TRUTH AND BELIEF: CASE STUDIES IN CONCEPTUAL ENGINEERING

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

In Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Eric Gordon Epstein

May 2017

TRUTH AND BELIEF: CASE STUDIES IN CONCEPTUAL ENGINEERING

Eric Gordon Epstein, Ph. D.

Cornell University 2017

Although the concepts of truth and belief are fundamental in philosophy, in recent years they have come under attack from various quarters. I argue that philosophers have been too quick to find these concepts problematic and in need of being replaced. For example, the Liar Paradox is sometimes taken to show that the concept of truth is inconsistent, and thus unsuitable for rigorous inquiry. But I develop a solution that gives a consistent account of this concept, allowing us to retain it in spite of the paradox. I argue that when the word 'true' occurs in such a sentence, it undergoes a one-off aberration in its reference, failing to refer to truth. Thus, Liar sentences and their kin fail to say what they pre-theoretically appear to say. However, there is no need to conclude that these sentences are meaningless; rather, I illustrate how these sentences come very close to saying what they appear to say, in spite of the aberrations they witness.

As with the concept of truth, *intentional* concepts like those of belief, meaning, and reference have been subject to skepticism and attempts at excision. I show that the content of the claim that the use of intentional concepts can be eliminated from scientific explanations depends on broader issues about how one conceives of explanations. Then I argue that intentional concepts, in particular the concept of belief, play an ineliminable role in the explanation of behavior: when we learn what someone believes, we get some information about how she would react to a variety of possible scenarios. This information that is useful in everyday life, and would be important in a science whose aim was to improve on our folk-psychological explanations. But so far, explanations

that avoid talk of beliefs have failed to replicate this distinctive kind of informativeness. Thus, we have reason to think that belief-attributions are indispensable when it comes to explaining people's behaviors, and so we have reason to retain the notion of belief.

BIOGRAPHICAL SKETCH

Eric Epstein was born on April 23, 1987 in Washington, D.C., to a middle-class, secular family of mixed Jewish and Puritan background. He was born with Type III von Willebrand Disease (hemophilia). Eric grew up in McLean, Virginia, where he developed a fascination with insects, poetry, the United States Civil War, and Scottish history and culture. He took poetry lessons and learned to play the highland bagpipes, studying with Sandy Jones for two summers at the North American Academy of Piping in Valley Crucis, North Carolina.

In middle school, long walks with his father led Eric to become fascinated with philosophy. He took philosophy lessons from Nathaniel Goldberg, then a graduate student at Georgetown University. His development was also profoundly influenced by his high school teachers Ken Kraner, Lewis Sinclair, Lois Cohen, and Faye Casio. Eric entered Cornell University in Fall 2005 with the intention of studying entomology, but was inspired to double major in mathematics and philosophy after taking introductory logic with Harold Hodes. Prof. Hodes went on to become Eric's undergraduate advisor and, eventually, dissertation committee chair. Eric spent his third year of college studying at Pembroke College, Oxford University, where he attended poetry workshops with the Oxford University Poetry Society and was tutored in metaphysics, epistemology, logic, and philosophy of language by Gabriel Uzquiano.

After receiving his B.A. in May 2009, Eric remained at Cornell University, entering the Ph.D. program in philosophy in the fall with interests in logic and the philosophy of language. In two seminars and a tutorial led by Matti Eklund, he wrestled with ideas about truth and meaning that became the subject of his dissertation, which developed under the guidance of Prof. Eklund, Prof. Hodes, and Prof. Will Starr. As a graduate student, Eric continued writing poems and playing bagpipes. His Jewish identity underwent a renaissance, and he took Hebrew classes, served as a teaching assistant for "History of the Israeli-Palestinian Conflict" under Ross Brann, and converted to Masorti Judaism with Temple Beth El of Ithaca. He met and married Yoon Choi, then a Ph.D. student in Communications. They live together with their two cats.

This work is dedicated to those teachers who have influenced me the most profoundly: Ken Kraner, Faye Cascio, Lois Cohen, Lewis Sinclair, Nathaniel Goldberg, Jens Schellhammer, Gabriel Uzquiano, Will Starr, Matti Eklund, Harold Hodes, and my parents, Kimberly N. Epstein and Jay S. Epstein.

ACKNOWLEDGMENTS

I am enormously grateful to Cornell University and the Sage School of Philosophy for generously funding this work over many semesters. Derk Pereboom, Tad Brennan, and Ted Sider deserve special mention for their helpful assistance and resourcefulness in the securing of funding. I am also grateful to the Department of Near Eastern Studies, the Department of German Studies, and the John S. Knight Institute for Writing in the Disciplines for appointments that enabled me to continue this work.

I offer my sincerest gratitude to Harold Hodes, Matti Eklund, and Will Starr for their extraordinary patience and meticulous comments on many drafts. I also owe great debts to Robert Rupert, Hartry Field, Richard Boyd, Vann McGee, Derk Pereboom, Karen Bennett, Julia Markovits, Agustin Rayo, Alex Byrne, David Kovacs, Lavinia Picollo, Yuna Won, Eric Rowe, Lucia Munguia, David Fielding, Philippe Lemoine, Katie Mathie-Smith, Adam Bendorf, Stephen Mahaffey, Ian Hensley, Ian McKay, Brandon Conley, David Gray Grant, Melissa Schumacher, and Bernhard Salow. I would also like to thank the Cornell University Philosophy Workshop, the "MITing of the minds" workshop in Fall 2012, the Israeli Philosophical Association, and the Logic Colloquium 2016 at Leeds University for giving me opportunities to present parts of this work. In particular, I would like to thank J.R.G. Williams, Volker Halbach, and Theodora Achourioti for their comments during the Logic Colloquium.

I am also grateful to my friends David Kovacs, Eric Rowe, Andrea Viggiano, Philippe Lemoine, Lucia Munguia, Adam Bendorf, Ian Hensley, Brandon Conley, Yuna Won, David Fielding, Katie Mathie-Smith, David Gray Grant, Lyndal Grant, Mary Centrella, Sam Taylor, Shay and Natalie Sastiel, Dovid and Miri Birk, Eli and Chana Silberstein, Kelsey Utne, Karl Rozyn, David Golding, Philip Collender, and Jeff Craley, for years of stimulating conversations and warm company. My greatest gratitude of all is due to my grandparents Peter and Ann Neilsen, my brother Robert Epstein, my parents Kimberly and Jay Epstein, and my wife, Yoon Choi, for their sound advice and loving support even in the most trying of circumstances.

TABLE OF CONTENTS

CHAPTER 1

COMPONENT CONTEXTS, REFERENCE DETERMINATION, AND THE LIAR PARADOX

1. Introduction

Many diagnoses of the Liar paradox are rightly criticized for being *ad hoc*. That is, these solutions turn on claims about the semantic behavior of some of the expressions involved in the paradox, but fail adequately to support these claims with independent linguistic evidence. Rather, the claims are motivated primarily by their purported effectiveness in helping us avoid paradox. On the other hand, however, there are good reasons to think that if one is to avoid paradox, at least some measure of *ad hoc*-ness is inescapable. The very reason that the Liar paradox is so compelling is that if the laws of logic and the expressions involved in generating the paradox are as they seem to be, then it follows that some contradictions are true (and thereby, via the classical inference rule *ex falso quodlibet*, that everything is true[1]). Any way to circumvent the paradox is bound to violate at least some of our pre-theoretical impressions, either about the behavior of the expressions involved, the laws of logic, or both.

Given that some measure of *ad hoc*-ness is inevitable, the charge of *ad hoc*-ness finds the most traction when directed against solutions that have significant consequences concerning the semantic behavior of non-paradoxical sentences, or concerning the ways that the expressions implicated in the Liar paradox behave when they occur in such sentences. For while the paradox itself gives us good reasons to suspect that something unusual is going on in the sentences involved, the same does not obviously apply to the rest of language. It is methodologically unsound to let one's account of the non-paradoxical elements of language be driven by one's attempts to solve the Liar paradox, without regard to the linguistic data. And indeed, as we will

[1] One can react to this situation by rejecting *ex falso quodlibet* (EFQ), but that commits one to denying that the word 'not' behaves as logicians have canonically taken it to behave. This canon is not simply a high-handed scholarly imposition, made without respect for everyday language; rather it is a way of making precise the everyday truism that one who accepts a contradiction might as well accept anything, a truism which in turn gives substance to the nearly universal conviction that it is bad to accept contradictions. The point is, philosophers who respond to the Liar paradox by rejecting *ex falso quodlibet* are, like their competitors, committed to some unusual claims about the behavior of some of the expressions involved in the paradox.

see below, views that employ this unsound methodology overwhelmingly do conflict with the data.

These reflections motivate a kind of diagnosis and solution to the Liar paradox and its ilk that I call *aberrationism*. According to aberrationism, what occurs in the Liar paradox is a one-off aberration in the behavior of the expressions involved. In non-paradoxical sentences, the expressions that are implicated in the paradox behave exactly as the linguistic data tell us they do; but in situations in which this behavior would lead to paradox, these expressions behave slightly differently—as slightly as one can get away with allowing, without lapsing again into contradictions. Like their competitors, aberrationist views are still *ad hoc* in that they justify their claims about the semantic behavior of the problematic expressions primarily by showing these claims to be effective in circumventing paradox. But unlike their competitors, aberrationist view refrain from making unsupported assertions about any phenomena that are not directly implicated in the paradox. Moreover, if they can find ways to minimize the severity of the aberrations that they posit when paradox is afoot, these solutions can come close to allowing that the expressions concerned behave as they pre-theoretically appear to behave.

