, an extension, "predicates". Specifically, I focus on the referent of names and predicates and remain neutral about the referent of other expressions if they exist at all. Facts about which object is the referent of which name and which set of objects is the extension of which predicate are semantic facts. And to assign every name in a language a referent and every predicate an extension is to have a semantic theory for that language.

There are disputes about whether this is the right way to characterize semantics. For example, it has been argued that the referents of predicates are intensions which are (mathematical) functions that take an object as argument and spit out a set of possible worlds (Lewis, 1970; Montague, 1974). For the most part, I will assume that the referent of a name is an object and the referent of a predicate is an extension. I will, however, return to these alternative characterizations of reference when it becomes relevant. I make these assumptions so I may investigate the question: in virtue of what facts do semantic facts hold, or what makes a semantic theory correct? Following David Lewis (1974), let us imagine a being who knows every non-semantic fact there is to know (i.e., every fact that is not about the meaning of words or sentences). It should be possible, in principle, for this Lewisian being to work out the right semantic theory. This possibility is evident by the fact that we, being far short of omniscient, do this frequently. None of us are born with a semantic theory in our head (i.e., we are not innately equipped with knowledge about the reference and extensions of terms in our language), but we are able to acquire such knowledge based on nothing but the instances of those noises and scribbles plus their immediate context; and when we disagree about what the right semantic theory is, we appeal to similar non-semantic facts to settle the disagreement. What is needed here is not to come up with an empirical story of how exactly we come to learn a language, but rather a metaphysical theory of what makes a semantic theory correct. In other words, we ask "what principle can the Lewsian being appeal to that connects semantic facts with non-semantic facts?" This is a metasemantic question: it asks about the metaphysics of semantics. In this thesis, I will not seek to answer the metasemantic question directly. Instead, I narrow my focus to two perplexing metasemantics-related phenomena, and a proposal aimed at accounting for those phenomena. I maintain that since the two phenomena have to be accounted for somehow no

matter what metasemantic answer one has in mind, the proposal will be a necessary though not sufficient component to an adequate answer to the metasemantic question.

Phenomenon 1: Indeterminacy

One proposed answer to the metasemantic question is interpretationism. According to a simplistic version of interpretationism, a semantic theory is true if it gets the truth-value of every sentence in our language right. However, there is a perplexing phenomenon: given the formal features of most natural languages, there are many incompatible semantic theories that agree on the truth value of every sentence, so interpretationism entails that many incompatible semantic theories will all be right, a paradox.

For example, suppose the Lewisian being attempts to deduce the right semantic theory for English. A crazy semantic theory that assigns to the name "Wittgenstein" the referent of Joe Biden can be just as truth-maximizing as the correct semantic theory that assigns to it the Wittgenstein (Putnam, 1981). This can happen as long as the being also assigns to predicates like "philosopher" the property of being Joe Biden or a philosopher who is not Wittgenstein in the crazy semantic theory. Many other predicates will have to be assigned different extensions as well, but this can be done. According to interpretationism, a semantic theory is correct as long as it gets the truth-value of whole sentences right. Therefore, under interpretationism, the crazy semantic theory is just as correct as the semantic theory that assigns "Wittgenstein" the intuitive referent - Wittgenstein. This is the first phenomenon: if interpretationism is the correct answer to the metasemantic question, the meaning of words will be indeterminate. It will be indeterminate whether "Wittgenstein" refers to Wittgenstein or Joe Biden.

Phenomenon 2: Disagreement

It is obvious that some disagreements are substantive whereas some are trivial. It is substantive to disagree about whether abortion is morally permissible; it is trivial to disagree about whether hotdogs are sandwiches. If someone told me that they believe abortion to be impermissible, I would regard them as making a substantive moral mistake: as speaking falsely. However, if someone denies that hotdogs are sandwiches, I will simply regard them as meaning something slightly different by the word "sandwich". One way to think about this phenomenon is to regard the disputants as forming distinct communities. In the case of the pro-life community and pro-choice community, everybody means the same thing by "morally permissible", but in the case of a hotdogs-inclusive community and a hotdogs-exclusive community, the two communities mean different things when uttering the same phoneme /sændwɪdʒ/.

More can be said about the phenomenon; in both cases of disagreement described above, the two disputant parties associate different theories with the same word. In the case of the disagreement between pro-life people and pro-choice people, the disagreement can be traced back to different theories of what is morally permissible. One party accepts the statement "abortion is permissible" in their theory, while the other party accepts its negation. Similarly, the apparent disagreement about hotdogs turns on theories about what counts as a sandwich. But if this analysis is correct, while the different theories of moral permissibility are competing theories about the same subject, the sandwich theories are compatible theories about different subjects. This calls out for an explanation: why having different theories about moral permissibility constitute substantive disagreement, while different theories about sandwich result in different referents of the word "sandwich".

The Proposal: Reference Magnetism

To summarize, to explain phenomenon 1, we need to explain what makes the correct semantic theory correct when there are other theories that satisfy the requirement set by our metasemantic theory (such as interpretationism) just as well. As for phenomenon 2, we need to explain why some words shift referents when the corresponding theory shifts, whereas the referents of other words remain stable. One proposal that allegedly does the explanatory job posits a ranking of properties in terms of eligibility to be referred to (Lewis, 1984). Specifically, those with a higher ranking are easier to refer to, while those with lower ranking are harder.

Here is how the proposal accounts for phenomenon 1: according to the proposal,, getting the right truth-value right is no longer our sole concern, we also need to make sure the properties that are assigned as the referents of predicates in our semantic theories are sufficiently easy to refer to. Therefore, if we can conclude that the property of being a philosopher is more eligible to be referred to than the property of being Joe Biden or a philosopher who is not Wittgenstein, then by the proposal, the semantic theory that assigns the predicate "philosopher" an extension of philosophers, instead of an extension oJoe Biden and philosophers who are not Wittgenstein, is the correct one. Then, because the correct theory must also get the right truth-value, and the sentence "Wittgenstein is a philosopher" is true, the name "Wittgenstein" must be assigned Wittgenstein as its referent as opposed to Joe Biden.

The proposal also accounts for phenomenon 2. Suppose we have a predicate "F" and two similar theories associated with it, T1 and T2. Specifically, T1 is a true theory that describes the property P1, and T2 is a true theory describing another property P2. According to the proposal, if P1 and P2 are roughly as eligible as each other, then disagreeing about whether T1 or T2 is the true theory of F "splits" the referent of "F": those who hold T1 refer to P1 with "F" while those

holding T2 refer to P2 by F. The reason is that since the properties are equally easy to be referred to, the referent of "F" is fixed by whatever theory one associates with it. On the other hand, if P1 is sufficiently more eligible than P2, then the referent of "F" will be fixed to P1 thus making T2 a false theory of F and the disagreement between those holding T1 and those holding T2 a substantive one. Therefore, if, for some reason, we reason that the property of moral impermissibility ranks sufficiently high on the eligibility scale compared to other candidates, then disputants who hold different theories of impermissibility will be talking about the same moral property. In contrast, if we are led to reason that neither the property of being a hotdog-inclusive sandwich nor the property of being a hotdog-exclusive sandwich ranks much higher than the other, we can conclude that the referent of "sandwich" is fixed completely by one's sandwich theory, so the person with a different theory refers to a different property. This proposal is called "reference magnetism", which will be the focus of this thesis.

Roadmap

The thesis is structured like this: in chapter 2, I will introduce the two phenomena with more details. Specifically, the first phenomenon will be given the name "the indeterminacy problem". I will focus on a version of this, called "the permutation problem," which was raised by Hilary Putnam (1981). The second phenomenon will be given the name "the problem of disagreement". I will focus on two instances of this problem: the moral twin earth case and the brains-in-vats case.

In chapter 3, I turn to the various theories of reference magnetism. These theories mostly focus on analyzing the nature of eligibility (i.e., what makes a property eligible). I will start with the canonical version of the theory which I will call the "theory-external analysis", before

turning to some variations. I will argue that none of these theories are satisfactory, in hopes of motivating the search for an alternative theory of reference magnetism.

In chapter 4, I present my alternative theory. I argue that facts about eligibility arise from our changing use of language over time. The theory argued for here is built on Jody Azzouni's solution to the rule-following problem (2017), so I will present Azzouni's work before developing it into an alternative to traditional formulation of the reference magnetism theory. Finally, I apply this alternative theory to the problems presented in chapter 2 and answer potential objections before concluding the whole thesis in chapter 5.

Chapter 2: Indeterminacy and Disagreement

In this chapter, I will introduce the two aforementioned phenomena under the headings "the problem of indeterminacy" and "the problem of disagreement" in more detail. I will start with the problem of indeterminacy in 2.1 and move to the problem of disagreement in 2.2.

2.1 The Problem of Indeterminacy

I will start with the problem of indeterminacy. The problem of indeterminacy follows a general formula: a metasemantic theory is introduced where a type of non-semantic facts T are alleged to determine semantic facts. Then, it is shown that T underdetermines semantic facts. Thus, it is concluded that if T is all there is to fix semantic facts, there would be mass indeterminacy.

2.1.1 Global descriptivism and Permutation Problem

Let me begin by introducing the theory of global descriptivism. Global descriptivism is a metasemantic theory that says the following: the best semantic theory for a language L is the one that maximizes the number of true sentences in a selected set of L sentences (Lewis, 1984). There is some disagreement about which sentences are selected, but here I follow Lewis in supposing that these are sentences that speakers of L hold true as a part of their theory about the world in the most general sense.

To motivate this global descriptivism, consider how we learn new words. Take the word "appurtenant" for an example. If we look up the meaning of the word in a dictionary, we will see that its meaning is explained by other words, specifically the words "supplying additional support". Moreover, when new words are introduced, it is often done by appealing to old ones

with which we are already acquainted. For example, I have once been told that the expression "to cap" just means the same as "to bluff". Call such sentences as "to be appurtenant is to supply additional support" and "to cap is to bluff" meaning-determining sentences. These sentences inform us of the meaning of certain expressions by appealing to expressions we already understand. To generalize, suppose "F" is a word in need of definition, and "$\forall x\ (Fx \leftrightarrow Gx)$" is a meaning-determining sentence, the extension of "F" (or the property referred to by "F") is whatever set of objects that satisfies "Gx". Other examples of meaning-determining sentences may include theoretical statements like "relativistic mass is mass times the Lorentz factor" or the classic "a bachelor is an unmarried male". Normally, it is supposed that this approach can only help fix the meaning of "new" words, whereas the meaning of the "old" ones will have to be fixed in some other way.

Global descriptivism, then, is motivated by the hope that, perhaps, old words and new words alike can be defined if the set of true meaning-determining sentences is large enough. Specifically, we can try to fix the meaning of "male" and "unmarried" by adding sentences like "Wittgenstein is male", "Anscombe is not male", "Wittgenstein is unmarried", "Anscombe is not unmarried" to the set of meaning-determining sentences. In actual practice, we select sentences that we are fairly sure to be true (like the marriage status of Wittgenstein), and try to make it such that for any two distinct predicates "F" and "G", either there is at least one open sentence $P(\phi)$ such that P(F) and ~P(G) are in the set (or vice versa). For example, for the predicates "red" and "green", we make sure that for some object x such that x is green and not red, the statements "x is green" and "x is not red" are included in the meaning-determining set so as to make sure that we do not assign the same extension to "green" and "red". Of course, every name and predicate we wish to define should appear in the meaning-determining set. The hope is that since

we are looking for the true semantic theory for every English predicate and name, the semantic theory under which the sentences in such a meaning-determining set come out true will be the right one[1].

The hope of global descriptivism is shattered by Putnam's permutation argument (1981). According to the argument, given a meaning-determining set of sentences, a crazy semantic theory can make exactly the same sentences true as an intuitive theory. Let's take a toy language (call it "L") for example. The meaning-determining set for L contains the following sentences:

- Adam is a dog
- Betty is a dog
- Carl is a dog
- Dave is a cat
- Eve is a cat

Let's stipulate the following to be the correct semantic theory for L:

- "Adam" refers to Adam
- "Betty" refers to Betty
- "Carl" refers to Carl
- "Dave" refers to Dave
- "Eve" refers to Eve
- "Dog" has the extension {Adam, Betty, Carl}
- "Cat" has the extension {Dave, Eve}

[1] However, given that the meaning-determining set may contain some false sentences (i.e., if Wittgenstein were secretly married), the constraint can be loosened: whatever semantic theory that maximizes truth will at least approximate the true theory.

It is obvious to see that the correct semantic theory makes every meaning determining sentence true. If global descriptivism is true, any semantic theory that makes exactly the same sentences true will be just as good as the correct interpretation. However, consider the following semantic theory:

- "Adam" refers to Eve
- "Betty" refers to Dave
- "Carl" refers to Carl
- "Dave" refers to Betty
- "Eve" refers to Adam
- "Dog" has the extension {Eve, Dave, Carl}
- "Cat" has the extension {Adam, Betty}

This semantic theory is false since we have stipulated that "Adam" in L refers to Adam and not Eve, "Betty" refers to Betty and not Dave, etc. However, this semantic theory still manages to make every sentence in the meaning-determining set true. Specifically, "Adam is a dog" is still true as "Adam" refers to Eve and Eve is in the extension of "Dog" (according to T2), "Betty is a dog" is true as "Betty" refers to Dave and Dave is in the extension of "Dog" (according to T2), etc. Call the stipulated semantic theory T1 and the crazy theory T2. We have the following set of inconsistent sentences:

1. T1 is an acceptable theory (or, more strongly, the true theory).
2. Two semantic theories are equally acceptable, if they make the same sentences true (because the quality of a semantic theory depend solely on how many sentences from the meaning-determining set it makes true)
3. Both T1 and T2 make every sentence the meaning-determining set of L true

4. T2 is an unacceptable theory (or, more weakly, a false theory)

Here, (1) and (4) are true by stipulation. (2) is simply a restatement of global descriptivism. Therefore, since (3) just follows from how we have constructed T2, (2), or global descriptivism, is the one that must go.

Of course, L is very impoverished as far as languages go: it has only five names and two predicates. Any real natural language will be much more complicated with an exponentially larger set of meaning-determining sentences. Therefore, in the following, I will also generalize the previous argument to any language with a universe of discourse that is larger than 1. Before I do that, I will introduce some necessary machinery. First, by the expression "an interpretation function of a semantic theory T", I refer to a function that takes a name or a predicate in a language and returns their corresponding referent or extension according to T. For example, let "||" be the interpretation function of the true semantic theory for English. Then, |"Wittgenstein"| interprets the name "Wittgenstein", so '|"Wittgenstein"|' refers to Wittgenstein himself. Similarly, |"dogs"| interprets the predicate "dog", so '|"dogs"|' refers to the set of actual dogs. Second, a permutation on a language is a bijective function that maps the universe discourse of that language onto itself. Here, to say that the function maps the universe of discourse of a language onto itself just means that it takes a member in the universe of discourse and returns another member in that same universe of discourse. Further, to say that the function is bijective just means that every element in the universe of discourse is mapped to a distinct element: in other words, every element is mapped to exactly one element. For example, suppose a function σ maps Wittgenstein onto Elon Musk, maps Elon Musk onto Wittgenstein, and maps everything else in the universe of discourse to itself. σ would be an example of a permutation on English. In this example, σ(Wittgenstein) is Elon Musk, σ(Elon Musk) is Wittgenstein, and for any x that is

neither Wittgenstein nor Elon Musk, σ(x) is x. This can be put as "the image of Wittgenstein under σ is Elon Musk, the image of Elon Musk under σ is Wittgenstein, and the image of everything else under σ is itself."

Now, the result for L can be extended. Take the English language for example, suppose again that "||" is the interpretation function for English according to the intuitively correct semantic theory, and σ is a permutation function that maps Wittgenstein and Elon Musk to each other but leaves everything else the same. Define a function ||* for English according to the following rule:

- For any name N, $|N|^* = \sigma(|N|)$
- For any predicate F, $|F|^* = \{\sigma(x) | x \in |F|\}$

We can see that ||* is an interpretation function of English corresponding to a new semantic theory. Call this new semantic theory "T2", and call the intuitive semantic theory that corresponds to the function || "T1". As the rules suggest, according to T2, a name in English refers to the image of its referent according to T1 under σ; and the extension of a predicate (according to T2) includes the image of everything that is in the extension of that predicate according to T1. Therefore, according to T2, "Wittgenstein" refers to Elon Musk, and "Elon Musk" refers to Wittgenstein. Similarly, the predicate "being a philosopher of language" now stands for the property of being Elon Musk or a philosopher of language who is not Wittgenstein, the predicate of "being an Austrian" stands for the property of being Elon Musk or an Austrian that is not Wittgenstein, the predicate of "being alive in the 21st century" stands for the property of being Wittgenstein or anyone who alive in the 21st century except for Elon Musk, etc. As we can see, this semantic theory is obviously wrong: it gets wrong the meaning of two names, "Wittgenstein" and "Elon Musk", as well as the meaning of countless predicates.

However, despite this, for every sentence that is made true by T1 in the meaning-determining set for English, T2 makes the same sentence true. In fact, it can be shown that T1 and T2 make exactly the same sentences true whether it is in the meaning-determining set or not. Specifically, if it can be shown that T1 and T2 make exactly the same atomic sentences true, then we know that they make true exactly the same compound sentences with only truth-functional connectives.[2,3]

Therefore, if we can show that T1 and T2 make exactly the same atomic sentences true, we will have shown that for every sentence that is made true by T1 in the meaning-determining set for English, T2 makes the same sentence and nothing else true. Here is how this is done: for any atomic sentence "ϕx", "ϕx" is true if and only if |"x"| ∈ |"ϕ"|. However, if |"x"| ∈ |"ϕ"|, σ(|"x"|) ∈ σ(|"ϕ"|) by construction. Moreover, since σ(|"x"|) just is |"x"|*, and σ(|"ϕ"|) just is |"ϕ"|*, σ(|"x"|) ∈ σ(|"ϕ"|) if and only if |"x"|* ∈ |"ϕ"|*. In other words, if the referent of "x" according to T1 falls under the extension of "ϕ" according to T1, the referent of "x" according to T2 will also fall under the extension of "ϕ" according to T2. This holds for any atomic sentence in English. Consider again the argument that global descriptivism cannot be true for a language such as L. The argument can be transplanted for English:

---

[2] This can be done by induction on the connectives. Here is a sketch for the conjunction "&" as an example.