In this essay, I will present and defend an aberrationist approach to the Liar paradox and its ilk. In Section 2, I define a number of central concepts, present a general articulation of aberrationism, and show how aberrationism is supposed to work as a solution. Then in Section 3 I distinguish a *moderate* version of aberrationism from a *radical* version that I take to be implicitly advocated in (Smith 2006). This distinction helps to clarify the nature of aberrationism, and to make clear exactly what aberrationism needs in order to work as a solution. Then I argue that the moderate version of aberrationism is preferable, using my own preferred way of developing it as an illustrative example. Next, in Sections 4-10 I defend aberrationism in

general against various competing approaches. Lastly, in Section 11, I say a few more things to motivate the two important claims that constitute aberrationism. As with all other prominent approaches, my principal justification for these theses is that they provide a more effective way to circumvent paradox than the competing solutions can offer, and that they provide the overall best explanation of how natural languages manage to be consistent in the face of the paradox. But in Section 11, I argue that these two claims also have some independent motivation.

2. Important Definitions

2.1. Liar Sentences and Russellian Propositions

For any sentence x, let us say that x is a *Liar sentence* if x is the negation of a sentence whose grammatical subject refers to x and whose grammatical predicate is an expression that refers to truth. (Henceforth, I'll use 'refer' so that it includes not only reference for singular terms, but also the reference-like relation in which predicates and relation symbols can stand to properties and relations. I'm neutral on whether this is in fact just reference, or whether it is *ascription* or something else, distinct from reference.)

Here is an example of a Liar sentence, which I will frequently refer back to in what follows:

(A) A is not true.

Let's check that A is a Liar sentence. Note firstly that 'A' refers to the sentence 'A is not true'. This is guaranteed by the stipulative way that I introduced the name 'A'. Moreover, the word 'true' refers to truth. Or at least, while these intuitive claims are not beyond question, a virtue of aberrationism is that it allows us not to question them.[2] Secondly, note that 'A is not true' is the

[2] See Sections 4, 9, and 10 for discussion of views that question these claims.

negation of the sentence ‘A is true’. So ‘A is true’ is a sentence whose grammatical subject refers to A and whose grammatical predicate is an expression that refers to truth. Thus ‘A is not true’ is the negation of a sentence whose grammatical subject refers to A and whose grammatical predicate is an expression that refers to truth. Since A = ‘A is not true’, it follows that A is a Liar sentence.

Loosely speaking, the central claim of aberrationism is that Liar sentences fail to say of themselves that they are not true. More generally, this idea can be applied to all of the many different Liar-like sentences: no Liar-like sentence succeeds in saying what it appears to say, or in saying anything else that would generate paradox. However, because the word ‘say’ can be interpreted in several different ways, a rigorous formulation of my claims will need to be more precise. To move toward that, I will introduce the notion of a *Russellian proposition* below.

To see the need for a more precise formulation, notice that it follows from my definition of Liar sentences that there are Liar sentences which have grammatical predicates other than ‘is not true’. For example, assume that truth is the most philosophically interesting property. Then consider:

(D) It is not the case that D has the most philosophically interesting property.

Clearly D’s grammatical predicate is something other than ‘is not true’. But as one can easily check, D is the negation of a sentence whose grammatical subject is an expression that refers to D and whose grammatical predicate is an expression that (we are assuming) refers to truth. So, D is a Liar sentence. Though note that D would not be a Liar sentence in a world in which justice rather than truth was the most philosophically interesting property.

Now ask: does D say of itself that it is not true? On one way of using ‘say’, it certainly appears to. This is the use of ‘say’ that I will employ throughout. Still, on another quite common

way of using 'say', D does not even so much as appear to say of itself that it is not true. Rather, all it says is that it lacks the most philosophically interesting property, whatever that property may turn out to be.

The notion of a *Russellian proposition* helps to clarify the sense of 'say' in which D appears to say of itself that it is not true. Think of Russellian propositions as ordered tuples: for any object *o* and property P, one can identify the Russellian proposition that *o* has P with the pair <*o*, P>. Now let's apply this to the case of D. Imagine a world *w* in which the most philosophically interesting property is justice. On one way of using 'says', in *w*, D says that D lacks the most philosophically interesting property, just as it says in the actual world, @. But in *w*, the Russellian proposition whose negation D expresses is <D, justice>, not <D, truth>. By contrast, assuming that the most philosophically interesting property in @ is truth, in @ D appears to express the negation of the proposition <D, truth>. In that sense, D appears to say of itself that it is not true.

I just wrote 'appears to say' rather than 'says', and for an important reason. Applied to D, the characteristic claim of aberrationism is that D does not succeed in expressing the negation of the Russellian proposition <D, truth>. Nor, recalling the earlier example of A, does A succeed in expressing the negation of <A, truth>. More generally, no Liar sentence S succeeds in expressing the negation of the Russellian proposition <S, truth>.[3] And even more generally, something similar holds for all other Liar-like sentences, such as Curry sentences, Yablo sentences, etc.[4] In Section 2.3 I will state aberrationism in its most general form, a form that allows it to apply to

[3] Of course, this proposition may nonetheless exist. See Chapter 2 Section 5 and also Section 5 below for more explanation of how my view can accommodate the existence of such propositions.
[4] See Chapter 2 Section 6 for an in-depth discussion of how best to define 'Liar-like sentence'.

sentences of all these different kinds. To that end, I will now introduce a few more essential concepts.

2.2. Occurrences

Again, aberrationism claims that for any Liar sentence S, S fails to express the negation of the Russellian proposition <S, truth>. It may seem that there are only two ways that a declarative sentence S' can fail to express the negation of the Russellian proposition <S', truth>:

(i) no expression in S' refers to S', or

(ii) no expression in S' refers to truth.

After all, it is via either (i) or (ii) that the vast majority of sentences S' manage to avoid expressing the negation of such a proposition. For instance, the English sentence 'Grass is not blue' fails to express the negation of <'Grass is not blue', truth> because the word 'Grass' refers not to the sentence 'Grass is not blue' but rather to grass, and because the word 'blue' refers not to truth but to blueness.

However, given the way that I defined Liar sentences, neither (i) nor (ii) holds of them. Taking A from above as an example, surely the name 'A' as I defined it does refer to A, and surely the word 'true' does refer to truth. At least, it is worth making an effort to retain these intuitive claims.[5] But in that case, how could A possibly fail to express the negation of the Russellian proposition <A, truth>? Aberrationism claims that this can happen, and in a way different from (i) and (ii):

(I) either the occurrence of 'A' in A fails to refer to A or the occurrence of 'true' in A fails to refer to truth, and

(II) which Russellian proposition (if any) A expresses the negation of is determined by the reference of the occurrences of 'true' and 'A' in sentence A, rather than by the reference of the expressions 'true' and 'A' *simpliciter*.

[5] See Section 10 for further discussion.

(I) and (II) invoke the notion of an *occurrence* of an expression, which may be unfamiliar. Since this notion is quite central to aberrationism, I will take a moment to clarify it.

First let's consider a few examples. A variable can occur multiple times in a formula, with some but not all of these occurrences being bound by a quantifier. E.g., consider the following formula:

x is yellow and for all x, if x is blue then x is great

The first occurrence of 'x' is unbound, but the second, third, and fourth occurrences are bound by the universal quantifier. For another example, there are three occurrences of the letter 'c' in the word 'occurrence', and two occurrences of the word 'an' in the sentence 'An occurrence of an expression is a lovely thing'.

Formally, one can represent an occurrence of an expression as an ordered pair of the expression and a *component context*. One can think of a component context as the result of using a hole-punch to punch out a single occurrence of an expression from the sentence or formula, replacing it with a blank. Thus the occurrence of 'true' in sentence A can be represented as the ordered pair <'true', 'A is not *'>, where I use '*' to designate a blank space. In what follows it will be harmless to identify occurrences with these ordered pairs. (Though we should carefully distinguish them from Russellian propositions, which we are also taking to be ordered pairs. This is easy, since unlike any component context, every Russellian proposition has at least one constituent that is either a property or a relation.)

It is easy to mix up occurrences and tokens, but in our discussion this distinction is important. To see the difference, consider that a sentence can exist and be part of a language, even if no one has yet tokened it—that is, uttered or written it. The same goes for occurrences of words in a sentence. E.g., as long as the sentence 'An occurrence of an expression is a lovely

thing' exists, so, too, do its two occurrences of 'an'. Those occurrences would have existed even if the sentence had never been tokened, just as the sentence itself would have. On the other hand, no (actual) tokens of an expression exist until someone actually writes or utters that expression. Similarly, tokens are tied to the particular time and place in which they are brought about. If I write the sentence 'An occurrence of an expression is a lovely thing' on the blackboard, I produce two tokens of 'an' that did not exist before. By contrast, the two occurrences of 'an' in that sentence existed before I decided to write anything on the blackboard. Again, their existence depends only on that of the word 'an' and the sentence in which they occur.

(I) and (II) are formulated with specific attention to the Liar sentence A. However, as I mentioned earlier, there are a variety of different paradoxes in the vicinity of the Liar that arise from sentences of similar kinds. Aberrationism applies to all of these paradoxes, and so must be formulated in a general way. Thus, in the next subsection I'll give a general formulation, leading up to it with several important definitions.[6]

2.3. Liar-like Paradoxes, Liar-like Sentences, and the Official Statement of Aberrationism

Let us begin by looking at the word 'satisfies', in logicians' English, which I'll define by the following rules:

(Sat-intro) $\dfrac{\varphi(a_1,\ldots,a_n)}{\text{satisfies } (\langle a_1,\ldots,a_n\rangle, \text{`}\varphi(x_1,\ldots,x_n)\text{'})}$

(Sat-elim) $\dfrac{\text{satisfies } (\langle a_1,\ldots,a_n\rangle, \text{`}\varphi(x_1,\ldots,x_n)\text{'})}{\varphi(a_1,\ldots,a_n)}$

[6] See Chapter 2 for a detailed application of aberrationism to a wide variety of Liar-like paradoxes, making frequent use of the generality of the definition.

Suppressing issues of context-sensitivity, here the 21st letter of the Greek alphabet is a substitutional variable ranging over all formulas of English, for each Arabic numeral, the 24th letter of the lowercase Latin alphabet subscripted by that numeral is a substitutional variable ranging over all objectual variables in English, and for each Arabic numeral, the 1st letter of the lowercase Latin alphabet subscripted by that numeral is a substitutional variable ranging over all English singular terms.

An important special case of (Sat-intro) and (Sat-elim) are the two rules (T-intro) and (T-elim) below. (T-intro) and (T-elim) "govern" the word 'true' in at least the sense that they are widely accepted and deeply entrenched. By and large we are disposed to follow them, and, when we are feeling helpful, to correct uses of 'true' that deviate from them. (T-intro) and (T-elim) will figure prominently in what follows, so it will be good to keep them in mind throughout.

(T-intro) $\dfrac{\varphi}{\text{`}\varphi\text{' is true}}$

(T-elim) $\dfrac{\text{`}\varphi\text{' is true}}{\varphi}$

Suppressing issues of ambiguity and context-sensitivity, here the 21st letter of the Greek alphabet is a substitutional variable ranging over all declarative sentences of English. A moment's reflection reveals (T-intro) and (T-elim) to be, modulo some trivial cosmetic differences, the special cases of (Sat-intro) and (Sat-elim) in which n = 0. This follows straightforwardly from the fact that declarative sentences are simply formulas with no free variables; so, for example, when n = 0, 'satisfies ($<a_1,\ldots, a_n>$, '$S(x_1,\ldots,x_n)$')' becomes 'satisfies ('S')'. The trivial cosmetic differences are those between, e.g., 'satisfies ('S')' and ''S' is satisfied', and between ''S' is satisfied' and ''S' is true'.[7]

[7] Setting the two rules side by side makes the point obvious. E.g.:

Now for just a few more definitions. Let us say that an expression *e* is an *alethic expression* if

(Base Clause) *e* is inter-substitutable with an expression that is governed by (Sat-intro) and (Sat-elim), for some n,

or

(Inductive Clause) the result of replacing each non-primitive component expression in *e* with its definiens contains an alethic expression.

By (Base Clause), 'true' is an alethic expression, since it is governed by the special case of (Sat-intro) and (Sat-elim) where n = 0. Likewise, by (Base Clause), any other expression that refers either to truth or to satisfaction is an alethic expression, since it is intersubstitutable with 'true' or 'satisfies', respectively. So, for example, if Brian is thinking about truth right now, then right now the expression 'the property that Brian is thinking about right now' is an alethic expression.

Let us say that a paradox is a *Liar-like paradox* if it makes ineliminable use of (Sat-intro) and (Sat-elim), perhaps modulo minor cosmetic differences. So in particular, classic versions of the Liar paradox count, as they make such use of (T-intro) and (T-elim). With this definition of 'Liar-like paradox' in hand, we can now say that a sentence is a *Liar-like sentence* if it gives rise to a paradox that is Liar-like.[8]

For our purposes, it will be safe to assume that every Liar-like sentence contains some alethic expression. Still, not every occurrence of an alethic expression in such a sentence need be involved in giving rise to the associated paradox. Let's refer to the problem-causing occurrences

---

(T-intro) $\dfrac{\varphi}{\text{`}\varphi\text{' is true}}$

(Sat-intro) $\dfrac{\varphi}{\text{satisfies(`}\varphi\text{')}}$

Each of these rules allows one to derive from a declarative sentence the result of applying a predicate to a quote name of that sentence.

[8] See Chapter 2 Section 6 for further explanation of the terms 'paradox', 'ineliminable use', and 'gives rise to'.

as *key occurrences*. Key occurrences are hard to define rigorously, but it is very easy to get the main idea. Consider (A_w):

(A_w) A_w is not true, and 'War causes suffering' is true.

Clearly, only the first occurrence of 'true' plays an essential role in generating the paradox associated with A_w. So only that first occurrence is a key occurrence.[9]

Given these definitions, we can now define *aberrationism* as the conjunction of the following two claims:

(Aberrations) For any Liar-like sentence, the key occurrences of its alethic expressions differ in reference from the expressions of which they are occurrences.

(Determined) When an occurrence of an expression in a Liar-like sentence differs in reference from the expression of which it is an occurrence, what Russellian propositions (if any) the sentence expresses, expresses the negation of, expresses the conjunction of, the disjunction of, etc., is determined by what the occurrence refers to, rather than by what the expression *simpliciter* refers to.[10]

A quick remark on my reason for including (Determined) here. For all (Aberrations) says, what a Liar-like sentence expresses, expresses the negation of, expresses the conjunction of, the disjunction of, etc. could be determined solely by the reference of its expressions *simpliciter*, not by the occurrences of these expressions in that sentence. In that case, even if an occurrence

[9] For present purposes, the following definition of 'key occurrence' will suffice. Given a Liar-like sentence S, fix the Liar-like paradox D to which S gives rise. Since D makes ineliminable use of (Sat-intro) and (Sat-elim), S must contain at least one occurrence of at least one alethic expression *e*. Let a *key occurrence* of *e* in S be an occurrence of *e* in S to which one of (Sat-intro) or (Sat-elim) is applied, or that is used to introduce a step to which the rule Contradiction Introduction is applied.

[10] As a historical note, when it comes to sentences that attribute propositional attitudes, (Frege 1892) is committed to something quite similar to (Aberrations) and (Determined). Frege holds that when a word occurs in the that-phrase of such a sentence, it refers to what is usually its sense (its *customary sense*), rather than what is usually its referent (its *customary referent*). But this amounts to the claim that the occurrence of the word in the that-phrase differs in reference from its other occurrences, or, if one prefers, from the word *simpliciter*. Moreover, in such cases it is reference by the occurrence in the that-phrase that contributes to the truth-conditions of the attitude-attributing sentence, not reference by the word *simpliciter* or by its other occurrences. So, for what it is worth, views of the sort I defend here are not without historical precedent. I am grateful to Harold Hodes for bringing this observation to my attention.

undergoes an aberration in its reference, that does not change what the sentence expresses, expresses the negation of, etc.; so, for example, Liar sentences succeed after all in saying of themselves that they are not true. Clearly, in that scenario, positing aberrations does nothing to help explain how natural languages manage to be coherent in the face of the paradox. For that to work, we need to claim that the aberrations have an effect on what proposition the sentence in question expresses, expresses the negation of, etc. (Determined) guarantees this.

Now that I have stated aberrationism precisely and in a way that enables it to apply to many different Liar-like sentences, let us see in greater detail exactly how it is supposed to work as a solution to the Liar paradox.

2.4. The Paradoxical Reasoning Connected with Sentence A

The motivating idea behind aberrationism is that if S is a Liar sentence, then the reasoning involved in the paradox to which S gives rise should best be interpreted as a *reductio* of the assumption that S expresses the negation of the Russellian proposition <S, truth>; and similarly for other Liar-like sentences.[11] To see how this idea works as a solution, consider the following derivation. (In what follows, any step prefixed by '|' is or occurs under an undischarged assumption.)

[11] Other theorists make similar moves. In Section 4 below I discuss (Goldstein 2009), which denies that any seemingly-paradox-inducing sentence expresses a proposition. This view is also discussed in (Parsons 1974) and (Glanzberg 2001), and has also been attributed to (Kripke 1975), on some interpretations.

1. A = ‘A is not true’	(definition of ‘A’)
2. | A is true	(Assume for *reductio*)
3. | ‘A is not true’ is true	(Substitution, (1), (2))
4. | A is not true	((T-elim), (3))
5. | Contradiction	(Contradiction introduction, (2), (4))
6. A is not true	(Negation Introduction, (1)-(4))
7. ‘A is not true’ is true	((T-intro), (6))
8. A is true	(Substitution, (1), (7))
9. Contradiction	(Contradiction introduction, (6), (8))

For this derivation to be sound, (4) must express the negation of what (2) expresses, and (6) must express the negation of what (8) expresses. (Otherwise, the instances of Contradiction Introduction fail to go through.) Given that (2) and (8) express the Russellian proposition <A, truth>, (4) and (6) must express the negation of that proposition.

However, on pain of absurdity, the derivation is not sound. As the best explanation of this, aberrationism proposes that (4) and (6) fail to express the negation of the proposition <A, truth>; that, aberrationists claim, is why the derivation does not go through. For (4) and (6) not to express the negation of the proposition <A, truth>, either the occurrence of ‘A’ in A must fail to refer to A, or the occurrence of ‘true’ in A must fail to refer to truth. (For if the occurrence of ‘A’ in A refers to A and the occurrence of ‘true’ in A refers to truth, then surely A expresses the negation of the Russellian proposition <A, truth>.) Aberrationism claims that of these two possibilities, the one that makes for a better explanation is that the occurrence of ‘true’ in A fails to refer to truth.[12]

As I have been stressing, aberrationists market their view as the best explanation of how natural languages avoid incoherence in the face of the Liar paradox. Of course, other philosophers disagree; any bit of Liar-like reasoning can be cast as a *reductio* of any number of claims. Thus, it needs to be shown that the explanation that aberrationism recommends is

[12] In Section 10 I will discuss views that instead target the occurrence of ‘A’.

genuinely better than the alternatives. That will be the business of Sections 4-10 below. However, before getting to that, I will clarify exactly what (Aberrations) and (Determined) commit us to; what exactly is and what is not essential for aberrationism's functioning as a solution.