- Base case: T1 and T2 agree on the truth-value of every atomic sentence.
- Inductive step: Assume that T1 and T2 agree on the truth-value of and. Then, T1 and T2 agree on the truth value of φ & ψ.

Thus, T1 and T2 agree on the truth value of any sentence that only has conjunction.

[3] We may have to "first-orderize" English by treating non-truth functional operators as higher-order predicates that take tuples of sentences as arguments and spit out a truth value. (For example, a counterfactual conditional takes its antecedent and its consequent as arguments and returns a truth value.) Then, we interpret these operators them the following rules:

- For any non-truth-functional operators O, |O| = |O|*

This may seem to jettison results in semantics, but the purpose here is only to show that such reassignment is possible. For permutated semantic theories that are more faithful to these results, see 2.1.2 in this thesis for applications to possible world semantics, Williams (2005) for applications to Montague grammar, and Williams (2008a) for applications to indexicals.

5. T1 is an acceptable theory (or, more strongly, true)
6. Two semantic theories are equally good, if they make the same sentences true (because the quality of a semantic theory depend solely on how many sentences from the meaning-determining set it makes true)
7. Both T1 and T2 make every sentence the meaning-determining set of L true
8. T2 is unacceptable (or, more weakly, false)

(5) is true since T1 is the intuitive interpretation for the English language. (8) is true since T2 assigns Elon Musk as the referent of "Wittgenstein", Wittgenstein as the referent of "Elon Musk", and gerrymandered properties to countless properties. (6) is again a restatement of global descriptivism. (7) is what I have shown in the last couple paragraphs. In fact, the result here can be fully generalized. What I have shown for (7) does not rely on any special property of English save one: it has a universe of discourse larger than 1. Therefore, for any language that refers to more than one object, global descriptivism cannot be true. As such, it is safe to suppose that global descriptivism, as stated, cannot be true for any natural language at all.

2.1.2 Permutation for Intensions

Global descriptivism is sometimes presented as a strawman version of more sophisticated metasemantic theories. Therefore, it is worth showing that the permutation argument extends to the more sophisticated theories as well. I will consider two variations of such theories: the first theory sets stronger constraints for the correct semantic theories: the correct theories must get the meaning of the meaning-determining sentences right where the meaning of a sentence is seen as the set of possible worlds in which the sentence is true. Therefore, according to this first theory,

the quality of a semantic theory depends on whether it gets the meaning of the sentences in the meaning-determining set right.

To start with, consider a meaning-determining sentence “to be cordate is to have a heart”. Suppose that everything that has a heart also has a kidney. The sentence cannot determine the meaning of “cordate” if the only constraint on the semantic value of “cordate” is that it makes the meaning-determining sentence true. This is because the set of every creature with a kidney, when assigned as the semantic value of “cordate”, also makes the sentence true (at least for the extensional fragments of English). Therefore, one way to amend the theory to avoid this problem is to regard the meaning of sentences not as their truth-values but the set of possible worlds in which they are true. The idea here is that even if, in the actual world, every creature that has a heart also has a kidney, it is possible that something can have a heart but not a kidney. This amendment gives the desired result, because the sentence “anything that is a cordate has a heart” is true in every possible world, but the sentence “anything that is a renate has a heart” is true only in some worlds. So the two sentences do not have the same meaning. Therefore, if it is required that the correct semantic theory must assign the meaning-determining sentence “to be a cordate is to have a heart” its correct meaning, the set of all possible worlds in this case, the theory cannot assign to “cordate” the property of having a kidney which only make the sentence true in some possible worlds.

A permutation argument can be formulated for semantic theories formulated like this as well (Hale & Wright, 2017). Here is how this is done: first, formulate T1, the intuitive semantic theory for English. T1 assigns to every name in English a function from a possible world to an object. This can be written as a set of ordered pairs where the first term is a possible world and the second term is the object the name refers to in that world. For example, using “||” as standing

for the interpretation function of T1 for English, we can write: |“Wittgenstein”| = {<W_1, Wittgenstein>, <W_2, Wittgenstein>, <W_3, Wittgenstein>...}. For convenience, say the following:

- |“N”| = {<W_i, N_i>} for every i, where N_i is the referent of the “N” in world i

Predicates, under semantic theories formulated like this, will also be assigned a function from an indexed possible world to the extension of that predicate in that world. For example, |“philosopher of language”| = {<W_1, {Wittgenstein, Kripke, Russell…}>, <W_2, {Kripke, Russell…}>}. (Here, W_2 is a possible world where Wittgenstein did not become a philosopher.) In other words,

- |“F”| = {<W_i, F_i>} for every i, where F_i is the extension of “F” at world i

Moreover, for two sentences to have the same meaning, it is not enough that they have the same truth value: they also need to be assigned the same truth value in every possible world. In other words, each sentence is assigned a function from an indexed possible worlds to a truth value according to the semantic value of its subsentential components which can be written as a set of tuples with the first term being an indexed possible world and the second term a truth value. According to the examples in this paragraph, the sentence “Wittgenstein is a philosopher of language” will be assigned the semantic value {<W_1, True>, <W_2, False>...} . And it is required that the correct semantic theory should assign the right function to the right sentences.

Now, a crazy semantic theory T2 can be formulated. To begin with, choose an arbitrary permutation function σ . Let “||” be the interpretation function for T1 and “||*” be the interpretation function for T2. T2 says the following:

- |“N”|* = {<W_i, $\sigma(N_i)$>} for every i
- |“F”|* = {<W_i, {$\sigma(x)$| x $\in F_i$ }>} for every i

Under T2, for any possible world W_x, any atomic sentence of the form "Fa" is true if and only if $\sigma(a_x) \in \{\sigma(x)| x \in F_x \}>\}$. By construction, this is true if and only if $a_x \in F_x$. Therefore, for any possible world x, any atomic sentence of the form "Fa" is true under T2 as long as it is true under T1. The extension to compound sentences remains the same. This means that in every possible world, a sentence is true under T2 if it is true under T1. Therefore, the permutation argument can be formulated for a third time:

9. T1 is acceptable (or, more strongly, true)
10. Two semantic theories are equally good, if they get the meaning of the same sentences right (because the quality of a semantic theory depend solely on the number of sentences from the meaning-determining set for which it gets the meaning right
11. Both T1 and T2 get the meaning of the exactly the same sentences right
12. T2 is unacceptable (or, more weakly, false)

(9) and (10) remain the same. (10) is the improved statement of global descriptivism, according to which the true semantic theory needs to get the meaning, conceived as functions from possible worlds to truth values, of sentences right. (11) is what is argued above: T2 and T1 assign the same meaning, again conceived as functions from possible worlds to truth values, to every sentence.

2.1.3. Davidsonian Interpretationism and the Foster Problem

The second theory I will consider is Donald Davidson's theory of interpretation (Davidson, 1973; 1974). The purpose here is to show that the permutation problem does affect real metasemantic theories.

According to Davidsonian interpretationism, the meaning of a sentence has to do with the intention of the speaker in uttering that sentence. However, according to Davidson, it is not possible to know the intention of the speakers without at the same time figuring out the meaning of their utterances (Davidson, 1990). Therefore, the meaning of sentences must be "solved" in some other way. Here, Davidson thinks that the meaning of a sentence is the condition under which it is true. More specifically, such truth-conditions are given by T-sentences of the form "S is true if and only if P" where S refers to a declarative sentence and P is a proposition. As such, a semantic theory of a language L successfully captures the meaning of L-sentences if it entails the T-sentence for every sentence in that language. In other words, the correct semantic theory for a language must entail propositions of the form "S is true if and only if P" for every sentence S in the language.

For such a theory to be successful, it must satisfy two additional constraints. According to the first formal constraint, the theory itself must be finite in order to explain our ability to understand long, novel sentences and to learn to navigate the complex system of natural language. To satisfy this requirement, it suffices that the semantic theory assigns semantic values to a finite lexicon of names and predicates (as opposed to whole sentences the number of which are infinite), assuming that there is also a finite set of rules of grammar. According to the empirical constraint, the theory must be empirically adequate. Here, to verify that a semantic theory is empirically adequate, we take on the substantive though defeasible assumption that the speakers have true beliefs and that the utterances made by the speaker are genuine. (This is the Charity assumption.) Then, a semantic theory is empirically adequate, if according to the semantic theory, the utterances made by the speaker are true. For example, suppose that the sentence type S_1 is typically uttered when P obtains. Then, assuming that the speaker has true

beliefs and speaks genuinely, we may defeasibly infer from the empirical generalization that S_1 is uttered whenever P obtains to the T-sentence "'S_1' is true if and only P".

To give another example, suppose that it is an empirical generalization that the sentence "the cat is on the mat" is normally uttered whenever the cat is on the mat. By the assumption of charity, we can defeasibly hold that the true semantic theory must entail the T-sentence "'the cat is on the mat' is true if and only if the cat is on the mat". How can a finite semantic theory entail it? Take an intuitive semantic theory for English. The theory would contain the following proposition as axioms

- "The cat" refers to the cat
- "X is on the mat" is true if and only if X is on the mat

The theory will contain many axioms like these, but assuming that we have a finite lexicon, the theory will still be finite. It is also easy to see that the two axioms above entail that "the cat is on the mat" is true if and only if the cat is on the mat.

Call a T-sentence "interpretative" if it specifies the meaning of the sentence on the left-hand side of the biconditional. A problem for Davidsonian interpretationism, then, is that the T-sentences generated by a finitely specified empirically adequate semantic theory may not be interpretive (Foster, 1976). To elaborate, the T-sentence "'the cat is on the mat' is true if and only if the cat is on the mat" is interpretive because the proposition indeed gives the meaning of the sentence. However, a T-sentence like "'Wittgenstein is philosopher of language' if and only if Elon Musk instantiates the property of being Elon Musk or a philosopher of language that is not Wittgenstein" is not interpretive as the proposition in the second half of the bi-conditional does not give the meaning of the sentence "Wittgenstein is a philosopher of language"[4]. Here,

[4] This is, to be sure, not the familiar form of the foster problem. However, as Williams (2008b) argues, the problem is essentially the same.

the problem is that Davidsonian interpretationism only requires that the T-sentences are true, but both interpretive and uninterpretive T-sentences can both be true as the above example shows.

Davidson, at times, appears to assume that no finite semantic theory of English will be able to generate T-sentences that are both true and uninterpretive for every sentence. (Davidson, 1973) This assumption, if Davidson indeed holds it, is mistaken: again, let σ be a permutation function for English that maps Wittgenstein and Elon Musk to each other and everything else to itself; and let I be the interpretation function for English according to an intuitive semantic theory T1. T1 will entail, for every atomic sentence "ϕx" in English, the T-sentence "'ϕx' is true if and only if |"x"| is a member of |"ϕ''|. This is the intended result, since according to T1, |"Wittgenstein"| is Wittgenstein, and |"philosopher of language")| is the set of all philosophers of language: the resulting T-sentence will be interpretive. However, we can construct another crazy semantic theory T2 corresponding to an interpretation function ||* such that, for every name "X", |"X"|* is σ(X), and for every predicate "ϕ", |"ϕ"|* is $\{\sigma(x)|x \in |\text{“}\phi\text{”}|\}$. Then, T2 entails for every atomic sentence "Fa" in English the T-sentence "'Fa' is true if and only if |"a"|* is a member of |"F"|*." This sentence is true if and only if the following sentence is true: 'Fa' is true if and only if σ(a)is a member of $\{\sigma(x)|x \in |\text{“F”}|\}$. By construction, this is true if and only if the sentence "'Fa' is true if and only if |"a"| is a member of |'F'|" is true. In other words, for any atomic "ϕx", T1 will entail a true T-sentence for it if and only if T2 does so as well. However, whereas T1 supposedly entail an interpretive T-sentence, T2 does not: T2 entails the T-sentence "'Wittgenstein is philosopher of language' if and only if Elon Musk instantiates the property of being Elon Musk or a philosopher of language that is not Wittgenstein". Extension to compound sentences is analogous to that in section 2.1.1. This shows that like global descriptivism, Davidsonian interpretationism underdetermines the right semantic theory for English: theories

that generate interpretive T-sentences and theories that generate uninterpretive sentences can equally satisfy the empirical constraint and the formal constraint.

2.2 the problem of disagreement

In order for two speakers to disagree with each other, it seems necessary that they are disagreeing about the same topic. In other words, the two disagreeing parties have to be making genuinely inconsistent claims. The problem of disagreement, then, is a challenge for any metasemantic theory to accommodate speakers' abilities to disagree with each other. As we will see, it is not obvious that our best metasemantic theories are able to meet this challenge.

Take global descriptivism as an example. Suppose two speakers A and B are disagreeing about whether P is true where A insists on P and B (apparently) insists on its negation. According to global descriptive, the right semantic theory should maximize the number of true sentences in the meaning-determining set. However, according to a well-known result in model-theory, there is a model for any syntactically consistent first-order theory (Henkin, 1949). For us, this means that if we construct a meaning-determining set for each of them that includes P (for A) or its negation (for B), there will be a distinct semantic theory for each of them that makes their utterance "P" (or "not-P") true. However, this will also entail that "P" uttered by A and "not-P" uttered by B are not really inconsistent (since they are both true), and thus A and B are not really disagreeing.

To generalize, the problem of disagreement is a challenge for any metasemantic theory to offer an account that can accommodate disagreement. Whereas the problem of indeterminacy only targets various stripes of interpretationism and descriptivism, the problem of disagreement affects (almost) every metasemantic theory. In the following, I consider two popular

metasemantic theories, and present scenarios or areas of discourse where they fail to accommodate disagreement. Specifically, my strategy is this: first, a reference-fixing mechanism is introduced. Then, a specific scenario is sketched where, intuitively, two speakers are having substantive disagreements. Finally, I argue that given the reference-fixing mechanism, the scenario in question is one where the speakers are not having genuine disagreement, so the reference fixing mechanism fails to accommodate genuine disagreement.

2.2.1 Moral Twin Earth

The first metasemantic theory proposes the following reference-fixing mechanism: a term T refers to X if and only if X is causally connected to the use to T in the appropriate way. There is a question of what the appropriate way is exactly: it may be that X has to be a cause of the event where X is named T (Kripke, 1980); it may that the use of T is probabilistically dependent on the presence of X (Dretske, 1981; Artiga & Sebastian, 2020); or it may that X causally regulates the use T such that the users of T converge on an approximately true theory of X over time (Boyd, 1988). The details of such a theory are not important for my purpose. I will assume that there is a causal mechanism M such that if T is related to X by M, then T refers to X according to the causal theory. Different versions of the causal theory will differ on what exactly M is, but qua causal theories, they all agree that M is causal.

The scenario, originally proposed by Horgan and Timmons (1991), is the following: imagine that there is a distant planet, call it “twin Earth”, that is almost identical to our Earth. The twin Earth is populated by people much like ourselves. We share almost identical physical features in almost identical physical environments. The people on twin Earth even speak a language that is almost identical to ours except in one feature that is to be specified. For

convenience's sake, let's assume that the words "moral goodness" in our language stand for the property of maximizing pleasure since that is the property that relates to our use of the words "moral goodness" by M. To fill in M a little for illustrative purposes, it may be that the event of naming an action type "morally good" is caused by the fact that the named action maximizes pleasure; it may that the probability of the presence of pleasure-maximizing action, as opposed to other types of actions, conditioned on the utterance "this is good" is the highest; or it may that over time, we would converge on using "morally good" to describe actions that maximizes pleasure. On twin Earth, the language spoken there is almost identical to English: in fact, the language is so similar that they use a string of phonemes that sound exactly like "moral goodness" to describe certain actions. Like us, they would use this exact string of phonemes to justify certain decisions, to praise and blame, and to describe actions like helping a sick man in need. To distinguish the two, I will henceforth italicize phrases used by the moral twin Earthers. However, there is one difference between us and our twin Earth cousins: their phrase *"moral goodness"* is related to the property of treating others as ends in themselves by M, not maximizing pleasure. Of course, the two moral theories here are not important as long as they are different theories.

According to the causal theories, T refers to X if X is related to T by M. By the stipulations in the moral twin Earth thought experiment, "moral goodness" in English refers to the property of maximizing pleasure, and *"moral goodness"* in twin Earth speak refers to the property of treating others as ends in themselves. Therefore, supposing that a moral twin Earth philosopher is joined by one of our own, and the two disagree on whether it is morally good to kill 1 to save 5, the above analysis would entail that the disagreement between the two is illusionary. Specifically, causal theories entail that the sentence "it is good to kill 5 to save 1" in

English expresses the proposition that the action of killing 5 to save 1 is pleasure-maximizing, and the sentence *"it is not good to kill 5 to save 1"* in twin Earth speech expresses proposition that the action of killing 5 to save 1 does not treat others as ends in themselves. The two propositions are not incompatible. However, this is a counterintuitive result. Suppose that the two philosophers had that disagreement in the context of having to decide whether they should murder an innocent person to save another five. Their disagreements on what to do are certainly real. Moreover, both philosophers agree to the following: only one of them recommends the morally correct action which is justified by her true utterance. However, on the causal theories, their utterances are both true. So what action do they license? This problem is not by any means restricted to wild thought experiments regarding distant planets. Moral terms from different cultures likely bear the appropriate causal relation to different properties. However, mundane moral disagreements across cultures are certainly not illusory.