3. Radical vs. Moderate Aberrationism

3.1. Smith

In (Smith 2006), Nicholas Smith defends an approach to the Liar paradox that I take to be best classified as a version of aberrationism. He claims that in uttering a Liar sentence like A from above (recall: 'A is not true'),

> either I do not refer to the sentence I utter, or I do not say of what I refer to that it is not *true*. I claim this is the *solution* to the Liar paradox. There is no paradox, because there is no Liar sentence—that is, no sentence which says of itself only that it is not true. When you try to construct such a sentence, you fail: either you do not refer to what you wanted to refer to (the very [sentence] you uttered), or you do not say of what you do refer to that it is not *true* (Smith 2006, p.182, emphasis his).

While Smith's talk of what one refers to by making an utterance suggests that his concern is with speech acts, at the same time he is clearly also committed to a strong claim about what sentences themselves can and cannot say: namely, the claim that no sentence can say of itself only that it is not true.[13] Moreover, in order for a sentence S to say of itself that it is not true, it would have to express the negation of the Russellian proposition <S, truth>. So, Smith is committed to denying that any sentence S can express the negation of <S, truth>. Even further, throughout his paper

[13] Thus, his claim is not merely, as one might have supposed from the talk of utterances, that it is impossible to make an utterance which says the same thing as a sentence which says of itself that it is not true. Note that if he takes this view and also admits that there are such sentences, then he is committed to the claim that there are some propositions that no utterance can express—namely, any propositions expressed by such a sentence. It would be uncharitable to attribute this claim to Smith, given things he says elsewhere (see his discussion of "incompleteness").

Smith uses ‘true’ and ‘A’ to refer to truth and A, respectively. So, he is implicitly committed to the claim that ‘A’ refers to A and ‘true’ refers to truth. Thus, the only way he can hold that A fails to express the negation of the Russellian proposition <A, truth>, and thus that A fails to say of itself that it is not true, is by endorsing aberrationism. So, I take Smith to be tacitly committed to aberrationism.

However, Smith couples his tacit commitment to aberrationism with the rejection of some highly plausible theses about language. In the rest of this section, I will show that these theses are in fact compatible with aberrationism. This observation will serve to clarify the nature of aberrationism. Moreover, it will be a convenient way to get my own cards on the table. I will defend the compatibility claim by presenting my preferred way to develop aberrationism, which retains the theses that Smith rejects. I will then use my own preferred view as an example to illustrate the advantages available to versions of aberrationism that retain these theses.

3.2. Semantic Supervenience and Semantic Regularity

Most philosophers who discuss the Liar paradox—and this includes me—are at least tacitly committed to holding that meaning and reference are part of the natural world. In particular, most philosophers assume that when our words refer, what they refer to is determined by our linguistic behavior involving these words, the mental states we use them to express, and how our uses of them are related to our physical environment. (Henceforth I will include all of these things when I speak of the *use* of an expression.) It follows that whenever an expression refers to something, there are some non-semantic facts, such as facts about how the expression is used, which determine that it refers to that thing. According to philosophers who accept this, these non-semantic facts determine that either within Liar sentences or at some step in any Liar-

like paradox, some of the expressions involved (or, in the aberrationist case, some of their occurrences) do not refer to what they would have to refer to, in order for a contradiction to follow. However, in several striking passages, Smith says things that can be interpreted as rejecting this view. For example:

> Sometimes our words do not mean what we want[14] them to mean: not due to hidden complexities of our semantic mechanisms, operating behind the scenes to produce unforeseen results—i.e., not for some principled, bottom-up reasons; but because our words *cannot* mean what we want them to mean, and so our semantic mechanisms simply break or malfunction, and some of our words get assigned meanings more or less randomly (p.195).

Here I understand Smith's talk of meaning as referring to reference. (Whether or not this was Smith's actual intention, the view that I wish to discuss is the one which makes the above claims about reference.)

On one rather strong reading of Smith's claim that some expressions "get assigned meanings more or less randomly" rather than for any "bottom-up reasons," he is committed to denying a claim that I will call *Semantic Supervenience*:

> (Semantic Supervenience) The semantic properties strongly locally supervene on the non-semantic properties.

For two kinds of properties A and B, the A-properties *strongly locally supervene* on the B-properties if and only if for any possible worlds w_1 and w_2 and any individuals x in w_1 and y in w_2, if x in w_1 is indiscernible from y in w_2 with respect to B-properties, then x in w_1 is indiscernible from y in w_2 with respect to A-properties.[15]

On this reading, in denying Semantic Supervenience one would have to claim that there are two possible worlds w_1 and w_2, and two expressions, x and y, such that x in w_1 is

[14] Admittedly, speakers' desires and intentions are only one factor that might contribute to determining the reference of their expressions. However, we will soon see that Smith is likewise skeptical about the determinative influence of the other putatively contributing factors.

[15] See (Kim 1987) and (McLaughlin and Bennett 2014) for this definition and further discussion.

indistinguishable from y in w_2 with respect to all non-semantic properties—in particular, properties having to do with use—but x in w_1 and y in w_2 do differ with respect to their semantic properties. Applied to the case of reference and put in more informal terms, the view would say that expressions can acquire referents in ways that are entirely unaffected by how they (the expressions) are used. It is in that sense that they "get assigned meanings more or less randomly," and not for "bottom-up" reasons. On this reading of Smith, a bottom-up reason why an expression means what it does would include non-semantic facts, such as facts about how the expression is used.

Now, it may be unfair to saddle Smith with the denial of Semantic Supervenience, since 'randomly' can be interpreted several ways and anyhow it is not clear to what extent his approach to the Liar paradox relies on his claim that sometimes reference is "determined more or less randomly." Still, it will be useful to have Semantic Supervenience (and its negation) on the table in what follows; what matters for my purposes is not exactly which view Smith holds, but rather the fates of the various views in the vicinity. In that connection, it is worth noting that (Kearns & Magidor 2012) find reasons independent of the Liar paradox to deny Semantic Supervenience. Philosophers sympathetic to their view might be further encouraged if the view turned out to provide a distinctive solution to the Liar that could not be had by other means. An important upshot of my discussion will be that this is not so; one can reap the benefits of aberrationism without embracing Semantic Supervenience.

A somewhat more moderate view—likely the better interpretation of Smith—is neutral about Semantic Supervenience, but instead denies a thesis that Smith calls 'Semantic Regularity':

(Semantic Regularity) There are *perfectly* reliable, principled relationships between our behavior, mental states and physical environment on the one hand, and what [our words mean[16]] on the other hand (p.188, emphasis mine).

Smith presents the denial of Semantic Regularity as being quite central to his solution to the Liar paradox. However, an important upshot of this section will be that one can adopt a similar solution while still reasonably claiming to retain Semantic Regularity and Semantic Supervenience.

I say 'reasonably' because as written, Semantic Regularity is obscure, owing to obscurity in the phrase 'perfectly reliable, principled relationships'. As I prefer to understand 'perfectly reliable, principled relationships', my own view is that there is indeed a relationship of this kind between the semantic and the non-semantic, but that the relationship is more complicated than one might initially have thought. In what follows, I will present my view in this way, since that best serves to highlight the differences between me and Smith. Once I have said more about what I take the relationship of the semantic to the non-semantic to be, it will become much clearer in what ways this relationship can be considered reliable and principled, and in what ways complicated.

Let us call any version of aberrationism that accepts both Semantic Supervenience and Semantic Regularity a *moderate* version. And let us call any version that rejects one, the other, or both a *radical* version. In crudest terms, the difference between moderate and radical aberrationism is that only moderate approaches hold that it is possible to have a (correct) theory of reference. (This, at least, is the issue that I take to be at stake in Smith's rejection of Semantic Regularity.) Even this crude way of putting the distinction helps to clarify aberrationism:

[16] Smith writes "what we mean by our utterances". Since my concern is with expression types rather than their tokens, I will address the version of Semantic Regularity that concerns types.

although aberrationism is committed to positing referential aberrations, and thus to denying that sentences always say what they pre-theoretically appear to say, it is not committed to the claim that what our words mean can float free of how we use them, in a way that precludes there being a correct theory of reference. One of the main goals of this section is to show that the fate of theories of reference is in fact a separate question, on which advocates of aberrationism can in principle take different stances.

In the rest of this section and in Chapter 2, I defend moderate aberrationism, using my own preferred variant as an example to illustrate its advantages over radical aberrationism. The point is essentially as follows. Because they doubt the robustness of the relation between semantic phenomena and non-semantic phenomena, radical versions of aberrationism are not in a good position to explain our very vivid sense of what Liar-like sentences seem to say by appealing to facts about language use, or even, setting explanation aside, to identify what these sentences in fact say by looking at facts about language use. By contrast, moderate aberrationism has more resources available on both these scores. On moderate views, when an occurrence of a word cannot co-refer with that word, it need not follow that the reference of the occurrence has nothing to do with how we use the word. Rather, there are patterns that link facts about a word's use to the reference of its occurrences. In particular, my own preferred moderate view claims that these patterns determine that the occurrence's referent is something similar to the referent of the word *simpliciter*.

While I will present my own preferred version of moderate aberrationism in some detail, readers should keep in mind that my goal here is not to defend that moderate view in particular, but rather merely to illustrate the advantages of moderate views in general over their radical cousins in a plausible, concrete way.