Therefore, we see that the causal theories fail the challenge: they undermine the idea that there are substantive disagreements in moral discourse. Before I move on, I will mention a solution to the moral twin earth problem to motivate the next victim of the problem of disagreement: conceptual role semantics. According to a version of conceptual role semantics, T refers to X if and only if X would make the use of T correct. To elaborate, for every term T, we can specify its use as a conceptual role. The conceptual role of T will include conditions that allow the speakers to use T (introduction rules) and the consequences of using T (elimination rules). According to this version of conceptual role semantics, the referent of T must make the rules correct (Wedgewood, 2001; Williams, 2018). In other words, the referent of T must be such that a speaker who follows the rules for using T is correct or rational in doing so. If we substitute this theory for causal theories and have a robust enough account of what counts as

correct/rational, there might be a solution to the problem. Suppose that the term "moral goodness" on our Earth has the following conceptual role (call it R):

- "Goodness" introduction: if a speaker approves of an action, the speaker is entitled to call that action "good".
- "Goodness" elimination: if a speaker sincerely calls an action "good", the speaker is committed to praise whoever does that action.

These rules are obviously too simplistic, and they likely get our moral psychology wrong. Nonetheless, suppose something like this is identified. It is expected that the term "*moral goodness*" used by the twin Earthers will have the same conceptual role (R). Moreover, suppose that there is a unique property P such that an individual following R is rational if and only if "moral goodness" refers to P. Then, given that the same R that governs our term "moral goodness" also governs twin Earthers's use of *"moral goodness", "moral goodness"* will have to refer to P as well if the twin Earthers are to be rational. Therefore, provided that (1) this particular version of conceptual role semantics is true, (2) the Twin Earther's term that is homophonic to "moral goodness" shares the same conceptual role, R, as our term "moral goodness", and (3) there is a unique property that makes R rational/correct, we and the moral twin earthers do refer to the same property when we make the sound "moral goodness". This gets the right result: there is a common property about which we are disagreeing with the twin Earthers, and when we differ on what that property is, we are having genuine disagreement.

2.2.2 Brains-in-vats

We see that although the causal theories fail to meet the challenge in cases of moral disagreement, conceptual role semantics[5] provides hope. In the following, I consider a case where conceptual role semantics fails to meet the challenge as well. To begin with, I will say more about conceptual role semantics. According to conceptual role semantics, the referent of T is whatever makes the use of T correct. I have mentioned that the conceptual role of T can be given as rules of introduction and elimination. Nonetheless, two more questions can be asked about this barebones description: first, what determines the conceptual roles of a term? Second, what is it for those rules to be correct? Depending on different theories that fall under the general umbrella of conceptual role semantics, a range of answers can be provided. To start with the second question, one criterion for correctness may be that the rules of the entire language do not lead to contradiction (Dummett, 1973). For example, no proposition can have the introduction rule of P while also having the elimination rule of not-P. (For example, define an operator "⋆" to have the introduction rule that if P, then ⋆P, and the elimination rule that if ⋆P, then not P. Such an operator, then, would lead to a contradiction.) A stronger criterion may be that if the input corresponding to the introductory rule is true, the output corresponding to the elimination rule must also be true (Horwich, 2004). In other words, suppose the introductory rule for P requires the belief that Q and the elimination rule outputs the belief that R, then it must be the case that if Q is true, R is true. However, as with the conceptual role of "moral goodness", since the input and output need not be belief (or propositions), truth predicates may not always apply. Therefore, an amendment is that if the input corresponding to the introductory rule is *correct*, the output corresponding to the elimination rule must also be *correct* (Wedgewood, 2001). Here, being correct can mean anything from believing a true proposition to simply being

[5] I will use "conceptual role semantics" and "inferential role semantics" interchangeably.

rational. To return to the first question, the dominant view seems to be that the conceptual role of a term is determined by either the causal role its tokens play in individual speaker's psychology (Block, 1986) or the linguistic norms within the community (Brandom, 1994).

Now, for the scenario. Imagine that every human being is in fact just a brain in a vat, and all of our sense impressions of the external world are fed to us by a powerful computer. Moreover, suppose that no one in human history has ever had any contact with the outside world except for the electric signals fed to us by the computer. Consider if this scenario were to obtain, what the referent of a term like "vat" would be according to conceptual role semantics. It is difficult to come up with the complete conceptual role of the word "vat". For my purposes, however, we can restrict our attention to rules that have to do with our perceptions and actions[6]. For example, as a first pass, we may consider the following rules for the term "vat" given that we are not brains in vats:

- "Vat" introduction: when a speaker sees a vat, the speaker is entitled to apply the term "vat".
- "Vat" elimination: when a speaker judges "X is a vat", and the speaker has an intention with the content "perform A if there is a vat", the speaker is committed to perform A.

If these rules do characterize our use of "vat", it is obvious what "vat" refers to; a vat. However, on most accounts of conceptual role, if we were brains-in-vats, the above rule could not have governed our use of the word "vat". This is because if human beings have always been brains-in-vats, then never in our history do we see a vat or perform any kind of action on a vat. Instead, we have, on numerous occasions, had the visual impression of a vat and the sensation of performing

[6] Not every conceptual-role theorist endorses the claim that actions and perceptions contribute conceptual or inferential roles (Chalmers, 2021; Harman, 1999). I will restrict my attention to theories that count actions and perceptions as contributing to conceptual roles.

some kind of action on a vat, but as the scenario stipulates, such sensations are the product of a powerful computer. As such, the tokens of "vat" would never have played the above causal role, nor would our community ever have the norms of using "vat" in the above way. Therefore, the above will not characterize the conceptual role of "vat" if conceptual roles are determined by causal roles or linguistic norms if we were brains in vats. Moreover, given that the above rule does not even capture true regularities of how the BIVs use the word "vat", it is unlikely any account of conceptual role semantics will entail that the above rules would be characteristic of the conceptual role of "vat".

So what would "vat" refer to if we were brains in vats? Supposing that there is a pattern of electric signal that is characteristic of vat-related impressions (call it "vat-signal"), then the following conceptual role is more likely to be correct (call it R1):

- "Vat" introduction: when a speaker receives vat-signal, the speaker is entitled to apply the term "vat"
- "Vat" elimination: when a speaker judges "X is a vat", and the speaker has an intention with the content "perform A if there is a vat", the speaker is committed to output signals associated with "performing A".

I claim that if we were brains in vats, the above rules would specify an operative norm within our communities: those who violate the above without good reason (evidence to the contrary in the case of introduction rule and overriding desire/reason in the cause of elimination rule) are deemed irrational by the community. Moreover, the above rules would also specify the causal role "vat" plays in our psychology: reception of vat-signal causes our judgement or belief involving token of "vat", and such judgement or belief causes our action (in conjunction with the right desire). Therefore, if we were brains in vats, according to conceptual role semantics, our

word “vats” would not refer to vats. This is because to maximize the rationality of those following the R1, the term “vats” must refer to vat-signals.[7]

This is not a surprising result. In the original thought experiment, Putnam argued that this is exactly the result if a causal theory of reference is true. (1981) I have only extended the same result to cover conceptual role semantics. However, there is a problem: suppose that we are brains-in-vats but we are not the only creatures in the universe. Somewhere in the galaxy, there is civilization that, by some miracles, speaks a language almost identical to English. Suppose that that civilization, in their travels, found the many envatted brains on Earth. Excited to communicate with other intelligent beings, they plug themselves into the super computer as well. After plugging themselves in, it is natural that they will try to reveal our sad truth to us. And, naturally, we will disagree with them at first. However, when they utter the string of phonemes homophonic to “you are brains-in-vats”, they mean that (P1) we the Earthlings are brains-in-vats. But, when we utter the sound homophonic to “we are not brains-in-vats”, we would mean (roughly) that (P2) we the Earthlings are not brains-signal-in-vat-signal. Both P1 and P2 are true. In other words, the aliens and we are in fact speaking different languages, and our apparent disagreement turns out to be illusory. P1 and P2 are not incompatible, and they are not about the same topic. This is not the right result: in the scenario, we are clearly having a substantive disagreement with the aliens, and if conceptual role semantics entails otherwise, something is wrong with conceptual role semantics. This reaction is shared by Thomas Nagel. In reaction to Putnam’s thought experiment, Nagel says:

> Such theories are refuted by the evident possibility and intelligibility of skepticism, which reveals that by “tree” I don’t mean just anything that is causally

[7] Vat-signals are what maximize rationality as opposed to vats since it is rational for a speaker to apply a word that refers to electronic signals when receiving electronic signals, and conversely it is irrational for a speaker to apply a word that refers to vat when receiving electronic signals.

> responsible for my impressions of trees...Since those things could conceivably not be trees, any theory that says they have to be is wrong. (Nagel, 1986, p. 73)

Here, Nagel is talking about the causal theory of reference, but the same lesson applies: if a theory of reference, be it causal or conceptual role, entails that skepticism (that we are brains-in-vats) is unintelligible because our concept/word "vat" cannot refer to vats if we really are brains-in-vats, then that theory must be wrong. Although, this is too quick: my aim here is not to argue that conceptual role semantics or causal theories are wrong, but only that they cannot be right *as they stand*. They both make certain disagreements illusory when those disagreements are perfectly substantive, so more needs to be said about them.

To take stock, I have presented two different problems in this chapter. In section 2.1, I presented the problem of indeterminacy that targets descriptivist and some interpretationist metasemantic theories. There, the problem is that the constraint provided by the metasemantic theory underdetermines the semantic theory: the intuitive semantic theory and its permuted cousin can perform equally well in getting the right truth-value (see 2.1.1), sets of possible worlds (see 2.1.2), and truth-conditions (see 2.1.3). In section 2.2, I presented the problem of disagreement. This problem presents a challenge to metasemantic theories of all stripes. In particular, I have argued that descriptivism, causal theories, and conceptual role semantics all fail to meet the challenge: they each fail to account for some apparently substantive disagreements.

A caveat: these two problems are meant to motivate a solution. However, that these "problems" are genuinely problematic is not uncontroversial: for example, Davidson himself appears to accept that referent of subsentential expression is inscrutable (1979). Moreover, there is no shortage of expressivist or relativist theories about morality that can accommodate the illusion of disagreement in the case of moral twin Earth, and Putnam himself fully endorses the

result that the skeptical hypothesis is self-refuting. However, I will set these aside and assume that the two problems above do call out for a solution.

Chapter 3: Reference Magnetism

In chapter 2, I laid out two problems. To recap, according to the first problem, for any language with a universe of discourse bigger than 1, we can construct a deviant semantic theory that preserves the semantic value of whole sentences. According to the second problem, some of our best metasemantic theories make certain substantive disagreements illusory because the reference-fixing mechanisms proposed by those metasemantic theories shift the topic of the disagreeing parties. In this chapter, I will introduce a solution to these problems, and evaluate some ways of sharpening it. The solution proposes an extra constraint on the right semantic theory, and it can be regarded as an added component of a reference-fixing mechanism. I will start, in section 3.1, by laying out the structure of this extra constraint. Then, in section 3.2, 3.3, and 3.4, I evaluate attempts at analyzing this added constraint: I will argue that none of those attempts are satisfactory.

3.1 Introducing Eligibility

To begin with, let's review the problem of indeterminacy: the problem arises for metasemantic theories according to which the only constraint for the correct semantic theory is that it must get the semantic value of whole sentences correct. However, as we saw, the constraint proposed by global descriptivism, intensional global descriptivism, and Davidsonian Interpretationism all fail to determine the correct semantic theory. By permuting the universe of discourse, we can construct deviant semantic theories that fulfill whatever constraints proposed by these metasemantic theories equally well as the correct semantic theory. Let's call the correct theory T1, and the deviant theory T2. To solve the problem of indeterminacy, the added constraint must be able to help us rule out T2.

Here is an observation: the referent of predicates assigned by deviant semantic theories thus constructed seem to often look disjunctive and complicated. Suppose that T1 is the intuitive semantic theory for English, and T2 is the result of permuting Wittgenstein with Elon Musk and leaving everything else the same. To preserve truth, truth-conditions, and sets of possible worlds assigned to whole sentences, the predicates of English must be reinterpreted under T2. Generally, if a predicate has an extension that includes Wittgenstein but not Elon Musk, under T2, the predicate must be reinterpreted to refer to a set that includes Elon Musk but not Wittgenstein (and vice versa). For example, under T2, the predicate "being a philosopher" must refer to the property of being Elon Musk or a philosopher excluding Wittgenstein, and the predicate of "living in the 21st century" must refer to the property of being Wittgenstein or anyone living the 21st century excluding Elon Musk. Compared with the property assigned to these two predicates under T1, the property of being a philosopher and the property of living in the 21st century, the properties assigned by T2 look more complicated for no good reason. This points to a possibility for the added constraint: if there is a systematic way of separating "normal" properties from

properties that are weird and gerrymandered for no good reason, then the constraint can simply say that when two semantic theories agree on semantic values assigned to whole sentences, the one that does not assign long and complicated properties to predicates is the correct one. Let's suppose that there is such a way: call properties that are long and complicated "gerrymandered", and call properties that are not gerrymandered "eligible". Section 3.2, 3.3, and 3.4 are all dedicated to theories about what makes a property eligible, but for now, I will simply lay out some more features of eligibility that are required to solve the problem of indeterminacy and disagreement.

To begin with, eligibility is not a matter of all-or-nothing. Specifically, consider the property of being a nucleon. Something is a nucleon just in case it is a proton or an electron. Intuitively, the property of being a proton or a neutron is not as eligible as the property of being a neutron. However, consider the property of being a proton or being a cat, call it "the property of being a catron". The property of being a nucleon certainly seems more eligible than the property of being a catron. Therefore, eligibility is a matter of degree where nucleons are more eligible than catrons but less eligible than the neutrons. More specifically, we can suppose that there is an eligibility score for every property we can refer to.

Second, the eligibility ranking of the properties assigned in a semantic theory can be summed. We can imagine that every property is assigned a numerical eligibility score where the more eligible the property the higher the score. Therefore, for any semantic theory, we can "add up" the eligibility scores of the properties that are assigned to predicates according to that theory. Call this the "overall eligibility" of a semantic theory. I will also define the eligibility of a semantic theory such that a semantic theory T1 is more eligible than T2, if the overall eligibility of T1 is higher than T2.

With these two features laid out, we can formulate the following constraint based on eligibility to solve the problem of indeterminacy. Again, the problem of indeterminacy arises because the constraint that the correct semantic theory must get the semantic value of whole sentences correct is not sufficient to pick out a unique semantic theory. However, among all theories that pass this first constraint, our observation has it that most of these theories assign gerrymandered properties as the referents of predicates whereas, intuitively, our predicates refer to very eligible properties. Therefore, to rule out these unintuitive theories, we simply need to add a constraint that requires the correct theory to maximize its eligibility score. The following constraint does the job:

> (TIEBREAKING*): Amongst the theories that get the semantic value of whole sentences correct, the most eligible theory is the correct one.

With this added constraint, the problem of indeterminacy seems resolved: the deviant semantic theory that is constructed as a result of permuting Wittgenstein with Elon Musk can be ruled out because the properties it assigns to predicates like "philosopher" and "living is the 21th century" are much less eligible than the properties assigned according to the intuitive semantic theory.

However, (TIEBREAKING*) is not sufficient to solve the problem of disagreement for two reasons. First, the problem of disagreement arises for different metasemantic theories. Whereas the metasemantic theories that suffer from the problem of indeterminacy determine reference holistically, this is not necessarily the case for causal theories and conceptual role semantics. This is to say that descriptivism and Davidsonian interpretationism fix the referent of names and predicates by first picking out the correct semantic theory and then looking at the referent assigned to names and predicates by the correct semantic theory. Causal theories, on the other hand, fix the referent of names and predicates by looking at individual names or predicates

and the entities or properties that are causally related to them in the appropriate way. In other words, they start at individual predicates and names, and then build up to the correct semantic theory, whereas descriptivism and Davidsonian interpretationism start by picking out the correct semantic theory and go from there. Therefore, to accommodate theories that do not fix reference holistically, the added constraint cannot appeal to semantic theories. To stay neutral about the right metasemantic theory, let's say that M is the right reference-fixing mechanism such that necessarily, a term T relates to its referent X by M. If causal theories are right, T relates to X by M iff there is an appropriate causal relation between T and X; if conceptual role semantics is right, T relates to X by M iff X makes the conceptual role of T correct. Therefore, we can supplement (TIEBREAKING*) by the following:

(TIEBREAKING): T refers to X if and only if amongst the properties that relate to T by reference-fixing mechanism M, X is the most eligible.

(TIEBREAKING) covers metasemantic theory that fixes reference for terms one at a time, but it is still not sufficient to solve the problem of disagreement. Recall that the problem of disagreement has the following form: we have a reference-fixing mechanism M and two apparently disagreeing parties A and B. However, despite the fact that A utters "not-P" when B utters "P", by M, "not-P" uttered by A is not incompatible with "P" uttered by B. Therefore, the problem is not that multiple properties relate to a single term by M, as it is the case with the problem of indeterminacy. Rather, the problem is that for A and B to have genuine disagreement, one of them has to refer to properties that do not relate to their utterance by M. Therefore, we need a constraint stronger than (TIEBREAKING), such as the following

(OVERRIDING): T refers to X iff X best balances the following two factors:

(1) X (approximately) relates to T by M

(2) X is eligible

Since the problem of disagreement arises for holistic metasemantic theories as well, we also need a counterpart to (OVERRIDING) that can apply to those theories. The following will do:

(OVERRIDING*): T refers to X iff T refers to X according to the semantic theory S where S is the theory that best balances the following two factors:

(3) The semantic values of whole sentences are (approximately) right under S

(4) S is eligible

With these added constraints, we have the start of a solution to the problem of indeterminacy and the problem of disagreement. However, this is all based on the assumption that there is a systematic way of distinguishing eligible properties from gerrymandered ones. In the next three sections, I will present some proposed analysis of eligibility. In section 3.2, I consider a family of theory-external views that analyze eligibility in terms of certain metaphysical notions. In contrast with them, the views presented in section 3.3 analyze eligibility in terms of semantic notions. In section 3.4, I consider a family of views that "reclocates" the application of eligibility. Aside from presenting these views, I will also say why they are unsatisfactory.

3.2. Theory-external Views

The notion of eligibility is first proposed by Lewis in response to Putnam's permutation problem (Lewis, 1983; 1984). Specifically, Lewis thinks that a property is eligible if and only if it is natural. To elaborate, in "New Work for a Theory of Universals", Lewis focuses on the ontological distinction between natural properties that carve nature at its joints and non-natural

properties that do not. Lewis puts this distinction to use in many areas of philosophy, one of them being that natural properties are eligible. In his own words: “eligibility to be referred to is a matter of natural properties" (Lewis, 1983, p. 371). The most salient distinction between natural properties and non-natural properties seems to be this: although any two things can be said to share an infinite number of properties, some things are objectively similar, and some are not. Natural properties, then, contribute to the objective similarities of things. For example, a neutron and a proton share the property of being part of the nucleus of an atom, and a neutron and a cat share the property of being either a neutron or a cat. However, a neutron and a proton are more objectively similar to each other than a proton and a cat are.

There are some ambiguities regarding what exactly this “naturalness” is (Schaffer, 2004). For now, I will assume, with the orthodox reading of Lewis, that a property is perfectly natural if it is metaphysically fundamental. According to Ted Sider, “fundamentality” is a primitive notion, and we should “regard as joint-carving the ideology that is indispensable in your best theory” (2011). In other words, we should think of the primitive categories employed by our best theory as referring to fundamental properties. Moreover, Sider is conservative about which of our theories deserve to count as “best”: specifically, only concepts used in fundamental physics carve nature at its joints.

Sider further makes the connection between fundamental properties and epistemic values. Specifically, just as tracking truth is an aim for our epistemic endeavors, so is tracking the fundamental structure of the world, and we do so by referring to fundamental properties. For example, call something “grue” if it is green before the year 3000 and blue after, and call something “bleen” if it is blue before 3000 and green after. Suppose that there is a community of scientists who explain the fact that emeralds reflect green light by appealing to the fact that

emeralds are grue before the year 3000 and bleen after 3000. According to Sider, these scientists' explanations, although true, fall short of our explanation: emeralds reflect green light because emeralds are green. Further, the reason our explanation is better than theirs is that our explanation refers to properties that are more fundamental.

Moreover, I have mentioned that eligibility is not a matter of all-or-nothing. Here is how we may implement this idea under the analysis of eligibility as fundamental. Call a language "canonical" if all the predicates in that language refer to metaphysical fundamental properties. Then, a property A is said to be more eligible than a property B if the shortest definition of a predicate referring to A in the canonical language is shorter than the shortest definition of a predicate referring to B in the canonical language. Here, the length of the definition can be measured by a weighted count of logical connectives. For example, a definition with 2 disjunctions will be shorter than one with 3 disjunctions (since there are fewer connectives), but nonetheless longer than one with 2 conjunctions (assuming that a disjunction is weighted more than a conjunction). And a predicate that designates a metaphysically fundamental property will be perfectly eligible, with a length of 0.[8] The overall eligibility, then, will be the numerical sum of the degree of eligibility (as a weighted count of connectives) of the predicates according to that theory.

To motivate this analysis of eligibility, like any other empirical theory, a semantic theory has certain desiderata. If Sider is right, then tracking the fundamental structure of the world will be one of them. Therefore, when formulating our best semantic theory, we should seek to appeal to fundamental properties, or at least properties not too (definitionally) distant from fundamental

[8] It has been questioned whether definitional length is the right way to measure degrees of naturalness (Dunaway, 2020). For example, Sider has suggested the overall naturalness of a semantic theory is measured by the degree of naturalness of the reference relation according to that theory (2011). However, this approach does not come with a way of directly comparing the naturalness of semantic theories, so I will have to set it aside for this thesis.

properties. Even if one is suspicious of the connection between fundamentality and epistemic value, it is uncontroversial that simplicity is such a virtue. However, a vexing question is "how to measure the simplicity of a theory?" A syntactic measure will not do by itself: for any arbitrarily complicated theory T, one can define a new predicate "T" such for all x, Tx if and only if x satisfies T. This maneuver simplifies T down to one sentence. A proposal, originating from David Lewis, is that the simplicity of a theory must be measured in language that only refers to fundamental properties (Lewis, 1983). If this is true, then the simplicity of semantic theory will also be measured in a language that only refers to fundamental properties. Moreover, since degree of eligibility, analyzed as fundamentality, precisely measures the syntactic simplicity of the property in a language that only refers to fundamental properties, the simplest semantic theory will also be the most eligible one.

Finally, note that this analysis of eligibility as naturalness is external: facts about eligibility have nothing to do with human activity, and everything to do with the world. However, the notion of "naturalness" is ambiguous. In fact, it is argued that there are two notions of "naturalness" implicit in Lewis's analysis pulling in different directions (Schaffer, 2004). In the following, I will consider both notions, and argue that eligibility cannot be either.

3.2.1. Mathematical World(s): Why External Analysis Fails

Although the analysis of eligibility as fundamental naturalness is well-motivated, it is likely mistaken. In this subsection, I present Robert Williams's objection to the proposal (Williams, 2007).

The argument hinges on two claims, so I will say more about what they are before going into the argument. Claim (1): the degree of eligibility of many properties differs from one

possible world to another. Specifically, although electrons and quarks figure in the laws of fundamental physics in the actual world, there is a possible world in which their behavior emerges from the behavior of more fundamental particles. Call those particles "fundamentrons". In a world where fundamentrons exist, they, and not electrons and quarks, will figure in the laws of fundamental physics. Therefore, in such a world, fundamentrons, and not electrons and quarks, will be metaphysically fundamental. Since the eligibility score (degree of eligibility) of any property measures the length of its definition in a language whose primitive predicates refer to metaphysically fundamental properties, the eligibility score of a property like H2O is lower in such a world than it is in the actual world, owing to the increased definition length of electrons and quark in terms of fundamentrons (while the length of the definition of H2O in terms of electrons and quarks remains the same). Claim (2): the eligibility score of purely mathematical properties remains the same in all possible worlds. This is because fundamental mathematical properties, whatever they are, must be necessarily fundamental.[9]

If these two claims are granted, we may construct a crazy semantic theory for English that not only gets the right semantic values, but also beats the intuitive semantic theory, one according to which "Wittgenstein" refers to WIttgenstein and so on, in overall eligibility. Here is how. We pick a bijection such that (1) any concrete object is mapped to a natural number from 1 to n where n is the number of concrete objects in the universe of discourse, (2) any natural number m is mapped to the number m+n, (3) and any other object is mapped to itself (Bays, 2007). Let σ be such a function. Then, we have something like:

- σ(Wittgenstein) = 1

[9] One may worry about whether the notion of fundamentality can be applied to mathematical properties. But see Sider (2009; 2013) and Lewis (1983) for examples.

- σ(Kripke) = 2
- …
- σ(Russell) = 78
- ...
- $\sigma(1) = 1+n$
- ...
- $\sigma(\pi) = \pi$

Then, following the same procedure in 2.1, we can construct a deviant semantic theory for English that makes exactly the same English sentences true as the intuitive theory. In this deviant theory, a predicate in English like "being a philosopher" will be assigned the property of belonging to the set {1, 2, 78...} and a predicate like "being prime" will be assigned the property of being a number x such that x-n is prime. Call this semantic theory "Math-World theory." This, then, is a semantic theory that exclusively assigns non-physical properties to English predicates.

As I have proven in 2.1, the resulting Math-World theory will get the same semantic value for whole sentences as the intuitive theory. Further, if claim (1) and claim (2) are granted, it is possible that the Math-World theory is more natural than the standard theory. Specifically, let $<W_1, W_2, W_3\ldots>$ be an infinite series of possible worlds that exhibits two features: first, every world in the series is a non-semantic macroscopic duplicate of the actual world, where (by stipulation) two worlds are macroscopic duplicates just in case that they share any non-semantic fact in which only properties less fundamental than electrons and quarks figure. Second, the series is ordered by the degree of fundamentality of electrons and quarks in each world. Specifically, the electrons and quarks become less fundamental along with the series. So, for any pair of successive worlds W_i and W_{i+1} in the series, the behaviors of particles that is fundamental

in Wi emerges from behaviors of more fundamental particles in W_{i+1}. Therefore, by claim (1), as the properties the standard theory assigned to predicates get less and less fundamental, the overall eligibility of the standard theory decreases as the series goes on. However, by claim (2), the overall eligibility of the Math-World theory does not change because it only assigns to predicates mathematical properties whose complexity scores are constant across worlds. As such, there will be some W_n in the series where the overall eligibility of the Math-World theory overtakes that of the standard theory. Further, for all we know, the actual world could be Wn, so it follows from (TIEBREAKING*) that it is epistemically possible that the true semantic theory for English is the Math-World theory. But this is absurd. Even if the actual world is not like W_n, we should still expect that in W_n, since it is a macroscopic duplicate of the actual world, there is a community of language users who speak a language that is very much like English, and that the true semantic theory for that language should be very much like the standard theory. However, (TIEBREAKING*) and the analysis of eligibility presented here has it that there will be worlds, such as Wn, in which the Math-World theory is true of languages that are phonetically and orthographically indistinguishable to English in an environment that is macroscopically identical to ours. But this is absurd too. Thus, according to this problem, (TIEBREAKING*) and the theory-external analysis of eligibility leads to absurd results, and one of them must go. I will now turn to some potential ways to resist this problem.

3.2.2. Consideration about Modality?

The first worry is that under the Math-World theory, a sentence like "Kripke is a philosopher" expresses the necessary truth that the number 2 is a member of the set {1, 2…}. Weatherson (2013), for example, briefly raises this worry. While Weatherson does not press this

point, one may worry that this result rules out the Math-World theory since the sentence “Kripke is a philosopher” only expresses a contingent claim. However, this is not the case. We may simply extend the math-world theory by mapping everything in the universe of discourse to a unique natural number *in every possible world*.[10] Then, we can follow the procedure laid out in (2.1.2) to construct the deviant theory. It is easy to see that the intensions (construed as functions from possible worlds to truth values) of whole sentences are preserved.

3.2.3. Alternative Theory-External Analysis?

In the paper where he raises the math-world problem, Williams also mentions a potential solution:

> if there can be ontologically emergent universals at a relatively macro level, there is no longer an argument that "underlying" reality will increase the logical distance between standard macroproperties and those that are perfectly natural or that correspond to universals. (p. 394)

Williams suggests that the problem may be avoided if we shift to a different conception of naturalness. The conception Williams points to here is scientific naturalness. In the following, I will introduce the idea of scientific naturalness before raising an objection against it.

Following Jonathan Schaffer, I distinguish between fundamental naturalness and scientific naturalness by the roles they play in our metaphysical theorizing. Specifically, fundamental properties form the ontological basis of the world, whereas scientifically natural properties figure in scientific laws at all levels of reality. In fact, I will assume that a property is

[10] Since everything in the universe of discourse is mapped to the same natural number in any possible world, rigidity is preserved.

scientifically natural if and only if it figures in a scientific law of nature that is not limited to fundamental physics. I take this bi-conditional to be true without assuming a direction of explanation: it could be that figuring in a system of regularities that exhibit enough theoretical virtue make a property scientifically natural, or it could be that referencing scientifically natural properties is necessary for a system of regularities to qualify as a law of nature (Eddon & Meacham, 2015). Either way, the bi-conditional is all I need. Given this bi-conditional, which properties are scientifically natural will covary with which systems of laws are good enough: it is commonly assumed that physics, chemistry, and biology qualify. Therefore, being an electron is both fundamentally and scientifically natural in the actual world, but whereas being a water molecule is scientifically natural (as it figures it laws of chemistry), it is not fundamentally natural as facts about water molecules depend on further facts about its microphysical structure.

Regarding how scientific naturalness enters into an analysis of eligibility, however, there is a divergence of opinion. Assuming that scientifically natural properties figure into the laws of nature, and assuming that the laws of nature are contingent, facts about which properties are scientifically natural are contingent. In light of this, according to one view, a property is perfectly eligible in a possible world W if it is scientifically natural in W (Williams, 2015). As such, facts about eligibility will be contingent as well. According to another view, a property is perfectly eligible in W if it is scientifically natural in the actual world (Dunaway & McPherson, 2016). According to this latter view, facts about eligibility are necessary facts. Let's call the first view the contingent view, and the latter the necessary view. Both views will be discussed in the next section.

Arguably, the scientific naturalness analysis is well-motivated as well. Specifically, just as we may measure the simplicity of a theory in terms of fundamental properties, we may also

measure it in terms of scientifically natural properties. As such, the simplicity of a semantic theory will also be measured by the definitional distance of the properties the theory assigns to predicates in terms of scientifically natural properties: this, again, is just a degree-of-eligibility measure.

Does analyzing eligibility in terms of scientific naturalness succeed in handling the math-world problems? I will argue: no. However, given the characterization of scientific naturalness above, Williams's argument strikes back: consider a possible world, W^, such that properties instantiated in W^ behave completely chaotically (and thus do not feature in any law of nature in that world).[11] However, for a period of around 150,000 years, nomological orderliness briefly arises from constant flux and properties in W^ appear to behave in a law-like way. Specifically, instances of a property in W^ that is uninstantiated in the actual world happen to behave like the water for that period. Call it "water^". In a similar manner, throughout those 150,000 years, exotic properties such as humans^, quarks^, and fish^ happen to behave like humans, quarks, and fish. Suppose that the species instantiating the property of humans^ evolved to speak a language that is orthographically and phonetically indistinguishable to English, namely English^. Further, suppose that their immediate physical environment is qualitatively indistinguishable from ours: where in the actual world, fish live in water and are constituted from quarks, W^ has fish^ in water^ made up of quarks^. We should expect that the true semantic theory for English^ should be similar to the true semantic theory for English. That is, for any English predicate P that refers to F, the predicate P^ that is indistinguishable from P in English^ should refer to the property F^. However, F^, along with water^, fish^, and quarks^ and their ilk are not scientifically natural – it's merely as if they were scientifically natural, for the limited span of 150,000 years. Moreover,

[11] Here, I assume a broadly best-system account of laws of nature.

we can construct W^ such that these properties have arbitrarily low scientific naturalness scores overall. We can do so simply by making the world arbitrarily chaotic outside the special time frame of 150,000 years. However, the overall naturalness of the Math-World theory remains constant even on a scientific conception of naturalness. As such, there will be a possible world where, on the scientific conception of naturalness, the naturalness score of the Math-World theory outcompetes that of the intuitive theory (where a predicate indistinguishable to "water" refers to water^). It should be clear by now that this line of thought repeats Williams's argument against fundamental naturalness.

3.2.4 Problem of Guiding

At this point, proponents of the theory-external analysis may reply that perhaps reference magnetism cannot save global descriptivism from permutations, but that this is the fault of global descriptivism. As previous sections showed, there are plenty of metasemantic theories that need reference magnetism and do not face the permutation problem. So, here is a more general objection to the theory-external analysis.

In "Putnam's Paradox", Lewis quotes Putnam's criticism of his view for being spooky and medieval and responds "Anyway, what's wrong with sounding medieval? If the medievals recognised objective joints in the world – as I take it they did, realists and nominalists alike - more power to them." (1984). Here are some ways that the "medieval-sounding" proposal may pose a problem.

One way to sharpen Putnam's complaint is to ask why we should *maximize* overall naturalness in the first place? All else being equal, why isn't the 2nd or the 17th most (overall) natural interpretation the correct one? The decision that the most natural

semantic theory is the right one seems arbitrary (Wright, 2012). For one, there seems to be prima facie cases where maximizing naturalness does not give us the right semantic theory. According to some theorists, considerations about degree of naturalness commit us to ruling out certain ethical theories such as particularism and Rossian pluralism due to their low degree of naturalness (Morriski, 2020). This, however, seems wrong. Further, it is often cited as a requirement that the meaning-determining fact "guides" the (correct) utterances of the speakers[12]. However, as speakers, we are normally not sensitive to the degrees of naturalness of the properties we talk about (except perhaps in the contexts of scientific and philosophical inquiries): we do not have any idea of what degrees of naturalness cars and tables have, nor do we care. As such, it is difficult to see how facts about naturalness may "guide" our utterances of words like "cars" and "tables".

What I offered here is, of course, not intended to be decisive against the theory-external analysis. However, in conjunction with the math-world problem, they provide, I hope, sufficient motivation for us to look elsewhere.

3.3 Theory-internal Analysis,

In the last two sections, I have discussed the theory-external analysis of eligibility. In this section, I will introduce the theory-internal analysis of eligibility and a problem it faces. Specifically, both analyses accept the bi-conditional that a property is eligible if and only if it is eligible according to the true theory of eligibility. According to the theory-external analysis, then, the direction of explanation flows from left to right: the fact that a particular theory of eligibility is true is explained by the fact it picks out properties that are eligible independent of

[12] This problem is raised by Merino-Rajme (2015). The requirement of guiding stems out of rule-following considerations (Wright, 2017; Verheggen, 2011).

our theorizing. A theory-internal analysis, on the other hand, has the direction of explanation flowing right to left: facts about which properties are eligible are explained by the fact that they are eligible according to our best theory of eligibility.[13]

An obvious problem for the theory-internal analysis, then, is that it seems like the problems in 2.1 can be repeated for a theory of eligibility. In other words, there will be competing interpretations of a theory of eligibility that disagree about which properties are eligible. Supposedly, facts about eligibility can help us decide which interpretation is the correct one, but here it is precisely facts about eligibility that hang in the air. Call this problem the "just-more-theory maneuver". This problem has been raised, more generally, for any analysis of eligibility whatsoever (Putnam, 1977; Taylor, 1992). However, I will argue that the scope of this problem is much smaller: it only affects theory-internal analyses, and even there, its effect is nuanced. In the following, I will first introduce two versions of theory-internal analysis currently in the literature in 3.3.1 before discussing the just-more-theory maneuver and related problems for theory-internal analyses in 3.3.2.

3.3.1. Theory-Internal Analysis

I will distinguish between two kinds of theory-internal analysis: according to the first, facts about eligibility depend on the theory of eligibility held by speakers of the object language. According to the second, facts about eligibility depend on the theory of eligibility held by the speakers of the metalanguage.

[13] An analogous case is Lewis's theory of the laws of nature where our best system of generalization of the goings-on in the world explains the laws of nature.