3.3. A Concrete Moderate Proposal

3.3.1. Principles, Respect, and Closeness

Let us temporarily set aside reference by occurrences and return to the more familiar conception of reference as a relation between expressions *simpliciter* and things in the world. Plausibly, what if anything an expression refers to is determined at least in part by how the expression is used. Moreover, it is quite plausible that, for many linguistic expressions, there are certain *principles*—inference rules or sentences (types) in which the expression occurs—which play an especially significant role in determining the reference of the expression. In general, the principles for an expression form a proper subset[17] of the totality of sentences and inference rules in which the expression figures. For example, they might include the sentences and inference rules which play widely accepted and deeply entrenched roles in governing people's uses of the expression.[18] For example, throughout his work David Lewis claimed that a *theoretical term*—a term that is introduced into a language by a theory—has the conjunction of the sentences of the theory as a principle, in my sense of the word 'principle'.[19] And Matti Eklund has suggested that the same holds for 'true' and the inference rules (T-intro) and (T-elim).[20]

To explain how an expression's being governed by a principle can influence the determination of its reference, I will need to introduce the notions of *respect* and *closeness*. To get a grip on these notions, it will help to consider some quite closely related ideas from (Lewis 1972), those of a *realization* and a *near-realization*. Lewis introduces these notions via an

[17] It is plausible that this can be a fuzzy set (or elsewise a set with vague boundaries), allowing for vagueness as to which sentences or rules are principles.
[18] See below for more on the relationship between this idea and the notion of meaning-constitutivity developed in (Eklund 2007).
[19] For example, see (Lewis 1970), (Lewis 1972), (Lewis 1994) p.58, (Lewis 1997), (Eklund 2002), and (Eklund 2007). See also (Braddon-Mitchell and Nola 1997) and (Melia and Saatsi 2006).
[20] See, e.g., (Eklund 2002) p.269 footnote 41.

example. He imagines a detective introducing a theory, in an attempt to explain the data at a crime scene. The theory, displayed below, introduces three terms, 'X', 'Y', and 'Z':

> X, Y and Z conspired to murder Mr. Body. Seventeen years ago, in the gold fields of Uganda, X was Body's partner…Last week, Y and Z conferred in a bar in Reading…Tuesday night at 11:17, Y went to the attic and set a time bomb...Seventeen minutes later, X met Z in the billiard room and gave him the lead pipe…Just when the bomb went off in the attic, X fired three shots into the study through the French windows…(p.250)

Then, Lewis imagines, we learn that the story "is true of a certain three people: Plum, Peacock and Mustard. If we put the name 'Plum' in place of 'X', 'Peacock' in place of 'Y', and 'Mustard' in place of 'Z' throughout, we get a true story about the doings of those three people." Lewis describes this situation by saying that "Plum, Peacock and Mustard together *realize* (or *are a realization of*) the detective's theory" (p.251, italics mine). Next, he imagines us learning that this triple *uniquely realizes* the theory—that is, this triple and no other realizes the theory. In that case,

> we would surely conclude that X, Y and Z in the story were Plum, Peacock and Mustard. I maintain that we would be compelled so to conclude, given the senses borne by the terms 'X' , 'Y' and 'Z' in virtue of the way the detective introduced them in his theorizing, and given our information about Plum, Peacock and Mustard. In telling his story, the detective set forth three roles and said that they were occupied by X, Y and Z. He must have specified the meanings of the three T-terms 'X', 'Y' and 'Z' thereby; for they had meanings afterwards, they had none before, and nothing else was done to give them meanings. They were introduced by an implicit functional definition, being reserved to name the occupants of the three roles. When we find out who are the occupants of the three roles, we find out who are X, Y and Z. (p.251)

In my terminology, the sentence that expresses the detective's theory counts as a principle for the expressions 'X', 'Y', and 'Z'. And *respecting* that principle is just like realizing it, in Lewis' sense. Similarly, *coming close to respecting* the principle is just like being a near realization of it.

With these comparisons in place, below are some more precise definitions of *respecting* and *coming close to respecting*.[21]

Suppose we are given an expression *e* and a principle P for *e*. Assume for simplicity that other than *e*, every singular term, function constant, relation symbol, and predicate in P is context-insensitive and refers to something. Given an entity *y*, if P is a sentence then *y respects* P if and only if, if every occurrence of *e* in P were to refer to *y* then P would be true.[22] (Compare Lewis' remark from above: "If we put the name 'Plum' in place of 'X', 'Peacock' in place of 'Y', and 'Mustard' in place of 'Z' throughout, we get a true story about the doings of those three people.")

The case of sentences is helpful for getting a grip on the notion of respect and connecting it with Lewis' ideas, but more important for my purposes is the case of respect as it functions for rules of inference. If P is a rule of inference, then *y respects* P if and only if *y* respects every instance of P. *y respects* a given instance N of P if and only if:

a) if P is an introduction rule, then: if every occurrence of *e* in N that is introduced by the application of P were to refer to *y*, then N would be truth-preserving, and

b) if P is an elimination rule, then: if every occurrence of *e* in N that is eliminated by the application of P were to refer to *y*, then N would be truth-preserving.[23]

[21] For an alternative definition, please see the Appendix.

[22] It is worth noting that in the case of sentences, the notion of respect is quite similar to the notion of satisfaction, introduced in Section 2.3. The difference is that respect is defined in terms of a modal condition on reference and in terms of truth, whereas the notion of satisfaction is introduced simply by stipulating that the rules (Sat-intro) and (Sat-elim) are henceforth to define the word 'satisfies'.

[23] Since I have characterized reference determination in terms of respect and then characterized respect in terms of subjunctive conditions that involve reference, one might worry that my remarks about reference amount to an explanatory circle. However, keep in mind that my aim in taking about reference is rather modest. I am not out to provide a full explanation, in non-semantic terms, of how reference is determined. I am just presenting a fact that I take to be partly explanatory of reference-determination. Please see the Appendix for an alternative, model-theoretic definition of respect that does not invoke reference, and an examination of the consequences of adopting this definition, vis à vis my diagnosis of the Liar paradox.

For my purposes, it will be harmless to ignore the case in which P is a rule of inference that is neither an introduction nor an elimination rule.

Now that I have introduced the notion of respect, I will introduce that of *closeness*, again drawing inspiration from (Lewis 1972). Here the relevant notion from Lewis is that of a *near-realization*, which he introduces after explaining the notion of a realization:

> A complication: what if the theorizing detective has made one little mistake? He should have said that Y went to the attic at 11:37, not 11:17. The story as told is unrealized, true of no one. But another story is realized, indeed uniquely realized: the story we get by deleting or correcting the little mistake. We can say that the story as told is nearly realized, has a unique *near-realization*. (The notion of a near-realization is hard to analyze, but easy to understand.) In this case the [terms 'X', 'Y', and 'Z'] ought to name the components of the near-realization. More generally: they should name the components of the nearest realization of the theory, provided there is a unique nearest realization and it is near enough. Only if the story comes nowhere near to being realized, or if there are two equally near nearest realizations, should we resort to treating the [terms 'X', 'Y', and 'Z'] like improper descriptions. (p.252)

Just as something can come close to realizing a theory and thereby be a near-realization of the theory, something can come close to respecting a principle.

Unfortunately, like the notion of a near-realization, the notion of coming close to respecting a principle is hard to analyze. One might initially hope to define relative closeness to respecting a principle in terms of relative number of instances satisfied: for all principles P and entities y_1 and y_2, y_1 comes closer to respecting P than y_2 does if and only if y_1 respects more instances of P than y_2 does. But (McGee 1992) demonstrates that this way of cashing out closeness fails.[24] Still, although the notion is difficult to analyze, (Eklund 2005) argues that

[24] To see why, let us consider the schema (T) below rather than (T-intro) and (T-elim), considering (T) as a principle for 'true':

(T) 'φ' is true if and only if φ

(McGee 1992) proves that there are some maximal consistent sets of instances of (T) that entail ''2 + 2 = 4' is true', and others that entail ''2 + 2 = 5' is true'. (These sets are maximal in the sense that any proper superset of them is inconsistent.) But, fixing such sets Δ and Γ, the point is this: when asking which objects come closer to respecting (T), other things being equal it would be perverse to allow that an object which satisfies all and only the instances of (T) that are contained in Γ comes as close as an object which satisfies all and only the instances of (T) that are

"reliance on a notion of closeness is necessary…from a number of…theoretical viewpoints," and "our grasp of the notion is sufficient to make reasoned judgments about closeness" (p.50). One convincing example that Eklund raises has to do with the idea that the task of a scientist—in Eklund's case a semantic theorist, though it applies across all the sciences—is to come up with a theory that best fits the data. The notion of closeness, Eklund points out, "is what underlies the talk of 'best fit'" (p.51).[25]

As a final remark concerning closeness, it is worth noting that it is judgments about relative closeness—about one thing's coming closer than another to respecting a principle—that are most important for my purposes, and that principled judgments about relative closeness can be easy to come by, even when judgments of absolute closeness are not. An example will help bring out the point. Say a sentence S is *grounded* if, given the truth values of all the sentences that contain no uses of 'true', one can determine the truth value of S in finitely many steps.[26] Thus, for example, Liar sentences are ungrounded, as are sentences such as 'This sentence is true'; in neither case does evaluation of the sentence in question bottom out in consideration of the non-linguistic part of reality. Now imagine that there is a property Q that respects all instances of (T-intro) and (T-elim) for grounded sentences, but fails for all ungrounded sentences. And now contrast Q with the property *being made of goat cheese*. The point is that while neither Q nor *being made of goat cheese* respects both rules, Q clearly comes closer to doing so.

contained in Δ. That is to say, when we assess how closely candidate referents for 'true' come to respecting (T), there needs to be some weighting that assigns greater importance to the respect of some, and less importance to the respect of other, instances of (T).

[25] See also Eklund's example of the two linguistic communities L and L* on pp.51-52.

[26] See (Kripke 1975) and Section 5 below for more on groundedness.

For example, *being made of goat cheese* fails to respect any instance of (T-intro) that involves a true sentence, regardless of whether that sentence is grounded. Consider for instance 'Grass is green'. That gives us the following instance of (T-intro):

Grass is green.
_____ (T-intro)
'Grass is green' is true.