An example of a theory-internal analysis where facts about eligibility depend on the theory of eligibility held by speakers of the object language is explored by Chalmers (2012). Let's call the theory "naturalness descriptivism". According to naturalness descriptivism, the meaning-determining set of the object language L will include a theory of naturalness, a set of sentences using a primitive predicate that has the set of natural properties as its extension. (Recall that a meaning-determining set of L is a set of utterances in L such that a good semantic theory of L should maximize the number of true sentences in the set.) In English, for example, a theory of naturalness will include such sentences as "electrons are perfectly natural" and "being green is more natural than being grue". In the case where speakers of the object language do not in fact have such a theory, we can attribute a theory to them based on their dispositions. This naturalness theory, then, plays two roles: first, it fixes the referent of the predicate "natural". Second, it tells us which predicates in the object language refer to natural properties and which ones do not. To illustrate this with a toy example, suppose that our naturalness theory includes the sentence "being green is a natural property" and "being grue is not a natural property". This theory, then, has two roles: first, it narrows down the extension of "natural". We now know that whatever property referred to by "green" is in the extension of "natural", but the property referred to by "grue" is not. Therefore, with a sufficiently detailed theory, we may hope to pin down the property referred to by "natural". Second, because a good semantic theory maximizes the number of true sentences in the evidence set, there is a pressure to assign a property in the extension of "natural" to the predicate "green", but there is no such pressure for "grue".

According to Chalmers, such a theory-internal analysis has two advantages over a theory-external analysis. Firstly, according to Lewis's theory-external analysis, we should indiscriminately assign the most natural referent to any predicate. More specifically, as long as

the overall naturalness of the semantic theory is maximized, it does not matter whether we have assigned the most natural referent to “electron”, “green”, or “grue”. However, this is counter-intuitive. We have independent reasons to think that some predicates refer to more natural properties: we think that the referent of “green” should be relatively natural, but no one will fault a semantic theory for assigning a gerrymandered property to “grue”. Further, we think that “grue” refers to an unnatural property not simply because charity trumps naturalness in this particular case, but because philosophers have stipulated that “grue” refers to an unnatural property to begin with.

Secondly, in the true semantic theory for a language, consideration of naturalness simply does not trump the informed reflection of the speakers of that language. For example, if upon reflection, we decide that the predicate “grue” refers to a gerrymandered property, then even if the semantic theory maximizes truths in the meaning-determining set and its own naturalness assigns a more natural referent to it, the referent of “grue” nonetheless remains gerrymandered. Therefore, our theory of naturalness seems to have more sway in the true semantic theory than simply assigning the most natural referent possible to every predicate without discrimination.

The second kind of theory-internal analysis locates the theory of eligibility not in the meaning-determining set of the object language but in the metalanguage. To modify the analysis above, note that we, the speakers of the metalanguage, also have a theory of naturalness and plausibly this theory should play a role in reference-fixing. This theory, of course, must fulfill some desiderata. For one, it must be true of our own metalanguage. This provides us a way to discover our theory of naturalness. Specifically, when the metalanguage and the object language are the same, the true semantic theory can be independently known. This, then, allows us to work backwards to fix a theory of eligibility in the metalanguage. Then, we can apply this theory of

eligibility to any other object language such that the semantic theory for any language is the one that maximizes truths in the meaning-determining set and its own eligibility score according to our theory of eligibility. One may wonder how this analysis differs from a theory-external one. The difference is this: in a theory-external analysis, the truth of a theory of eligibility depends on external facts about eligibility, whereas, here, we have a theory of eligibility formulated in the metalanguage and, if the analysis is right, the property of eligibility is whatever satisfies this theory. Again, the direction of explanation is reversed.

An example of the second kind of theory-internal analysis is explored by Williams (2015). According to Williams's suggestion, instead of maximizing its own naturalness, a good semantic theory should maximize its familiarity in the metalanguage. Specifically, a property is familiar if it is referred to by a familiar predicate in the metalanguage, and a predicate is familiar if it is commonly used by speakers of the metalanguage. (Alternatively, we can be more liberal about which predicates count as familiar. For example, we may hold that a predicate is familiar as long as it is in the lexicon of the metalanguage. For the purpose of my discussion, this distinction does not matter too much.) Therefore, according to this suggestion, the true semantic theory should balance maximizing truths in the meaning-determining set and assigning familiar properties to predicates of object language (i.e. making the object language and the metalanguage co-refer as much as possible). This is an example of a theory-internal analysis that locates the theory of eligibility in the metalanguage. The easiest way to work out a theory of eligibility is simply to make every property that is actually referred to by an English predicate eligible. In other words, since we know that "whale" refers to the property of being a whale and not being a whale or a fish, we know that the former is more eligible than the other.

3.3.2. Just-More-theory

An objection to the theory-internal analysis, initially aimed at the eligibility constraint in general, is the just-more-theory maneuver (Putnam, 1977; Putnam, 1980). In this section, I discuss this objection in detail. I will argue that this objection (1) rules out any theory-internal analysis that locates the theory of eligibility/naturalness in the object language, (2) has no force against a theory-external analysis, and (3) affects the kind of theory-internal analysis that locates the theory of eligibility/naturalness in the metalanguage only if we also hold certain assumptions about nature of semantic concepts. I regard (1) and (2) to be relatively easy to show, so I will argue for them first, before dealing with the more nuanced (3).

To start with the kind of theory-internal analysis that locates the theory of eligibility in the object language, here is how a just-more-theory maneuver would go: take any object language L. If speakers of L already have a theory of eligibility, then include that theory in the meaning-determining set. If not, come up with such a theory based on their disposition and add the theory to the meaning-determining set. However, adding the theory of eligibility does not avoid either the problem of indeterminacy or the problem of disagreement.

Here is why: let's call the meaning-determining set without the theory of eligibility "S" and the meaning-determining set after the theory is added "S+". Let T1 and T2 be two semantic theories that agree on the truth-value of every sentence in S. Then, we can construct semantic theories T1+ and T2+ which extend T1 and T2 respectively while making exactly the sentences true in S+. To elaborate, the added theory of eligibility will be a set of sentences of the form "E(φ)" where "E" is the eligibility predicate and "φ"is some predicate. A semantic theory, then, needs to assign an extension to "E" which will be a set of sets. Then, to preserve equivalence in S+, T1+ will assign to "E" the extension $\{X||F|^1 = X$ and "E(F)" is in the theory of eligibility$\}$

and T2+ will assign to "E" the extension {Y||F|^2 = Y and "E(F)" is in the theory of eligibility}, where ||^1 is the interpretation function for T1 and ||^2 is the interpretation function for T2. Thus, T1+ and T2+ will agree on the truth value of every sentence in the added theory of eligibility, and since the two theories agree on the truth-value of every sentence in the original set S, the two agree on the truth-value of every sentence in S+. As for the problem of disagreement, we can easily see that disputing linguistic communities, the earth people and the twin earthlings for example, will likely have different theories of eligibility. Thus, the dilemma remains: the twin earthlings will reason, by their theory of eligibility, that the earthlings are referring to the property pleasure-maximizing when speaking of "good", while we will reason, by our theory of eligibility, that the twin earthlings are referring to the property of treating others as ends not means when they speak of "good". It is hard to see how this makes any headway.

This objection, however, does not apply to a theory-external analysis (Lewis, 1984; Gardiner, 1995). If a theory-external analysis is true, the fact that our theory of eligibility comes out true under both T1+ and T2+ says nothing about whether T1+ or T2+ actually managed to assign eligible or natural or fundamental or what-have-you properties to predicates. Further, despite the communities' diverging theories of eligibility, there is no guarantee that either theory is true, and the referent of their terms will be eventually determined by what is in fact eligible.

Finally, consider the kind of theory-internal analysis that locates the theory of eligibility in the metalanguage. It has been argued that the just-more-theory maneuver poses a problem for this type of analysis as well (Taylor, 1991). However, I argue that whether this is true depends on whether we accept an important assumption. Specifically, call the view according to which the fact that the sentence "F is eligible" has the content it does is explained by the fact that "F is eligible" represents F as eligible "inflationism". In other words, according to inflationism, facts

about representation broadly considered explain facts about content. And call the denial of inflationism "deflationism". (This way of drawing the distinction is taken from Simpson (2018) who discusses the distinction between moral representationalism and moral expressivism.) Only for those who accept inflationism does the just-more-theory maneuver pose a problem for the kind of theory-internal analysis under consideration here. Specifically, here is why: assume for reductio that inflationism is true. By the theory-internal analysis of eligibility, the true semantic theory of an object language L will be partially explained by the content of the theory of eligibility in the metalanguage L_1. Because of inflationism, the content of "F is eligible" in L_1 is explained by the fact that "F is eligible" represents F as eligible. But the fact that "F is eligible" represents F as eligible will further be explained by the true semantic theory for L_1, which is partially explained by facts about eligibility, which is explained by the content of the theory of eligibility in L_2 that is the metalanguage for L_1. But, by inflationism, the content of the eligibility theory in L_2 is explained by facts about representation, which is explained by the true semantic theory of L_2 which…This ends in an infinite regress. However, if we give up inflationism, then the crucial premise needed for the infinite regress is denied. This preserves the theory-internal analysis that locates the theory of eligibility in the metalanguage from the attacks of the just-more-theory maneuver.

3.4 Relocating Eligibility

The eligibility constraint comes to us by Lewis's proposal in "Putnam's Paradox" and "New Work for a Theory of Universals". In those papers, Lewis appears to advocate for a metasemantic theory where the eligibility constraint and global descriptivism jointly pick out the true semantic theory. However, according to some recent discussions, what Lewis offers in

"Putnam's Paradox" and "New Work for a Theory of Universals" is not more than a heuristic for the theory that I will call "Lewisian Interpretationism" (Schwartz, 2013; Weatherson, 2014). Instead, the discussion points to "Radical Interpretation" (Lewis, 1974) and "Languages and Language" (Lewis, 1975) as presenting the theory that Lewis really endorses, a head-first approach of interpretationism. In this section, then, I aim to discuss Lewisian interpretationism, as it is worked out by its latest defender Williams (2020) and cast some doubts on it.[14]

3.4.1 (Neo)Lewisian Interpretationism

In this subsection, I present Lewsian Interpretationism a la Williams while pointing out places where other Lewsian interpretationists disagree with Williams. I will also derive the naturalness constraint as a result of Lewsian Interpretationism at the end.

The Lewisian interpretationists answer the metasemantic question in two steps. First, the Lewsian interpretationists fix the correct interpretation of the mental content of the language users. Specifically, an interpretation of mental content is a function that maps a language user at an instance in time to a pair <*belief*, *desire*> where "*belief*" stands for a set of propositions that the language user takes to be true (or, alternatively, a function from propositions to credence scores) and "*desire*" stands for a set of propositions that the language user desires to be true (or, alternatively, a function from propositions to utility scores). According to Lewisian interpretationism,

> (Stage-1) An interpretation <*belief*, *desire*> of the mental content of a language user A at time t is correct if <*belief*, *desire*> maximizes A's rationality at t.

[14] I focus on Williams since it is the most detailed and clearest formulation of the theory. I will note how other Lewisian interpretationists, including Lewis himself, differ from Williams along the way.

Here, a language-user's rationality is evaluated along three dimensions. The language user is rational along the first dimension if they act in a way that is rationalized by their belief and desire set. Specifically, if A performs the action going to the fridge at t_1, then an interpretation of A at t_1 that attributes A with the belief "I (A) can get milk if I go to the fridge" and the desire "I (A) want milk" will obviously rationalize the action. The language-user is rational along the second dimension if their beliefs conform to epistemological norms. For example if extreme Cartesian foundationalism is true, then the language user is rational only if they believe only propositions that are entailed by logical truths. Finally, the language user is rational along the third dimension if their desires conform to norms of substantive practical rationality. Such norms might include (but are not limited to) moral norms. Therefore, a speaker who murders for fun fails the standard of rationality along this third dimension. Moreover, a speaker who desires a saucer of mud also fails along this dimension, since (plausibly) it is a non-moral norm that one should not desire a saucer of mud. According to (Stage-1), then, an assignment of belief-desire pairs should maximize the language-user's rationality along these three dimensions.

The second step of Lewisian interpretationism goes as follows: once the mental content of every individual language-user in a linguistic community is fixed, the correct semantic theory for that community is simply picked out by convention. Specifically,

(Stage-2) T is the correct semantic theory for the linguistic community C if according to T, speakers of C

(a) only utter what they believe as a matter of convention

(b) come to believe what others have uttered as a matter of convention

Here, a semantic theory, along with information about the syntax and the meaning of the logical operators of the language, can be seen as a function from an uttered sentence to a proposition. A

convention, here, is a kind of regularity within the community propped up by norms. Specifically, it is proposed that a regularity R within C qualifies as a convention if it fulfills the following criteria:

(1) (nearly) Everyone in C conforms to R

(2) (nearly) Everyone in C believes that everyone conforms to R

(3) The belief that everyone in C believes that everyone conforms to R is a good reason to conform to R

(4) There is a preference in C for conformity to R

(5) There is an alternative regularity R' such that had it satisfied (1) and (2), it would also have satisfied (3) and (4)

(6) (1) to (5) are common knowledge

(1) - (6) are tentative, but let's suppose that something like this is true. To summarize, then, a semantic theory of L is correct if (a) for any sentence S uttered by A, the semantic theory maps S to a proposition that A has high credence in, and (b) for any sentence S uttered to A, the semantic theory maps S to a proposition that A comes to have high credence in.

So far, (Stage-1) and (Stage-2) are (almost) universally held by Lewisian interpretationists even though the specific analysis of a convention and first-order theories about rationality may differ. However, Williams's Lewisian Interpretationism makes three potentially controversial moves. To start with, in order to evaluate the rationality of beliefs and actions, Williams needs an account of the contents of perception (since perceptions confer rationality on beliefs) and an account of intentional action (to distinguish them from involuntary behavior). Here, Williams adopts Karen Neander's theory of perception and extends it to intentional action. Without going into the details, the idea is that a state of the agent's perceptual system, or a state

of the agent's motor system, has a content, or intention, [event E occuring] if the perceptual, or motor, system has the function of being in that state in response to, or in order to bring about, [event E occuring]. This is by no means shared by all Lewsian interpretationists: Lewis himself, for example, does not seem to endorse a teleological approach to perceptual content, and neither do other contemporary Lewisian interpretationists (Pautz, 2021).

Secondly, Williams invokes a fine-grained language of thought as the medium of our belief and desire. Specifically, according to Williams, our beliefs and desires are attitudes towards propositions composed from a rich lexicon of names, predicates, modifiers, and operators in a rule-governed manner. I think that the main motivation is this: supposing that (Stage-1) successfully assigns to each individual in a community a determinate mental state, then there is still the problem of assigning a semantic theory to the whole community. If our beliefs and desires are attitude towards coarse-grained truth-conditions (i.e., a truth-value or a set of possible worlds), so the belief that John is an unmarried male and the belief that John is a bachelor have the same content, then the problem from section 2.1 resurfaces. We can match each sentence with a truth-value or a set of possible worlds through (Stage-2) but doing so undetermined the right semantic theory. Thus, by invoking the language of thought, we can match each sentence with a fine-grained content, and the possibility of underdetermination is ruled out. However, this is controversial. Lewis, for one, thought of beliefs and desire as having coarse-grained truth-condition rather than fine-grained content (1994).

Finally, Williams seems to assume that there is no limit on the kinds of beliefs/desires we can have without a public language. In particular, since mental content is fixed entirely independently of linguistic content, agents may form beliefs with such complex content as "in immediate self-consciousness the simple ego is absolute object, which, however, is for us or in

itself absolute mediation, and has as its essential moment substantial and solid independence" even if the agent has no public language. (Quote from *Phenomenology of Spirit* by Hegel) Again, this is not accepted by every Lewisian Interpretationist. Adam Pautz, for example, thinks that agents with no public language can only form beliefs/desires with "thin" content, that is, content employing only observational concepts and simple logical/mathematical concepts (2021). I will return to these points of controversy later when raising objections to Williams's theory.

What is the upshot of all this? If Lewisian Interpretationism is true, then the naturalness constraint falls out of the theory. Specifically, the naturalness constraint can be derived from (Stage-1) in conjunction with an assumption about the epistemic value of natural properties. Specifically, if it is an epistemic norm that we should hold beliefs in terms of predicates that are either fundamentally or scientifically natural, then an interpretation of an agent's mental content maximizes the rationality of that agent if it assigns beliefs in terms of predicates that are either fundamentally or scientifically natural whenever possible. For example, all else being equal, we should assign the belief "all emeralds are green" rather than "all emeralds are grue", since believing the former makes one more rational.

3.4.2 Why Eligibility is not Derived from Lewisian Interpretationism

In this subsection, I discuss some problems faced by Lewisian Interpretationism.

Problems for (Stage-1)

I will start by noting two problems for (stage-1). To begin with the first, it has been questioned whether Williams's interpretationism is really equivalent to a very substantive kind of

inferentialism (Chalmers, 2022).[15] This is because according to Williams's theory, an interpretation "makes" a speaker rational by making their inferences rational. This requires us to first identify the core inferential roles of each concept and then figure out what semantic values would make those inferences rational. However, if this is the case, the BIV case from 2.2 should apply to Williams's interpretationism as well. Consider again BIVs. The interpretation that makes their inferential practice using, say, the concept "vat" most rational is one that assigns "vat" the semantic value of vat-signals.[16] This, again, undermines their disagreement with speakers that are not envatted. (See 2.2.2 for details.)

According to the second problem, Williams's interpretationism still fails to fully determine semantic values (Buchanan & Dogramaci, 2022; Hattiangadi, 2020). Consider the following case. Suppose that Karl 1 is an emerald enthusiast. Having dug around for emeralds for years and discovered a dozen green emeralds, Karl 1 forms the belief that all emeralds are green. Intuitively, it is possible for Karl 1 to withhold their belief about the color of emeralds in general at this point. Suppose that is what Karl 2 is like: having discovered the same number of green emeralds as Karl 1, Karl 2 withholds their belief. We may suppose that Karl 1 and Karl 2 made their respective judgment ,or the lack thereof, after having gathered the same amount of evidence, and both acted and continue to act in the same way driven by a love for emeralds regardless of their color. The problem for the Lewsian Interpretationist is this: which Karl is more rational? If our theory of epistemic rationality has it that Karl 1 is more rational, then by (Stage-1), Karl 2 is impossible. This is because by (Stage-1) the mental content of an agent is the belief-desire pair that maximizes the rationality of the agent. Therefore, since Karl 1 and Karl 2

[15] It should be noted that even Williams himself agrees with this diagnosis (2022).
[16] A similar argument is put forth by (Dickie, 2022).

have the same evidence, same action and same desire, if the belief that all emeralds are green is more rational given their shared evidence, belief, and desire, then both Karl 1 and Karl 2 will be assigned that belief. However, Karl 2, by stipulation, does not have that belief. This is a contradiction. The same applies if our best epistemic theory has it that Karl 2 is more rational: in that case, Karl 1 would be impossible. If, however, our epistemic theory says that Karl 1 and Karl 2 are equally rational, then the belief of an agent that has Karl's desire, action and evidence would be simply underdetermined by Lewsian Interpretationism.

Problems for (Stage-2)

In "Meaning without Use" (1992), Lewis appeals to the notion of naturalness not at the level of assigning mental content, but the level of assigning linguistic content (choosing a semantic theory) given fixed mental content. Specifically, for every speaker in the linguistic community, assuming that we already know their beliefs, we may collect the sentences they utter, and look for patterns that match sentences with (the content of) beliefs. However, since Lewis thought that the content of beliefs is a set of possible worlds, we only have a match between sentences and sets of possible worlds. Therefore, to avoid the permutation problem in section 2, Lewis still needs the naturalness constraint independent of his interpretationism. Williams avoids this by giving beliefs a language of thought that has its own rules of composition and fine-grained concepts. Then, by (Stage-2) the best semantic theory will be the one that inherits the compositional rules and concepts of the language of thought as much as possible. However, as I have mentioned, this is controversial. If this controversial thesis is discarded, the permutation problem in section 2 re-emerges.

Moreover, consider the similarly controversial assumption that we may form beliefs involving abstract concepts without a public language. Presumably, if the assumption is relaxed, then we will not be able to form beliefs about, for example, electrons without a public language. If so, we may consider an alternative semantic theory for English where "electrons" are assigned the property of being a donkey in a distant possible world or electrons in nearby possible worlds. Assigning this property to "electrons" matches our electrons-related utterances with our beliefs as assigned by the rationality-maximizing interpretation since we have no beliefs concerning electrons in those distant possible worlds independent of our public language, the content of which still hangs in the air. Therefore, a semantic theory that assigns the property of being a donkey in a distant possible world or electrons in nearby possible worlds to "electrons" will be a permissible theory (one just as good as the intuitive theory) on (Stage-2). However, this is absurd.

Therefore, unless we think in a robust language of thought completely independent of our public language, there will be unacceptable indeterminacy at the level of language. One solution to this, of course, is to reintroduce the eligibility constraint at (Stage-2) such that semantic theories that assign gerrymandered properties to "electrons" and permuted semantic theories are ruled out for their low eligibility score.

Chapter 4: Stabilized Standard Account of Reference

In chapter 2, I have discussed two problems for any metasemantic theory: the indeterminacy problem and the disagreement problem. In chapter 3, I have argued that reference magnetism, traditionally formulated, is not a viable solution to the problems. Under one approach discussed in chapter 3, namely William's Lewisian Interpretationism, the role of eligibility, or naturalness in Williams's case, is derived from other more general considerations about rationality (2020). Although I have argued that Williams's approach ultimately does not succeed, I think he is right that the constraints of eligibility on reference are ultimately derived from other considerations. In this chapter, I give my own account about the considerations from which eligibility is derived. To the extent that Williams's interpretationism can be said to be an explanation of reference magnetism, the same thing can be said of my story.

The story I tell builds on Jody Azzouni's solution to the rule-following problem (2017). Therefore, I will start by introducing Azzouni's solution in 4.1, before developing it into a full account in 4.2. Then, in 4.3, I return to the theory of reference magnetism and show how a kind of reference magnetism falls out of the theory developed in 4.2. Finally, I consider some outstanding objections in 4.4 before concluding the chapter.

4.1 The Story of Crusoe

In this section, I give an overview of the proposal that I will build on later. To do so, I start by introducing the rule-following paradox, before presenting Azzouni's solution to it.

4.1.1 Azzouni's Solution to the Rule-Following Puzzle

In *The Rule-Following Paradox and its Implications for Metaphysics*, Jody Azzouni presents a solution to the rule-following problem (2017). Briefly, according to the rule-following problem, facts about a speaker's mental state cannot determine the meaning of the words they are using. As such, the problem can be seen as a challenge to any kind of metasemantic theory that only appeals to the speaker's mental states as the relevant meaning-determining metasemantic facts (Kripke, 1982). Although there are subsequent developments of the problem that generalize the challenge to other metasemantic theories, these developments need not concern us here. In his book, Azzouni focuses on the disposition of individual speakers to apply words as the typical target of the rule-following paradox. Specifically, according to the simplified version of this theory, the meaning of "W" uttered by A is the set of things that A is disposed to apply "W" to. The original rule-following problem, then, issues at least the two following challenges.[17] First, it is simply not the case that we have a disposition to apply the relevant predicate to everything under its extension. For example, suppose that you are faced with a thousand ducks swimming in the lake. The lake in front of you falls under the extension of "having a thousand ducks in it", but unless you have superhuman vision, you would not have a disposition to apply that predicate to the lake in front of you: the number of ducks is simply too large for you to process. Second, we are often disposed to apply words to the wrong things. For example, just as I have a disposition to apply "duck" to ducks under normal lighting situations, I am also disposed to apply "ducks" to geese on a foggy day. There is much more to the rule-following problem than what I have described so far, but this suffices for me to set up Azzouni's solution to the problem.

[17] There is a third challenge regarding the normativity of meaning according to many. (See, for example, (Ginsborg. 2011)) I will ignore this challenge partly because it would take us too far afield and partly because reference magnetism has never claim to solve this problem either so I will relieve myself of the duty as well.

Azzouni, instead of amending the dispositional theory in light of the challenges, tells a story about an *individual* language-user and shows how semantic facts arise from their dispositions alone. Let's call this individual "Crusoe".[18] There are several features about Crusoe's linguistic dispositions and practices. First, if we group the words uttered or written by Crusoe by their phonetic and orthographic features, we will find that Crusoe's disposition to use phonetically and orthographically identical tokens of expressions changes over time.[19] For example, at t1, we may observe that Crusoe is disposed to apply the word "coconut" to coconuts, coconut-looking rocks, and yaconuts (an imaginary fruit that looks like coconut but tastes yucky). Then, at t2, we may observe that Crusoe is disposed to use "coconut" for coconuts and yaconuts but not coconut-looking rocks, and finally at t3, Crusoe applies the word to coconuts exclusively.

Second, Crusoe's disposition to apply words is, by and large, not transparent to Crusoe on introspection alone, but despite this, Crusoe is able to distinguish cases where his application of words leads to success and cases that don't. For example, Crusoe is able to tell that at t1, when he collects what he calls "coconuts", many turn out to be useless rock or rather disgusting fruit; in contrast, his collection of what he calls "coconuts" at t3 has mostly nourishing coconuts. Crusoe is able to recognize this difference and reinforces the kind of disposition he had at t3 over the kind he had at t1 and t2. Again, Crusoe needs not to have introspective access to his disposition to do this: he may simply recognize that at t1, he had consumed certain substance before the activity of collecting what he calls "coconuts", or that he had attended to certain

[18] In Azzouni's book, this individual is called "Crusoe 5", but we need not bother ourselves with Crusoes 1 to 4.

[19] The proposal works back if, instead of grouping words by their phonetical and orthographical features, we have a theory of word individuation ahead of time. I will, however, not enter the discussion about what is right way to individuate words to not add to the already bloated thesis, but only note that most theorists think that individuation of words is independent of their meaning (Kaplan, 1990; Stoljnic, forthcoming). C.f, (Williams, 2020) and (Schroeter & Schroeter, 2015).

details of the fruit at t3 that he had not at t2 or t1. Having recognized this, Crusoe can reinforce the relevant disposition by refraining from the particular substance and consciously attending to the particular detail of the fruit.

Finally, Crusoe regards his previous self as committing a mistake when he regards coconut-looking rock and yaconut as what he now calls “coconut”. In other words, Crusoe has the expressive resources such as “is false”, and because the change of disposition is not introspectively transparent to Crusoe, Crusoe attributes mistakes to his earlier self instead of simply having disposition that leads to less success. Crusoe can even employ a truth-predicate/false-predicate to utter sentences as “I spoke falsely when I called that a ‘coconut’”.

Based on these three features, the appearance of a familiar linguistic practice emerges: since Crusoe is not aware of his constantly changing disposition to use “coconut”, he regards himself as having been using the same word all along. It appears to Crusoe that, in this apparently stable language spoken by him, “coconut” refers to coconut all along. Further, when Crusoe is disposed to apply the word to yaconut and coconut-looking rock at t1, it appears to Crusoe later that he had a disposition to apply the word to the wrong thing (as opposed to having a disposition to use a different word correctly).

Azzouni’s crucial move, then, is to vindicate this appearance. To quote Azzouni’s own words,

> The perspective in terms of [private-language-practice coherence-inducing] dispositions describes what might be called the “engineering” of reference. But, despite this enhanced perspective, our words nevertheless refer to what they refer to, and part of the explanation of how they manage this involves the trajectory over time of our plpci dispositions. (p. 112)

Here, “the perspectives in terms of dispositions” refers to our knowledge that Crusoe’s disposition to use “coconut” changes over time. Despite this knowledge, Azzouni claims, the fact

that “coconuts” refers to coconuts does not change over time. The challenge, then, is to say how the constant semantic facts can emerge from the flux of meaning-determining facts. Azzouni suggests that meeting this challenge “involves the trajectory over time of our plpci dispositions”, but remains cryptic about the detail of such an account.

The big picture, however, seems clear enough. Under a traditional account, a speaker uses a name/predicate correctly just in case they apply the term to its referent/instances of its extension.[20] In other words, the standard of applying a word is explained by the semantic value of the word. Under Azzouni’s solution to the rule-following puzzle, however, the direction of explanation flows the other way. Specifically, the standard of application is explained by the relevant meaning-determining facts (facts about Crusoe’s dispositions for example), and the semantic value of a term is then somehow explained by the standard(s) of applying it over time. The details of exactly how these explanations are supposed to work aside, if Azzouni is right about this, then the changing meaning-determining facts will not be a problem for stable semantic facts, and the rule-following puzzle is avoided. To elaborate, since the referent of a word stays constant despite changing dispositions to apply it, the fact that I do not have the disposition to apply the predicate “having a thousand ducks in it” to some of its instances at a particular time (just as Crusoe does not have the disposition to apply “coconuts” to some coconuts at t1), or the fact that I sometimes have a disposition to apply “duck” to goose (just as Crusoe has the disposition to apply “coconut” to yaconut), does not affect the semantic facts about the predicate “having a thousand ducks in it” or “duck”.

4.1.2 What Comes Next

[20] See (Boghossian, 1989) for example.

I have now introduced Azzouni's solution to the rule-following problem. For my own purposes, I will appropriate his central move, namely that meaning-determining facts such as those proposed by interpretationist theories, causal theories, or inferential-role theories do not ground semantic facts directly. Rather, they ground the standard of application at a specific time, and those standards of application over time ground semantic facts. In the rest of this chapter, I will then flesh out this move into an account of reference that I will call the "stabilized-standard account" (or "SSA"). However, for now, I will lay out some challenges for Azzouni's picture that SSA will have to overcome.

First, note that Azzouni's account of meaning applies to Crusoe, an individual speaker without a community. And Azzouni only considers facts about dispositions to be the relevant facts that fix the standard of application. However, my aim is to explain certain semantic properties of natural language which is spoken by a community of speakers by supplementing a wide range of metasemantic theories not limited to the dispositional theory of meaning. Thus, the first challenge for my SSA will be to generalize Azzouni's solution to apply to a language spoken by a community, and formulate it in a way that is friendly to all stripes of metasemantic theories in the same way a theory of eligibility is. Call this the generalization challenge.

Second, according to Azzouni, the speaker uses a word correctly (according to its standard of application) just in case that the speaker is disposed to use the word that way. This, however, seems too easy; in fact, it almost seems that a speaker cannot fail to use a word correctly. If this is the case with Azzouni's account and, by extension, SSA, then we are in serious trouble because (1) speakers are simply not infallible like that and (2) part of the goal for Azzouni's account is to vindicate Crusoe's attribution of error to his past self, but if his past self is correct, what kind of error is Crusoe attributing and how can such attribution be vindicated?

The challenge may be phrased this way: we need there to be some external constraint on the standard of application at any time so that at a later time, Crusoe may attribute errors to previous standards of application based on that constraint. Call this the external constraint challenge.

Third, as I have mentioned, Azzouni does not give many details about how standards of application give rise to semantic facts. At some points, Azzouni seems to suggest that the semantic value of a term is determined by the standard of applying that word that eventually stabilizes (pp. 114, 116). However, if this is the case, an immediate problem is "what about terms that we keep using in new ways or terms that, by nature, have infinite application? Do they ever stabilize?" My SSA will take Azzouni's suggestion that semantic values are determined by stabilized standard of application, so answering this question will be a challenge to SSA as well. Call this the stabilization challenge.

4.2 SSA

In this section, I will sharpen what I have presented in 4.1. My aim here is not explication but appropriation of Azzouni's solution to the rule-following problem to my own end: eventually to argue for a theory of eligibility that constrains semantic facts. For now, I will provide more detail for SSA and hopefully meet the three challenges I laid out above.

To start with, under SSA, the semantic value of a term is given by the stabilized standard of applying it, so if according to the standard of using the word "coconut" that stabilizes, the word "coconut" is correctly applied to all and only coconuts, the extension of "coconuts" will be the set of all coconuts. Further, the standard of applying a word at a particular time is determined by what has been proposed as meaning-determining facts by various metasemantic theories. Thus, the standard of using "coconuts" at a particular time may be determined by the best

interpretation of the speaker's language at that time, the appropriate causal relation that obtains between "coconut" and a piece of the world at that time, or the inferential patterns of "coconuts" at that time. I will remain neutral about which one of these options is the correct one and speak of a generic "standard-determining fact" as a placeholder.

With this formulation of SSA on the table, I will now turn to three challenges I laid out at the end of the last section and see how SSA deals with them.

4.2.1 The Generalization Challenge

To begin with, I will say how SSA generalizes Azzouni's proposal in two dimensions: first, it can be taken as a supplement to many metasemantic theories and not just the dispositional theory. Second, it applies to a natural language spoken by a community of speakers not only to the "private" language spoken by a lone Crusoe. I do this by showing that the three central features of Crusoe's language that makes Azzouni's proposal, and SSA by extension, viable can be translated to (1) a wider range of metasemantic theories (2) for a community of speakers. Specifically, the three features are (i) Crusoe's disposition changes over time, (ii) the disposition change is subpersonal, and (iii) Crusoe nonetheless takes notice of the result of the changing disposition and attributes errors to himself when his disposition did not lead to success.

To start with (1), I cannot exhaustively show that the standard-determining facts proposed by every metasemantic theory will exhibit these three features, but I will briefly give some examples of how, if we take some prominent metasemantic theories as supplying us with standard-determining facts, the three features are satisfied. To start with simple versions of interpretationism (see 2.1), according to it, the standard-determining facts are facts about what

beliefs we hold, and what is the best interpretation of those beliefs.[21] However, it is plain to see that our beliefs and desire change constantly thus satisfying the first feature. Further, it is also the case that we do not have transparent access to our beliefs (Schroeter, 2012; Boghossian, 1994). This may seem surprising, but part of the reason is that beliefs are individuated by their contents which depend on the best interpretation of those beliefs, but we do not have transparent access to what best interprets our own beliefs (Carruthers, 2011). Finally, when our beliefs change, even when we do not notice such change, the result of such belief change is opaque since having different beliefs lead us to take different actions. Thus, based on the success or failure brought by the different decisions we (and other people but more on that later) make, we can attribute errors. This secures the third feature.

Also consider causal theories (see 2.2 for examples). Whatever is the appropriate causal relation, be it "indication" (Dreske, 1981), evolutionary function (Milikan, 1981), or causal "regulation" (Boyd, 2002), it is unlikely that we have introspective access to these causal mechanisms. This secures the second feature. Further, it is likely that for any given word, the facts about what it indicates, its function, or what regulates its use change over time. Take "fish" for example. There was a time in human history, before the advent of modern biology, when the use of the word "fish" probabilistically depended on the presence of fish-shaped aquatic creature including whales (in other words, the conditional probabilities of there being a fish-shaped creature when someone uttered "fish" was greater than there being any other creature including whale-exclusive fish), the function of the word "fish" was to pick our fish-shaped creature rather than fish (so it was picking out the former that contributed the continued use of the word "fish"),

[21] The story for Lewisian Interpretationism is more complicated. Whether that theory exhibits this feature depends on whether the objections to it in 3.4.2 goes through.

and it was fish-like creature rather than fish that causally regulated our use of it (so our procedures for recognizing what we called "fish", those we thought are experts of what we called "fish" all ensured that the term "fish" was correctly predicated of fish-like creatures not just fish). Thus, given that the word today clearly indicates fish, has the function of picking out fish, and is regulated by fish, the first feature is satisfied as well: the proposed standard-determining fact changes over time. Lastly, what a word indicates, the function of the word, and what regulates our use of it all contribute to our success as a result of having a word that indicates what it does, has the function that it has, and is regulated in the way that it is. Thus, we can attribute errors to ourselves (and others) based on such success (or failure).

Consider, finally, inferential-role theories (see 2.2 for examples). Again, it is plain that the inferential patterns of a word that we use change over time, and since we do not keep conscious track of the inferences we make with most words, such changes happen without our noticing it. This secures the first two features already. Finally, since perception and action features as part of these inferential patterns, we can again take notice of the result of such changes and attribute mistakes when such changes result in failure.