If the occurrence of 'true' in the conclusion were to refer to *being made of goat cheese*, then the conclusion would say that the sentence 'Grass is green' is made of goat cheese. This inference would then fail to be truth-preserving, since it would have a true premise and a false conclusion. It is now plain that the same goes for any other true sentence that one substitutes for 'Grass is green'. By contrast, in the instance of (T-intro) just displayed, if the occurrence of 'true' were to refer to Q then by our assumption about Q, both the premise and the conclusion would be true and so that instance of (T-intro) would be truth-preserving. And in fact, the same goes for all other grounded sentences. Even if it is unclear exactly what would make Q come closer than *being made of goat cheese* comes to respecting (T-intro), that it does come closer is hard to dispute.

3.3.2. A Lewisian Story about Reference Determination

With the notions of principles, respect, and closeness in place, I can now give a semi-detailed sketch of a concrete view that illustrates my main point against radical versions of aberrationism. That point, again, is that one can reap the benefits of positing one-off aberrations while retaining Semantic Regularity. On the view I am sketching, the reference of any occurrence of 'true' is significantly influenced by how we use the word 'true'—in particular, by

the central role of (T-intro) and (T-elim) in governing our uses of this word. That holds even for occurrences that cannot, on pain of contradiction, refer to truth.

For the moment, let us continue with our familiar talk of reference as a relation between expressions *simpliciter* and objects. Now consider the following Lewisian claims about reference:[27]

i. For some expressions *e* and principles P involving *e*, *e* refers to that entity *y*,[28] if there is one, which comes closest to respecting P, and
ii. if there are several entities $y_1,\ldots y_n$ which respect P or come equally close to respecting P, then *e* is indeterminate in reference as between $y_1,\ldots y_n$, and
iii. if nothing even comes close to respecting P, then *e* altogether fails to refer.

In broadest outline, my plan is to extend (i)-(iii) to the case of reference by occurrences. The idea is as follows. Truth is the property that comes closest to respecting (T-intro) and (T-elim).[29] So most occurrences of the word 'true' refer to truth. However, on pain of contradiction, this cannot hold for the key occurrence of 'true' in any Liar sentence. Still, it need not follow that there is no principled, reliable relationship between the reference of this occurrence and (T-intro) and (T-elim). Rather, such an occurrence refers to that property, if there is one, which comes next closest (after truth) to respecting these rules. If there is more than one such property, then the occurrence is indeterminate in reference. If there is no such property, then the occurrence fails to refer.

[27] See (Lewis 1970) and (Lewis 1972), including the passages quoted above in the main text, for expressions of (i). Lewis advocates (ii) in (Lewis 1994), but then adopts a more qualified position in (Lewis 1997). His later position is that for *e* to be indeterminate, the candidate referents have to be sufficiently similar. Otherwise, if multiple very different entities come equally close to respecting P, then *e* has no referent whatsoever. I find this modification plausible, but for simplicity I omit it from the main text, as it will not feature importantly in what follows. (Once one sees what the candidate referents for occurrences of 'true' in Liar sentences are, it becomes plausible that these candidates are similar enough so that these occurrences of 'true' are indeterminate. See below for my description of the candidate referents.) For other discussions of Lewisian views about reference for theoretical terms, see (Eklund 2002), and (Eklund 2005).

[28] One might think that '*e* refers to *y*' immediately entails 'every occurrence of *e* refers to *y*'. But, of course, a central claim constitutive of aberrationism is that this is not so. See Section 11 for further discussion.

[29] See Section 3.3.3.2 below for an argument.

Plainly, however, these claims about the key occurrences of 'true' in Liar sentences do not immediately follow from (i)-(iii), since (i)-(iii) only address reference by expressions *simpliciter*, not reference by occurrences. Thus, I propose to supplement (i)-(iii) so as to address reference by occurrences:

> iv. Suppose E is an occurrence of *e*. If there is some entity *y* such that *e* refers to *y*, then E refers to *y*, unless this would lead to a contradiction. If this would lead to a contradiction, then E refers to that unique thing, if there is one, which is next in line for coming closest to respecting P. If multiple things $y_1,\ldots y_n$ are next in line, then the reference of E is indeterminate as between $y_1,\ldots y_n$. If nothing comes sufficiently close, then E fails to refer.

(iv) is clearly in the same Lewisian spirit as (i)-(iii). And note that with acceptance of (iv), we are able to say that while 'true' *simpliciter* refers to truth, the key occurrences of 'true' in Liar sentences do not.[30] As (iv) specifies, the "next-in-line rule" applies to expression-occurrences, and it is only occurrences of 'true' in Liar sentences and their ilk that cannot, on pain of contradiction, refer to truth. Thus, all the rest of the occurrences of 'true' in the wide variety of non-paradoxical English sentences are free to refer to the thing that comes closest to respecting (T-intro) and (T-elim), namely, (as we've assumed) truth itself; and the occurrences of 'true' in Liar-like sentences are free not to refer to truth.[31]

---

[30] Here again, I assume that truth comes closest to respecting (T-intro) and (T-elim). See Section 3.3.3.2.

[31] Here I gloss over a complication. Consider the instances of (T-intro) and (T-elim) that involve sentence A. Let E_0 be the underlined occurrence of 'true', and E_1 be the non-underlined occurrence:

(IN)

A is not <u>true</u>
_____ (T-intro)
'A is not <u>true</u>' is true

(OUT)

'A is not <u>true</u>' is true
_____ (T-elim)
A is not <u>true</u>

Now, (iv) says that E_0 refers to truth unless that would lead to a contradiction. But whether it would lead to a contradiction depends also on what E_1 refers to. And vice versa. That is, one can avoid contradiction by allowing either E_0 or E_1 to fail to refer to truth; one isn't forced to choose. So, (iv) doesn't yield a definite verdict on the

3.3.2.1. Applying (i)-(iv) to the Case of 'true'

As we have seen, (Lewis 1970) advocated something like (i) as, in the first instance, a characterization of reference for "theoretical terms"—terms that are introduced into a language by a scientific theory. This is also made clear by the example from (Lewis 1972), in which the terms 'X', 'Y', and 'Z' are introduced by the detective's theory about the crime. That said, Lewis later extended his view so as to apply to certain folk terms, such as the names for colors.[32] Even so, the question naturally arises whether Lewis' views about reference, and by that token, (i)-(iv), apply to 'true'.

One thing that makes this concern especially pressing is the widespread belief that "descriptivist" theories of reference such as the Lewisian one I have presented apply only to some kinds of expressions, if any. (For concreteness, let's say a theory of reference for expressions of a given kind (or occurrences thereof) is *descriptivist* if it holds that the referent of an expression of that kind (or occurrence thereof) is that entity, if there is one, which most closely fits (that is, respects) a certain description. If 'description' in the relevant sense also applies to inference rules, then it's clear that the conception of reference that I have advocated for the word 'true' and its occurrences is descriptivist in this sense.) In particular, for example, while descriptivism about proper names has historically had some supporters, at the time of writing the mainstream consensus is that this view is false.[33]

---

reference of either. To get the desired conclusion (namely, that E_0 but not E_1 fails to refer to truth), something needs to be added to (iv). For the sake of brevity, I will leave that for another occasion, and simply assume that it is reference by E_0 that needs to change.

[32] See (Lewis 1997).

[33] For supporters of descriptivism, see (Frege 1892), (Russell 1905), and (Searle 1958). For problems with the descriptivist account of proper names, see (Marcus 1947) and (Kripke 1980). For a sketch of a popular alternative account, see (Kripke 1972), beginning p.91. For yet a further alternative, see (Graff Fara 2015), which argues that names are predicates of a certain distinctive kind.

Of course, the case of proper names need not be an immediate concern, since 'true' is not a proper name but rather (on most accounts) a predicate. Moreover, the most prominent non-descriptivist accounts of reference for proper names sound implausible when applied to 'true'. For instance, on the view sketched in (Kripke 1980), a name acquires its referent in what he calls an "initial baptism," an event in which the name is first deliberately applied to an object that is identified in some other way, e.g., via ostension. Reference to that object by that name is then sustained by causal relations that relate that initial event to subsequent uses of the name. Kripke's account sounds most plausible as applied to names for objects that can be readily identified, e.g., by ostension, not for unobserved entities that are introduced by scientific theories or abstract properties like truth. Thus, causal theories of reference à la Kripke pose less of a threat to my claims about 'true' than one might have thought.

A consideration that positively supports the idea of (T-intro) and (T-elim) being a reference-determining theory is the centrality of these rules in governing our uses of 'true'. The use of 'true' in accordance with these rules is deeply entrenched and widely accepted, more so than any other patterns of use for it. Indeed, if someone failed to find these rules generally compelling, that would raise *prima facie* doubts about her grasp of the word. And it is reasonable to suspect that the uses of an expression that are most deeply entrenched and widely accepted play the most significant roles in determining the expression's reference.

A closely related idea, defended in (Eklund 2002) and (Scharp 2013) Chapter 2, is that (T-intro) and (T-elim) are *meaning-constitutive* for 'true'.[34] That is, using 'true' to mean what it means in English consists in standing in some special, receptive cognitive relation to (T-intro) and (T-elim). For instance, one might hold that using 'true' to mean what it means in English

[34] For further discussion of related issues, see (Eklund 2002) and (Scharp 2013) pp.43-56 and p.62.

consists in being disposed to accept these inference rules. On another variant of the view, using 'true' to mean what it means in English consists in finding these inference rules *primitively compelling*—finding them compelling, and not in virtue of finding anything else compelling.[35] Whatever in the end meaning-constitutivity amounts to, one who likes the idea can identify the reference-determining principles for an expression as those sentences or inference rules that are meaning-constitutive for it.