An exhaustive treatment of every metasemantic theories is untenable, but I hope that I have shown that there is no problem generalizing Azzouni's proposal beyond the dispositional theory of meaning as the central features of Crusoe's language are preserved. In the following, I will refer to the feature of a language for which the standard-determining facts (whatever they are) vary over time the "variation feature"; and I will call the feature of a language for which such variation is not transparently accessible to its speaker(s) the "opacity feature"; and I will call the feature of a language where speaker(s) attributes errors based on the result of varying standard-determining facts the "error-attribution feature".

Finally, I claim that a language spoken by a community of speakers can readily exhibit variation, opacity, and error attribution. Take English for an example. Clearly, the standard-determining facts of many English words vary across time (just think about "fish"). Further, there are likely sub-communities of English speakers for whom the same word is used with different standards of application due to varying standard-determining facts (think about how Democrats and Republicans use "critical race theory"). Thus, the variation feature is preserved not just across time, but also extends to cases across subcommunities. Further, opacity is preserved since community-level facts about causal relations, inferential patterns, and best interpretations of collective beliefs are certainly not accessible to introspection. Finally, it is also obvious that we attribute errors to one another whether to the ancient fisherman calling whales "fish" or to people calling critical race theory evil. Thus, the error-attribution feature is also preserved.

4.2.2. The External-Constraint Challenge.

I now turn to the challenge that there must be some external constraint by which we can judge the standard of applications. I have already suggested that by the error-attribution feature the varying standard of applications can lead to varying levels of success in using that word. To answer the external-constraint challenge, I exploit this point. I do this with the help of a substantive anthropological assumption that if a standard of application is maximally success-inducing, then our standard of application will over time converge towards that standard.

In the following, I try to make explicit the way the external constraint of success, with the help of the anthropological assumption, guides the changing standard of applying a word. I do this with the help of a phenomenon identified by David Plunkett and Tim Sundell called

"metalinguistic negotiation" (Plunkett & Sundell, 2013).[22] According to Plunkett and Sundell, two speakers are engaging in metalinguistic negotiation if they (1) disagree about the standard of applying a word, and (2) express their disagreement implicitly, specifically by using the word whose standard of application is the subject of disagreement instead of mentioning it. For example, the disagreement between us and past speakers who refer to whales with "fish" is likely a case of metalinguistic negotiation. We and the speakers past (1) disagree on the standard of applying "fish" by the variation feature of our language, and (2) this disagreement is expressed implicitly since by the opacity feature, we do not realize that our standards of applying the word in fact differ from the standards employed by the speakers past; in other words, we disagree with them by attributing error to them using the word "fish" instead of mentioning it to talk about its standard of application.

Further, Plunkett and Sundell stress that metalinguistic negotiation can often be substantive since there is often a fact of the matter about which standard of application induces more success (Sundell, 2012; Plunkett & Sundell, 2013; Plunkett, 2015). For example, one central function of the word "fish" is to describe the world. Thus, a standard of applying it that carves the world at its joint (thus whale-exclusive) will fulfill its function better and induce more success for the language-user than one that fails to carve the world at its joint (thus whale-inclusive). In contrast, since the function of the word "morally good" is to describe the moral properties of the world, then the standard of using the word according to which the word tracks

[22] They credit this proposal to Barker (2002).

an objective moral property, however naturalistically complex it is, will be better than a standard under which the word tracks a property that is a metaphysically natural property[23].

If all this is granted, then, by the anthropological assumption that our standard of application evolves to maximize our success, the stabilized standard of application, that ultimately fixes the semantic values of our language, will also be the one that brings the most success. And in cases of metalinguistic negotiation, the speaker whose standard of application leads to more success will likely come out victorious. Thus, when we attribute error to communities of past speakers who deviate from the stabilized standard (thus uttering strictly false sentences), the attribution is usually vindicated: not only are they uttering falsehood, but they are also failing to maximize success.[24]

Thus, the external-constraint challenge is met. Before moving to the next challenge, I should also note that although I have worked with the metalinguistic negotiation to illustrate my point, there is no reason to commit to metalinguistic negotiation as the only way standards of application changes. It is simply the most natural way given the opacity and error-attribution features of our language (Plunkett & Sundell, 2021). If a speaker realizes that they are employing different standards of application, they may "negotiate" the standard going forward explicitly in the way that philosophers of race negotiate the definition of racial terms (Haslanger, 2012); and

[23] There is no commitment to moral realism. If, for example, some sort of expressivism is true, and moral talk only help coordinate collective action, the use of moral terms that is more conducive to collective action will be better.
[24] One objection may be that the anthropological assumption is too strong, and if it is relaxed, then there may be cases where speakers say things that are strictly false, but there is no basis for the community to attribute error to the speaker since the standard of application of the community turns out to induce less success. However, if another scientist uses their word in a way that better describes the world, why would we attribute error? Imagine, for example, if the scientists had reacted to Einstein's discovery of special relativity by saying that mass is defined by p = mv, so the way Einstein uses "mass" is strictly wrong. Thus, in such cases, the attribution of error is simply mistaken despite the (strictly speaking) falsity of their utterances.

speakers may also copy more successful individuals without engaging anything as dramatic as disagreement and attributing error.

4.2.3 The Stabilization Challenge

To answer the final challenge, I will start by saying what it means for a standard of application to be stabilized. I propose that if the standard of using W according to the relevant standard-determining fact at t2 is viable as the standard for using W according to the relevant standard-determining facts that obtain at an earlier time t1, then the standard of applying W is stable from t1 to t2 (although it may have changed in between). Thus, the standard of applying W is stabilized at t if there is no later time t' such that the standard of using W at t' is not the standard of applying W at t according to the relevant standard-determining fact. For example, suppose that at no later time will the relevant standard-determining fact for "fish" change in a way such that our current use of the word "fish" violates that later standard. Then, the standard of using "fish" is stabilized. This already handles the stabilization challenge. Plausibly, "fish" is a term that has infinite applications: it can be correctly applied to all the fish past, present and future as well as all possible fish. However, even if the relevant standard-determining fact right now has it that "fish" can only be applied to finite things, since there is no later time where the standard of applying "fish" changes so drastically that application of the word "fish" according to current practice violates that future standard, our finite use is stable.

For cases where the standard-determining facts change (1) so drastically that uses according to the current standards are incorrect and (2) so constantly that (1) always obtains, I am committed to say that the standard does not stabilize. But there is no problem. This simply

describes vague or indeterminate terms. For example, “sandwich” is a word that likely satisfies (1) and (2), so there are many bread-related foods for which it is vague whether they count as sandwiches. As such, the stabilization challenge is met.

4.3 Emergent Eligibility

In chapter 2, I motivated the thought that popular metasemantic theories need to be supplemented by a theory of eligibility where the degree of eligibility of different properties plays a role in determining the semantic value of a word. In chapter 3, I considered various analyses of the notion “eligibility”, and argued that none of them is viable. So far in chapter 4, I have developed an account of reference where the referent of a term is determined by its stabilized standard of application. In this section, I will show how it all comes together: a theory of eligibility emerges from the SSA. Specifically, I will lay out this emergent theory of eligibility before turning to the problem of indeterminacy and the problem of disagreement again and show that the problems that have motivated the theory of eligibility in the first place can be solved with SSA.

4.3.1 Building Eligibility

In this section, I will show how something like eligibility emerges from SSA by showing that SSA can recover the tie-breaking and overriding function of a theory of eligibility. Then, I formulate a theory of eligibility based on SSA.

To elaborate, suppose that a generic metasemantic theory T gives the right standard-determining fact, where a predicate P refers to a property F if P relates to F by M under T.

According to the kinds of theory of eligibility I have considered in chapter 3, facts about eligibility help determine semantic facts in two ways: first, in cases where a predicate P bears the relation M to two distinct properties, facts about the eligibility of those properties break the tie between the two properties such that the semantic value of the predicate is the set of things that instantiate the more eligible property. Second, in cases where a predicate bears the relation M to a property F', but there is a property F that approximately bears the relation M to P such that F is sufficiently more eligible than F', then, the semantic value of P may be the set of things that instantiate F instead of those instantiating F'. As I discussed in 3.1, these two ways for facts about eligibility to constrain semantic facts are called "tie-breaking" and "overriding" respectively. In the following, I will show how SSA recover them.

To start with, recall that, under SSA, generic metasemantic theories are partial theories: they supply *standard*-determining facts that will be only *meaning*-determining when supplemented by SSA. Suppose again that T gives the right standard-determining facts. Thus, under SSA, we treat facts about M at any given time t as providing the standard of application at t. Then, here is how SSA breaks the tie between potential semantic values that T underdetermines. Suppose that P bears the relation M to distinct properties F and F* at a time t1. Then, according to SSA, the tie between {x|Fx} and {x|F*x} as potential semantic value is broken in favor of the property that P bears M to, when facts about what P bears M to have stabilized. Therefore, suppose that at t2, the standard of applying P stabilizes, and P bears M to F instead of F* at t2, then according to SSA, {x|Fx} is and was the correct semantic value of P. To see how SSA overrides T, suppose that at t1, the predicate bears M to F*, but at t2, when the standard of applying P has stabilized, P bears M to F, then according to SSA, the semantic value of P is {x|Fx} instead of {x|F*x} even at t1. As such, SSA overrides the metasemantic facts

proposed in T. Of course, T is generalized: facts about the relation M are just stand-ins for facts about any appropriate causal relation, any inferential roles, or any rationality-maximizing interpretation we are independently motivated to accept.

There is thus a clear symmetry between SSA and a theory of eligibility: both supplement a metasemantic theory by breaking the ties when the metasmeantic theory yields indeterminate results by itself and sometimes overriding the result produced by metasemantic theory. We may even restate SSA in the following theory of emergent eligibility:

- (EE) A property F is perfectly eligible with respect to a language L if there is a predicate P in L such that according to the stabilized standard of applying P, P can be predicated of anything that instantiates F and nothing else.

I claim that (EE) is the correct theory of eligibility.

If (EE) is to be a restatement of SSA, there are two further features to note: first, it is usually assumed that facts about eligibility are balanced with the meaning-determining facts proposed by the metasmeantic theory it supplements. In other words, the semantic value of a predicate is usually allowed to favor a less eligible property even if there is a more eligible property in its vicinity. However, if (EE) is to be a reformulation of SSA, this is not allowed: the property that a predicate designates just is the one that the predicate is correctly applied to when the standard of application is stabilized or, in other words, the perfectly eligible property. This may seem too strong for a theory of eligibility, but this is not unheard of. Consider the theory-internal analysis of eligibility Williams (2015) suggests (see 3.3.1 for the details). According to that theory, the perfectly eligible properties are the ones that are designated by familiar predicates in the metalanguage. Therefore, when the metalanguage and object language are identical, considerations about eligibility always trump other considerations.

Second, under SSA, if it is indeterminate whether a predicate P designates the property F or F* even when the standard of applying P has stabilized, then we would have a case of genuine semantic indeterminacy, perhaps a vague predicate or an ambiguous one. Therefore, it is plausible that whatever indeterminacy SSA leaves behind are acceptable cases of indeterminacy. It is the type of indeterminacy where there is no fact of the matter about whether a particular person falls under the extension of "bald", not the type where it is indeterminate whether "Wittgenstein" refers to Wittgenstein or Elon Musk. If the latter type of indeterminacy remains, then we may need another theory of eligibility on top of SSA. Then, (EE) will not be a genuine alternative to the existing theories of eligibility. Therefore, in the following two sections, I will take up the problem of indeterminacy and the problem of disagreement again, and show that SSA, and therefore (EE), can solve these two problems.

4.3.2 The Problem of Indeterminacy, Revisited

To see how SSA deals with the problem of indeterminacy, I will consider Williams (2007)'s math-world problem which is arguably a stronger version of the problem. (To revisit the details, see 3.2.1.)

Recall that according to the metasemantic theory known as "global descriptivism", a semantic theory T for L is true as long as under T the number of true sentences that features in the total theory held by the linguistic community of L is maximized. Now, consider Williams' Math-World theory. Under that semantic theory, every proper name in English is assigned a natural number as its referent, and the extension of a predicate like "table" will be the set of numbers that are assigned as the referent of the names of which the English speakers predicate "table". Therefore, the problem goes, the Math-World theory will, by construction, be such that

every sentence that is held true by the linguistic community is true under it. This makes the Math-World theory *possibly* the true semantic theory for English.

Here is, then, how SSA deals with the problem. Suppose that, at the moment when the current version of Math-World theory is formulated (call it "t1"), the standard of applying "table" is stabilized. Therefore, by SSA, the semantic value of "table" will be a set of natural numbers. Call this set A. However, suppose at a later time t2, a new table is brought into existence. Naturally, English speakers will talk about this new table using the predicate "table". However, this puts a dent in the Math-World theory. To elaborate, suppose this new table is dubbed "Steve". Then, whatever number the Math-World theory assigns as the referent of "Steve", that number will not be in A. As such, the sentence "Steve is a table" will be false under the Math-World theory. One reply may be that there is a revised Math-World theory that we may formulate at t1 where "table" is assigned the set A^ as the semantic value, where

- $A^{\wedge} =_{df.} A \cup \{|\text{“Steve”}|\}$

where "||" is the interpretation function of the revised Math-World theory that assigns to "Steve" a natural number. But, as new tables are continually being made, any revised Math-World theory will be ruled out.

There is, however, another way of revising the Math-World theory that poses a more serious problem for how SSA deals with the problem of indeterminacy. Specifically, let B designate the set of all tables that are already in existence at t1. Then, the Math-World theory may assign the set

$$A^{*} =_{df.} A \cup (\{x|\text{table}(x)\}/B)$$

or the union set of A and the set of all tables except for those already in existence. As such, any table produced in the future will fall under this extension, but no current table does. This is just

as crazy as the original Math-World theory, and it seems to bypass SSA. To diagnose the problem, I think that global descriptivism in fact cannot even determine a standard of application. In other words, given a set of sentences that the speakers hold true at a particular time, GD cannot even determine what is the correct way of applying a particular name or predicate at that time. To use the table example again, the current sentences that we hold true do not determine whether A^* or {x|table(x)} is the semantic value of "table".

With this diagnosis in hand, SSA has another response: at any given time, the standard of application is determined by not only the truth-maximizing interpretation, but also facts about how successful the speakers' linguistic activities are. To elaborate, recall that in 4.2, I have argued that there is a fact of the matter about how successful a standard of application is. As such, even though the set of beliefs we hold true, by themselves, underdetermine the standard of application under GD at a given time, since there is still a fact of the matter about how much success we are able to garner with the way we use language, and only one of those underdetermined standard of applications accurately reflect the amount of success we do get, that is the standard of application operating at that time. Having fixed the standard of application at a time, the semantic theory follows.

4.3.3 The Problem of Disagreement, Revisited

Now, consider the problem of disagreement according to which our best metasemantic theories seem to misrepresent cases of substantive disagreements between linguistic communities as merely talking past each other. The examples I have considered include the case of moral twin earth and the case of a community of brains in vats encountering the outside world for the first time. The way SSA handles these cases depends on the theory of word-type individuation we

adopt. Specifically, a theory of word type individuation tells us when token expressions are of the same type. Therefore, I will first consider how SSA deals with these cases if according to the theory of word type, the two communities are using the same word with the same meaning; then, I will consider what happens if, according to the theory of word type, the two communities are not using the same word with the same meaning.

To start with the first case (where disagreeing parties use the same word), the disagreement problem is immediately solved. Specifically, if the theory entails that our moral term and the moral term of the twin earthers are in fact of the same type, or that BIV's word "brains" and "vats" are of the same type as those of their saviors, then the immediate consequence is that they utter sentences with incompatible content. The disagreement is immediately real and substantive.[25] However, one problem remains: in these cases, we not only have the expectation that the disagreement is substantive, but we also have specific expectations about the topic of disagreement. For example, in the case of moral twin earth, the intuition is that we, and the twin earthers, are disagreeing about moral goodness (as opposed to the property of maximizing utility assuming that it is not identical with moral goodness), and in the case of BIVs, the BIVs and their saviors are disagreeing about the external world (not the strings of electronic signals that correspond to BIV's concept of the external world).

Does SSA get the topic of disagreement right as well as the fact of disagreement? We can't be sure, but we have good reason to think that it does, especially if certain anthropological assumptions are borne out. To elaborate, assuming that there are general anthropological reason that the more success-inducing standard of applying a word is more likely to be spread, and a word that refers to moral goodness is more success-inducing in guiding our action and a word

[25] Laura Schroeter (2014), for example, takes a route similar to this.

that refers to the external world is more success-inducing in our theorizing about the world, then it is likely that we would eventually settle on standard of application where we have a word referring to moral goodness as opposed to utility maximization and a word referring to the external world as opposed to a peculiar type of electronic input.[26] Thus, as the twin earthers and the BIV saviors are using the same type of word with the same meaning as we and BIVs do, we and the moral twin earthers will eventually agree that "morally good" designates moral goodness, and the BIV and their saviors will eventually agree that "external world" designates the external world and not a type of electronic input. I hold that this is the most likely case. As such, under SSA, our expectation about the topic of disagreement is very likely to be right, though it is not guaranteed (as the anthropological assumption is not guaranteed to be correct).

The case just considered, where the theory of word type yields that the two disagreeing communities are using words of the same type, is relatively easy. However, in the setup of the case of the moral twin earth and the BIV encounters, it is wildly unlikely that our preferred theory of word type would yield this result. After all, if we really meet a group of aliens, no matter how similar our and their linguistic practices are in terms of the phonemes uttered and the shapes scribbled, we would be very skeptical of the claim that we are speaking the same language. Indeed, it has been argued that it is mistaken to expect our word "morally good" is the same word with the same meaning as twin earther's word "*morally good*" (Dowell, 2016).