I find the idea of meaning-constitutivity appealing, and I am sympathetic with the idea that if an expression has principles that are meaning-constitutive for it, then these principles play a more significant role in determining its reference than any other sentences or inferences in which it features do. However, since the notion of meaning-constitutivity is controversial, I emphasize that I do not obviously need to invoke it here. What I need is for 'true' to have its referent be determined by (T-intro) and (T-elim) in the manner described by (i)-(iv). I see no reason why this claim cannot be based entirely on the centrality of these inference rules in governing our practices with 'true', leaving open the question of meaning-constitutivity.

Before laying out my preferred view in further detail, I should stress a few important points concerning the flexibility of moderate aberrationism. Readers should keep in mind that moderate aberrationism is defined only by the acceptance of (Semantic Supervenience), and (Semantic Regularity) (along, of course, with (Aberrations) and (Determined)). It need not be committed to the Lewisian views about reference that I have just laid out. Moreover, even if it is committed to these Lewisian views as applied to some terms, it need not be committed to their application to 'true'.[36] Likewise, even fans of moderate aberrationism who endorse my

[35] The idea of someone's finding something primitively compelling is due to (Peacocke 1992), Chapter 2.
[36] That said, to remain faithful to (Semantic Regularity), moderate aberrationists who do not apply the Lewisian views to 'true' will need some other principled account of what determines the reference of the key occurrences of alethic expressions in Liar-like sentences. They will need this in order to remain faithful to (Semantic Regularity).

application of the Lewisian views to 'true' might disagree with me about which properties come closest to respecting (T-intro) and (T-elim). One can see, then, that within moderate aberrationism there is room for a great variety of positions. Again, my only goal here (Section 3.3) is to develop a sample moderate view, one that can compellingly illustrate the advantages available in principle to moderate aberrationism. For all I say here, other moderates may have different ways to secure these advantages, and still others may not be able to secure them. Much rests on the details of the particular views.

3.3.3. The Reference of the Occurrence of 'true' in A

We will see below that some of the advantages of moderate aberrationism over its radical cousins rest on moderates' views about of what Russellian propositions Liar-like sentences express the negation, the conjunction, disjunction, etc., and thus in turn on their answers to questions such as, 'To what (if anything) does the occurrence of 'true' in A refer?'. In this subsection, I will develop my own preferred answer: such occurrences are indeterminate in reference as between two quite truth-like properties, *ascending truth* and *descending truth*, which are described in (Scharp 2013). Neither property respects both (T-intro) and (T-elim), but each comes equally close to doing so, and both properties come closer than anything else (besides truth itself). In that sense, ascending truth and descending truth are "tied for second place" when it comes to respecting (T-intro) and (T-elim). It then follows from the Lewisian views about reference from Section 3.3.2—that is, from (i)-(iv)—that when an occurrence of 'true' in a Liar-like sentence cannot refer to truth, its reference is indeterminate as between ascending truth and descending truth. Given this, it then follows from (Determined) that all Liar sentences are indeterminate in content: for any Liar sentence S, it is indeterminate of which of the propositions

<S, *ascending truth*>, and <S, *descending truth*> S expresses the negation. Similar remarks apply to all other Liar-like sentences. (The role of (Determined) here is to get us from indeterminacy in the reference of the occurrence of 'true' to indeterminacy as to what Russellian proposition is such that its negation is expressed.)

The view just described serves as an illustration of how a moderate aberrationist might reasonably answer the question of what Liar-like sentences say. In Section 3.3.4, I will explain why this sort of answer is better than those available to radical aberrationism. Before that, however, I will say some things to make my illustration more concrete and plausible. I will describe ascending truth and descending truth in some further detail, and justify some of my claims about these properties. This will also enable me in Section 3.3.3.2 to make good on my promise to show that truth is the unique property which comes closest to respecting (T-intro) and (T-elim).

3.3.3.1. Ascending Truth and Descending Truth

In (Scharp 2013), Kevin Scharp introduces two predicates, 'ascending true' and 'descending true'. These two predicates are defined by a list of 20 axiom schemata in which they feature.[37] He calls the set of all instances of these axiom-schemata *ADT*—"the theory of ascending and descending truth." In particular, two of the axiom-schemata that help to define these predicates are closely related to the rules (T-intro) and (T-Elim):

(A1) If S then 'S' is ascending true

(D1) If 'S' is descending true then S

[37] See his p.154 for a full list.

Here, setting aside issues of context-sensitivity, the 19th letter of the uppercase Latin alphabet ranges over declarative sentences of English.

Importantly, however, no instance of either (A1#) or (D1#) below is in ADT:

(A1#) If ‘S’ is ascending true then S

(D1#) If S then ‘S’ is descending true

In practical terms, the inclusion of (A1) and (D1) and exclusion of (A1#) and (D1#) means that speakers can always use ‘ascending true’ in accordance with (T-intro) but not always (T-elim). Similarly, they can always use ‘descending true’ in accordance with (T-elim), but not always (T-intro). So, the idea is, instead of having a single word, ‘true’, that we pre-theoretically take to obey both rules, we now have two different words, each of which we treat as obeying one rule but not always the other.

It is worth noting that similar moves can be made with ‘satisfies’ in place of ‘true’. That is, just as Scharp proposes to divide the principles for ‘true’ between two separate predicates, an analogous move can be made with ‘satisfies’, giving rise to two predicates, ‘ascending satisfies’ and ‘descending satisfies’, each of which can be applied, for any n, to an n-tuple of objects and an n-ary formula. Thus, Scharp’s ideas, and my appropriation of them, can be extended to Liar-like sentences which speak of satisfaction rather than truth.

So far, I have only been describing Scharp’s predicates. But the claims that I want to make concern the properties to which these predicates purport to refer. (See below for more on whether there are any such properties.) Again, those claims are as follows:

1. There are two properties, *ascending truth* and *descending truth*, to which the predicates ‘ascending true’ and ‘descending true’ respectively refer.
2. Ascending truth respects (T-intro) but not (T-elim), and descending truth respects (T-elim) but not (T-intro).
3. Ascending truth and descending truth both come close to respecting (T-intro) and (T-elim).

4. In fact, they come equally close.
5. Nothing else, besides truth itself, comes closer to respecting these rules.

As we saw earlier, given (1)-(5) and the Lewisian views about reference (i)-(iv) from Section 3.3.2, any key occurrence of 'true' in any Liar sentence is indeterminate in reference as between ascending truth and descending truth, and likewise for all other Liar-like sentences. I will now make a few remarks on (1)-(5).

In his book, Scharp proves the consistency of ADT relative to set theory by constructing a set-theoretic model, M_2, of ADT.[38] Assuming that set theory is consistent, a proof that an expression defined by a list of axioms has an extension in a set theoretic model serves as evidence that if the expression were added to English, defined by those axioms, it would refer. At least, this is how philosophers routinely reason. In particular, then, the fact that 'ascending true' and 'descending true' have extensions when interpreted in M_2 serves as evidence that they refer when added to English and defined by the adoption of the elements of ADT as axioms. Moreover, Scharp is careful not to impose any expressive limitations on the language of M_2 that might be relevant to its extensibility to a full natural language such as English. Indeed, he objects to other theorists' failure to take such precautions.[39] In light of these observations, I propose to accept (1) on the basis of the evidence Scharp provides.

Quite similar is the matter of thesis (2), concerning the non-satisfaction of (T-intro) and (T-elim). Scharp claims to have proven that all instances of (A1) and (D1) are true-relative-to-M_2, whereas only some instances of (A1#) and (D1#) are true-relative-to-M_2.[40] That serves as evidence that ascending truth, if it exists, respects all instances of (A1) and (T-intro), but only

[38] See his Section 6.6 (pp.157-169) and an appendix (p.178) to his Chapter 6.
[39] See p.156.
[40] See his Section 6.A.4, beginning bottom p.186. He does not give a full proof, but rather provides only the proof for (D1) in a footnote.

some instances of (D1#) and (T-elim). Likewise, we have similar evidence that descending truth, if it exists, respects all instances of (D1) and (T-elim), but only some instances of (A1#) and (T-intro). In particular, we will soon see that α and δ below are counterexamples to (A1#) and (D1#), respectively:

(α) α is not ascending true

(δ) δ is not descending true

More generally, I will soon (Section 3.3.4.3) show that any sentence S which says of itself that it is not ascending true is (contrary to what it says) ascending true, and therefore false. Similarly, any sentence S which says of itself that it is not descending true is not descending true, and is thus true. These observations help give a better grip on ascending and descending truth.

Let us now discuss (3), the claim that ascending truth and descending truth come close to respecting both (T-intro) and (T-elim). As a first thing to note, both properties respect all instances of each rule which involve sentences that are "safe" (p.186).[41] A sentence is *safe* if it is either descending true or not ascending true. Respect of both rules when it comes to safe sentences is significant, since the safe sentences include most sentences that anyone would ever want to use. In particular, Scharp points out, "every sentence that is grounded (in something like Kripke's sense) is safe" (p.170). A sentence is *grounded in Scharp's sense* if "its ascending truth value and descending truth value are completely determined by the ascending truth values and descending truth values of sentences that have no occurrences of 'ascending true' or 'descending true'." All sentences that are grounded in Scharp's sense are safe. Moreover, "many ungrounded sentences, like 'every sentence is either ascending true or not ascending true' and 'no sentence is both descending true and not descending true' are safe." So, the range of sentences for which

[41] Scharp proves that all sentences which are safe have the same ascending truth and descending truth values, and that (A1#) and (D1#) hold for all such sentences.

ascending truth and descending truth respect both rules is significant; it includes most sentences that anyone would normally want to use, and more besides.