I will now consider the case if our word individuation theory does not work out in our favor. In such cases, there is no guarantee that the disagreement is substantive, but just as in the

[26] Again, there is no reason to commit to moral realism or even realism about structure. It may be that moral vocabulary only contributes to cooperation, and the "projectability" or the explanatory force of a term is only a function of its entrenchment in our ideology or our cognitive architecture. Then, the prediction will be that our use of moral words will settle on a standard that maximizes cooperation and our use of explanatory words will settle on a standard that fits most comfortably with our ideology and cognitive architecture.

case before, it is highly likely that it is. Here is why. Although we cannot say for sure that the stabilized standard of applying the word "moral goodness" for moral twin earthers is the same as ours, we can be pretty confident that it is. For one, we can be confident that our term refers to moral goodness. Also, we can be pretty confident that a species like the moral twin earthers would not achieve the sophistication that they do without some kind of cultural or cognitive mechanism that constrains their use of language to be more and more success-inducing. Therefore, we can be confident that their evaluative term "*moral goodness*" will find its way to moral goodness as well. Again, this is all tentative, but I think that optimism is justified.

Thus, SSA does deal with the two problems for eligibility pretty well. As such, I claim that SSA, and (EE) which is just a reformulation of SSA, is a genuine alternative to existing theories of eligibility.

4.4 Outstanding Objections

In this section, I will consider and respond to some outstanding objections to SSA. Specifically, I will consider the objection that the causal origin theory poses a counterexample and the objection that SSA cannot handle cases where standards of applications just do not stabilize.

4.4.1 Causal Origin Theory

To start with, SSA is intended to apply to any languages that exhibit variation (the standard of application changes over time), opaqueness (such changes are not transparent to language users), and error-attribution (speakers attribute error to other speakers who apply words according to a different standard). I have also claimed that SSA can be used to supplement any familiar metasemantic theory. One objection, then, is that SSA cannot be used to supplement the

causal origin theory of reference since if the standard of applying a word is determined by the causal origin of that word, then our language loses its variation feature. Taking the word "fish" as an example, according to this objection, if the original baptismal event involved a person dubbing a fish with the kind term "fish", then it seems like the standard of applying "fish" has stayed the same throughout the ages: both the ancient fisherman and we have the same standard of applying "fish" since such a standard, in both cases, is grounded in the same baptismal event.

To answer this question, let's start by considering why the causal origin theory needs to be supplemented by SSA: the "qua" problem (Devitt & Sterelney, 1987). Here is the problem. Suppose that the baptismal event for "fish" is one where a particular fish named "Nemo" is dubbed. Then, because Nemo exhibits both the property of being a fish and the property of being a fish-shaped aquatic creature, it is indeterminate whether we should construe the cause of the baptismal event as a fish or a fish-shaped aquatic creature. The implication, then, is that it is indeterminate, given the baptismal event, whether "fish" is correctly applied to fish or the kind of things that are either fish or whales or something else.

A solution to the "qua" problem, then, rests on the observation that Nemo the fish causes our use of "Nemo" the word in virtue of having some relevant property, and presumably it is in virtue of being a fish, rather than a fish-shaped sea creature, that Nemo has caused our use of the word "Nemo" (Miller, 1992; Deustch, 2021). There are two ways to develop the observation: according to one approach, the causally relevant property of Nemo in this case is the property of being a fish because had Nemo not been a fish, we would not have acquired the capacity to track fish as a result of dubbing Nemo (Miller, 1992). According to another, the property of being a fish is relevant because had Nemo not been a fish, the baptismal event would not have happened (Deustch, 2021).

Both approaches, however, need SSA. Here is why. Consider the first approach: we did not always have the capacity to track fish, instead of fish-shaped creatures. Ancient fishermen do not. Instead, they had the capacity to track fish-shaped creatures, and they apparently acquired this capacity because of their use of the word "fish". The question, then, is "who's tracking capacity counts? Ours or the ancient fishermen's?" This is where SSA provides a natural answer: the capacity the speakers acquire at a particular time t by using the word "fish" tells us the standard of applying that word at t, but the extension of "fish" is determined by the stabilized standard and thus the capacity of those speakers whose use of the word has stabilized. This also secures the variation feature, as our tracking capacity clearly varies with time and across subcommunities. Consider, also, the second approach. Here, the baptismal event is one where someone dubs Nemo with a string of phonemes S. However, there is a "qua" problem for S as well: S instantiates both the property of being a token expression that tracks fish and the property of being a token expression that tracks fish-like creatures. Thus, it is unclear which is the relevant property of S that is essential to the baptismal event. It is clear that this repeats the problem for the first approach, and SSA comes in in the same way.[27]

4.4.2 Reference without Stabilization

I have claimed that under SSA, terms whose standards of application do not stabilize do not have determinate semantic values. There is, however, a complication I failed to discuss. If human civilization goes out of existence before the vast majority of our terms settle on a stable standard of application, then the vast majority of our terms do not have determinate semantic value. This level of indeterminacy is unacceptable. My reply to the problem, then, is that in such cases, we

[27] Note that in both approaches, it is not that the causal relations in the past are changed by events in the future, but that the future events change what the relevant causal explanans are (from the capacity to track fish-like creatures to tracking fish for example), so the explanandum changes as well.

may appeal to the truth of such counterfactuals as "had human civilization not gone out of existence, then the standard of applying their term T would eventually stabilize on S." To give a standard modal gloss, in such doomsday cases, the semantic value of a term T is given by the stabilized standard of applying T in the closest possible world where we survive long enough for the standard of applying T to stabilize. Thus, the problem is solved: even if human civilization goes out of existence before our standard of application stabilizes, there would be no mass indeterminacy.[28]

There is however a further worry: it seems that in some cases the relevant counterfactual may not fix the semantic value. Specifically, according to this further worry, counterfactuals about what would happen if the standard of application stabilizes may change with time. For example, at one point it may seem like the standard of applying the word "planet" would stabilize on an extension including Pluto, but now that is not the case. Thus, suppose that in the actual world, the standard of application does eventually stabilize. Then, the worry goes, there may be a time t in our history such that at t, in the closest possible world where the standard of application stabilizes, the standard of applying a word W may be different from the standard of applying W that actually stabilizes. However, stated as such, the worry is not so worrying: if the standard of application in the actual world does eventually stabilize, then the actual world would of course be the closest possible world where the standard of application stabilizes. If the standard of application in the actual world does not stabilize, then there is a unique (set of)

[28] A rejoinder may be that according to certain views about time, such as presentism, there are no future facts. To respond, if that is the case, then whatever makes statements about the future true will also make true statements about stabilized standards of applications. If, however, it is insisted that statements about the future do not have determinate truth value, then, sure, SSA falls to mass indeterminacy. But that is just more reason to think that statements about the future have truth values now.

closest possible world where the standard of application does stabilize, and the semantic value of relevant terms is fixed by what happens in those possible world.

In this chapter, I have discussed Azzouni's solution to the rule-following problem, and developed the SSA based on that solution. As I have argued in 4.3, the upshot of the SSA is that it presents a genuine alternative to existing theories of eligibility: it solves the problems they promise to solve, and it is free of the problems that afflicted them.

Chapter 5. Conclusion

In this thesis, I have explored the metasemantic theory of reference magnetism. I started by considering two problems, a theory of reference magnetism aims to solve: the disagreement problem and the indeterminacy problem. Then, I have surveyed the different theories of reference magnetism on offer and argued that none of them is satisfactory. Finally, I presented an alternative theory based on Azzouni's solution to the rule-following problem that explains why there appears to be the phenomenon of reference magnetism.

In conclusion, if I am right, what lies behind reference magnetism is nothing but our continual negotiation about the way we talk.

References:

Azzouni, Jody. *The rule-following paradox and its implications for metaphysics*. Vol. 382. Heidelberg: Springer, 2017.

Artiga, Marc & Sebastián, Miguel Ángel (2020). Informational Theories of Content and Mental Representation. *Review of Philosophy and Psychology* 11 (3):613-627.

Barker, Chris (2002). The dynamics of vagueness. *Linguistics and Philosophy* 25 (1):1-36.

Bays, Timothy. "The problem with Charlie: some remarks on Putnam, Lewis, and Williams." *The Philosophical Review* 116.3 (2007): 401-425.

Block, Ned. "Advertisement for a Semantics for Psychology." *Midwest studies in philosophy* 10 (1986): 615-678.

Boghossian, P., 1989a. "The Rule-Following Considerations," *Mind*, 98, 507–549.

____. "The transparency of mental content." *Philosophical perspectives* 8 (1994): 33-50.

Boyd, Richard N. "How to be a moral realist." *Contemporary materialism*. Routledge, 2002. 318-367.

Brandom, Robert. *Making it explicit: Reasoning, representing, and discursive commitment*. Harvard university press, 1994.

Bridges, J. (2014) 'Rule-Following Skepticism, Properly So Called', in J. Conant and A. Kern (eds) Varieties of Skepticism: Essays after Kant, Wittgenstein, and Cavell. Berlin: De Gruyter.

Buchanan, Ray, and Sinan Dogramaci. "Representation and Rationality." *Philosophy and Phenomenological Research*. Forthcoming.

Carruthers, Peter. *The opacity of mind: An integrative theory of self-knowledge*. OUP Oxford,

2011.

Chalmers, David (2012). *Constructing the World*. extended edition. Oxford University Press.

____. (forthcoming). Inferentialism, Australian style. *Proceedings and Addresses of the American Philosophical Association.*

____."Interpretivism and Inferentialism." *Analysis* (2021): 524-535.

Davidson, D. (1979). 'The inscrutability of reference'. The Southwestern Journal of Philosophy, pages 7-19. Reprinted in Davidson, Inquiries into Truth and Interpretation (Oxford University Press, Oxford: 1980) pp.227-242.

____. (1967). "Truth and meaning." *Synthese* 17 (1):304-323.

____. "Radical interpretation." *Dialectica* (1973): 313-328.

____. "Belief and the Basis of Meaning." *Synthese* 27.3 (1974): 309-323.

____. *Truth and Predication.* Harvard University Press. (1990).

Deutsch, Max. "Is There a "Qua Problem" for a Purely Causal Account of Reference Grounding?." *Erkenntnis* (2021): 1-18.

Devitt, Michael and Kim Sterelny, 1987, *Language and Reality: An Introduction to the Philosophy of Language*, Cambridge, MA: MIT Press.

Dickie, Imogen. "How Wrong Could We Be." *Analysis*. Forthcoming.

Dorr, Cian & Hawthorne, John (2013). Naturalness. In Karen Bennett & Dean Zimmerman (eds.), *Oxford Studies in Metaphysics: Volume 8*. Oxford

Dowell, J. L. (2016). The Metaethical Insignificance of Moral Twin Earth. In Russ Shafer-Landau (ed.), *Oxford Studies in Metaethics volume 11*. Oxford: Oxford University Press. pp. 1-27.

Dretske, Fred I. "Knowledge and the Flow of Information." (1981).

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

Dunaway, Billy, and Tristram McPherson. "Reference magnetism as a solution to the moral twin earth problem." *Ergo, an Open Access Journal of Philosophy* 3 (2016)

Dunaway, Billy. *Reality and Morality*. Oxford University Press, 2020.

Eddon, Maya, and Christopher JG Meacham. "No work for a theory of universals." *A companion to David Lewis* 57 (2015): 116.

Foster, J. A. (1976). 'Meaning and truth theory'. In G. Evans and J. McDowell, editors, Truth and Meaning: Essays in semantics, pages 1-32. Clarendon Press, Oxford.

Gardiner, Mark Q. (1995). Just more theory? *Australasian Journal of Philosophy* 73 (3):421 – 428.

Ginsborg, Hannah. "Primitive normativity and skepticism about rules." *The Journal of Philosophy* 108.5 (2011): 227-254.

Hale, Bob, and Crispin Wright. "Putnam's model-theoretic argument against metaphysical realism." *A Companion to the Philosophy of Language, 2nd edition, Oxford: Wiley-Blackwell* (2017): 703-730.

Harman, G. 1999: *Reasoning, Meaning and Mind*. Oxford: Oxford University Press.

Haslanger, Sally (2012). *Resisting Reality: Social Construction and Social Critique*. Oxford University Press.

Hattiangadi, Anandi (2020). Substantive Radical Interpretation and the Problem of Underdetermination. *Analysis* 80 (4):822-833.

Hawthorne, John & Lepore, Ernest (2011). On Words. *Journal of Philosophy* 108 (9):447-485.

Henkin, L. (1949). The completeness of the first-order functional calculus. *The Journal of*

Symbolic Logic, 14(3), 159-166.

Horgan, Terence & Timmons, Mark (1991). New Wave Moral Realism Meets Moral Twin Earth. *Journal of Philosophical Research* 16:447-465.

Horwich, Paul. "A use theory of meaning." *Philosophy and Phenomenological Research* 68.2 (2004): 351-372.

Kaplan, David. "Words." *Proceedings of the Aristotelian society, supplementary volumes* 64 (1990): 93-119.

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

____. *Wittgenstein on rules and private language: An elementary exposition*. Harvard University Press, 1982.

Lewis, David. (1970). General semantics. *Synthese* 22 (1-2):18--67.

____. "Putnam's paradox." *Australasian journal of philosophy* 62.3 (1984): 221-236.

____. (1974). Radical interpretation. Synthese 27 (July-August):331-344

____. "New work for a theory of universals." *Australasian journal of philosophy* 61.4 (1983): 343-377.

____. (1975). Languages and language. In Keith Gunderson (ed.), *Minnesota Studies in the Philosophy of Science*. University of Minnesota Press. pp. 3-35.

____. (1994). Reduction of mind. In Samuel Guttenplan (ed.), *Companion to the Philosophy of Mind*. Blackwell. pp. 412-431.

____. (1992). Meaning without use: Reply to Hawthorne. *Australasian Journal of Philosophy* 70 (1):106 – 110.

Merino-Rajme, C. (2015) 'Why Lewis' Appeal to Natural Properties Fails to Kripke's Rule-Following Paradox', Philosophical Studies, 172/1: 163–75.

Miller, R. B. (1992). A purely causal solution to one of the *Qua* problems. *Australasian Journal of Philosophy, 70*(4), 425–434.

Millikan, R., 1984, *Language, Thought and Other Biological Categories*, Cambridge, MA: MIT Press.

Mokriski, David (2020). The Methodological Implications of Reference Magnetism on Moral Twin Earth. *Metaphilosophy* 51 (5):702-726.

Montague, Richard (1974). *Formal Philosophy: Selected Papers of Richard Montague*. New Haven: Yale University Press.

Nagel, Thomas (1986). *The View From Nowhere*. Oxford University Press.

Neander, Karen. *A mark of the mental: In defense of informational teleosemantics*. MIT Press, 2017.

Pautz, Adam (2021). Consciousness meets Lewisian interpretation theory: A multistage account of intentionality. In Uriah Kriegel (ed.), *Oxford Studies in Philosophy of Mind*.

Plunkett, David & Sundell, Timothy (2013). Disagreement and the Semantics of Normative and Evaluative Terms. *Philosophers' Imprint* 13 (23):1-37.

____. (2021). Metalinguistic Negotiation and Speaker Error. *Inquiry: An Interdisciplinary Journal of Philosophy* 64 (1-2):142-167.

Plunkett, David (2015). Which Concepts Should We Use?: Metalinguistic Negotiations and The Methodology of Philosophy. *Inquiry: An Interdisciplinary Journal of Philosophy* 58 (7-8):828-874.

Putnam, Hilary. Reason, Truth and History. Cambridge, 1981. University Publishing Online.

Web.

____. (1977). Realism and Reason. *Proceedings and Addresses of the American Philosophical Association* 50 (6):483-498.

____. (1980). Models and reality. *Journal of Symbolic Logic* 45 (3):464-482.

Schwarz, Wolfgang. "Against magnetism." *Australasian Journal of Philosophy* 92.1 (2014): 17-36.

Sider, Theodore. *Writing the Book of the World*. OUP Oxford, 2013.

____. "Ontological realism." *Metametaphysics*. OUP, 2009. 384-423.

Simpson, Matthew (2018). Solving the problem of creeping minimalism. *Canadian Journal of Philosophy* 48 (3-4):510-531.

Schaffer, Jonathan. "Two conceptions of sparse properties." *Pacific Philosophical Quarterly* 85.1 (2004): 92-102.

Schroeter, Laura, and François Schroeter. "Rationalizing self-interpretation." *The Palgrave handbook of philosophical methods*. Palgrave Macmillan, London, 2015. 419-447.

Schroeter, Laura (2014). Normative Concepts: A Connectedness Model. *Philosophers' Imprint* 14.

____. "Illusion of transparency." *Australasian Journal of Philosophy* 85.4 (2007): 597-618.

Stojnić, Una. "Just words: Intentions, tolerance and lexical selection." *Philosophy and Phenomenological Research* (2021).

Sundell, Timothy (2012). Disagreement, Error, and an Alternative to Reference Magnetism. *Australasian Journal of Philosophy* 90 (4):743 - 759.

Taylor, Barry (1991). 'Just more theory': A manoeuvre in Putnam's model-theoretic argument for antirealism. *Australasian Journal of Philosophy* 69 (2):152 – 166.

Verheggen, C. (2011) ‘Semantic Normativity and Naturalism’, Logique et Analyse, 54/216: 553–67.

Weatherson, Brian. "The role of naturalness in Lewis's theory of meaning." *Journal for the History of Analytical Philosophy* 1.10 (2013).

Wedgwood, Ralph (2001). Conceptual role semantics for moral terms. *Philosophical Review* 110 (1):1-30.

Wildman, Nathan. "On shaky ground? Exploring the contingent fundamentality thesis." *Reality and Its Structure: Essays on Fundamentality, eds. R. Bliss and G. Priest* (2018): 275-90.

Williams, J. Robert G. (2018). Normative Reference Magnets. *Philosophical Review* 127 (1):41-71.

____. "The price of inscrutability." *Nous* 42.4 (2008a): 600-641.

____. "Permutations and Foster problems: two puzzles or one?." *Ratio* 21.1 (2008b): 91-105.

____. "Eligibility and inscrutability." *The Philosophical Review* 116.3 (2007): 361-399.

____. "Lewis on reference and eligibility." *A companion to David Lewis* (2015): 367-381.

____. *The metaphysics of representation*. Oxford University Press, USA, 2020.

____. "Reply to Critics." *Analysis* 81.3 (2022): 536-548.