Scharp himself sums up the situation thus:

> only sentences that contain 'ascending true', 'descending true', or 'safe' might turn out to be unsafe and, even among those, only sentences that would be paradoxical if 'true' were substituted in for these terms might be unsafe. (p.186)

It is worth flagging the second conjunct here. This is a substantial claim that needs careful justification, which Scharp does not provide in his book and which I am inclined to doubt. We will return to this matter later, in the discussion of *wholesale indeterminism* in Section 9. For present purposes, however, even if Scharp's definition of safety counts a few pre-theoretically unproblematic sentences as unsafe, and even if ascending truth and descending truth fail to respect the instances of (T-intro) and (T-elim) involving these sentences, that need not prevent these properties from coming close to respecting both rules. For in this case, non-respect would still be restricted to a small range of sentences. As I'll explain in the next section, near-respect of both rules is sufficient to render ascending truth and descending truth very truth-like. This is important, since it allows Liar-like sentences to come very close to saying what they appear to say.

Now for (4), the claim that both properties come equally close to respecting both rules. Here I will be brief. The claim is plausible from the outset, given the symmetry of the axioms that define these properties. Moreover, we just saw that each property respects all instances of one rule, and all the safe instances of the other. That is enough for present purposes, since the goal is merely to give an illustration. I leave to another occasion the task of exhaustively verifying that each property comes equally close.

Finally, we come to (5), the claim that nothing (besides truth) comes closer to respecting both (T-intro) and (T-elim). While I find this claim quite plausible, I need not defend it here.[42] For even if something(s) else besides truth comes closer, that does not threaten the point I am trying to make in favor of moderate aberrationism. In that case, advocates of such views can still retain my Lewisian views about reference, and so claim that what Liar-like sentences say is powerfully influenced by how their component expressions are used, in particular by our use of 'true' in accordance with (T-intro) and (T-elim). Moreover, they can still claim that Liar-like sentences come close to saying what they appear to say. The reason is that nothing could come close to respecting (T-intro) and (T-elim) without thereby being quite similar to truth. Respect of (T-intro) and (T-elim) is the hallmark feature of truth; it is arguably the feature that contributes most significantly to making truth a useful property to attribute.[43] So, any property that comes close to respecting it is similar to truth in the respect that matters most to us.

3.3.3.2. Aberrations and Respect

Up to this point, the reader may reasonably have been wondering: does truth even respect (T-intro) and (T-elim)? If not, does it come closer than ascending and descending truth? These

[42] Scharp repeatedly emphasizes that ascending and descending truth are to be embraced as a team. Matti Eklund raises the question whether, *even when taken separately*, these properties still tie for second-closest when it comes to respecting (T-intro) and (T-elim). In response, it helps to note that all other truth-like properties that have been described fail to respect some instances of either (T-intro) or (T-elim) which involve sentences that we would pre-theoretically judge to be non-paradoxical. (For instance, truth_{mp} fails to respect the instance of (T-intro) involving the sentence 'If this sentence is true then this sentence is true'.) Because identifying all and only the paradoxical sentences is notoriously difficult, *a fortiori*, targeting all and only these sentences for failures of (T-intro) or (T-elim) is a daunting challenge. In the above discussion of my claim (3), we had Scharp claiming that his properties succeed in meeting this challenge: ascending truth fails to respect (T-elim), and descending truth fails to respect (T-intro), on all and only the sentences that would be paradoxical if they contained 'true' rather than 'ascending true' or 'descending true'. If indeed Scharp is right about this, then it is plausible that his properties really do tie for second place for respecting (T-intro) and (T-elim), even when taken separately; for (prior to (Scharp 2013), we are assuming,) respecting the rules on all pre-theoretically non-paradoxical sentences has yet to be achieved. That said, Scharp does not provide a proof, and given the difficulty of the challenge, there is room for skepticism.

[43] See (Quine 1970) pp.10-12 for more on this.

questions matter, since if it does not come closer, then that is very bad for the Lewis-inspired version of moderate aberrationism that was laid out in the previous subsections—for then by the Lewisian theory (i)-(iv) no occurrence of 'true' refers to truth. Rather, all occurrences of 'true' are indeterminate between ascending and descending truth. (That is, assuming there's no yet further property that comes closer than these do to respecting (T-intro) and (T-elim). See below for more on this question).

Since non-Liar-like sentences pose no obstacle to respect, whether truth respects (T-intro) and (T-elim) depends entirely on whether the instances of these rules that involve Liar sentences would be truth-preserving, if the occurrences of 'true' therein were to refer to truth. But as I'll now explain, according to any aberrationist, no such instances are possible. For according to any such approach, no key occurrence of an alethic expression in a Liar-like sentence can co-refer with that expression, on pain of contradiction.[44] Now, I claim, in assessing whether truth comes closer to respecting (T-intro) and (T-elim) than ascending and descending truth do, we should surely count only the genuinely possible instances of these inference rules.[45] If we make our assessment in this way, then truth clearly wins out. Whereas we will soon see (in the next subsection) that the sentences α and δ from Section 3.3.3.1 prevent ascending truth and descending truth, respectively, from respecting both (T-intro) and (T-elim), sentence A, failing as it does to express the negation of the Russellian proposition <A, truth>, does not prevent truth from respecting both (T-intro) and (T-elim). So, by (i)-(iii) the word 'true' refers to truth.

[44] As explained at the end of Section 2, the full justification for this claim will come only once I have shown that this is the best solution out of all of the alternatives.

[45] Here I legislate away a potential complication. The claim that truth satisfies (T-intro) and (T-elim) is a hypothesis about impossible things, analogous to the claim that if there were round squares, then there would be a formula for determining their area. If such claims are all false, then it is false that truth satisfies (T-intro) and (T-elim), and so it may be that by some measure of closeness either ascending truth or d-truth comes closer. Alternatively, if such claims are all indeterminate, then it is indeterminate whether truth satisfies (T-intro) and (T-elim). In the latter case the question of which property—truth, ascending truth, or d-truth—comes closest is hard to settle.

At this point I am in a position to claim that while ascending and descending truth come close to respecting (T-intro) and (T-elim), they do not come as close as truth itself. Thus, if indeed ascending truth and descending truth are "tied for second place," then by (iv) the occurrence of 'true' in any Liar sentence is indeterminate in reference, with ascending and descending truth serving as candidate referents. (More generally, for any *e* and $y_1, \ldots, y_n$, if it is indeterminate whether *e* refers to y_1, to y_2, …, or to y_n, let's say that $y_1, \ldots, y_n$ are *candidate referents for e*.) The fact that all these candidate referents are quite truth-like gives us a respectable sense in which Liar-like sentences come quite close to saying what they appear to say.

3.3.3.3. The Status of A and Other Liar-like Sentences

Suppose I am right that the occurrence of 'true' in A is indeterminate in reference as between ascending and descending truth. Then, I have claimed, A is indeterminate in content: it is indeterminate of which Russellian proposition A expresses the negation. But, one wonders, is it therefore indeterminate what truth value A has? I will now consider two prominent views concerning the truth values of sentences that contain indeterminate terms, and show that both views yield the verdict that A is indeterminate as to its truth value.

On one available view, any sentence that contains some indeterminate terms is indeterminate in truth value.[46] If indeed the occurrence of 'true' in A is indeterminate in reference, then this view has the immediate consequence that A is indeterminate in truth value. There is a slight complication here given that we are dealing with indeterminacy in an

[46] Kleene's "weak" truth tables assign values in this way. This is the right way to assign truth values if one conceives of indeterminacy as being akin to incoherence. The idea is that the result of using a logical connective to combine a sentence with something incoherent is itself incoherent.

occurrence rather than in a word *simpliciter*, but we are already assuming that the semantic status of the sentence is determined by that of its occurrences. (Keep in mind: here we are figuring out the consequences of a given moderate aberrationist view.)

On a different view about indeterminacy, a sentence containing exactly one indeterminate term is true (false) if, for every candidate referent for that term, if the term referred to that candidate referent, the sentence would be true (false). If this holds for some candidates but not others, then the sentence is indeterminate in truth value. (This view can be easily recast in terms of occurrences rather than terms *simpliciter*.) Now our question is, which description fits A?

Start by supposing that the occurrence of 'true' in A refers to ascending truth, so that A expresses the negation of the Russellian proposition <A, ascending truth>. We will now figure out the status of A in that case. Consider first the sentence α:

(α) α is not ascending true

Scharp explains that from the construction of α and the fact that 'ascending true' respects (T-intro), we can prove both 'α is ascending true' and ''α is not ascending true' is ascending true'.[47]. On pain of contradiction, then, we cannot infer from ''α is not ascending true' is ascending true' to 'α is not ascending true'—for the latter would then contradict 'α is ascending true'.[48] Returning to the case of A, we can mimic the argument just given. Assuming that the occurrence of 'true' in A refers to ascending truth, from the construction of A we can prove both 'A is true'

[47] See his p.151. He writes: "we can prove A(a) and A(⌜~a⌝)." To keep things as informal and simple as possible, I am writing 'It's not the case that α is not ascending true' rather than ⌜~α⌝ and then substituting 'α is ascending true' for 'It's not the case that α is not ascending true'.

[48] Here is the proof:

1.	α = 'α is not ascending true'	definition of 'α'
2.	α is ascending true	proven elsewhere
3.	'α is not ascending true' is ascending true	proven elsewhere
4.	If 'α is not ascending true' is ascending true then α is not ascending true	Instance of (A1#)
5.	α is not ascending true	*modus ponens*, (3), (4)
6.	Contradiction	Contradiction Introduction, (2), (5)

and ‘‘A is not true’ is true’. And then on pain of contradiction, ascending truth cannot respect the following instance of (T-elim):

‘A is not true’ is true (T-elim)

A is not true

Now, it turns out that this failure of (T-elim) gives us some useful information about the status of A. To see the point, note the following fact about respect. When a property P fails to respect an instance of (T-elim) involving a sentence S, that is because S has P (and so ⌜‘S’ is true⌝ is true, if the occurrence of ‘true’ therein refers to P), but S is not true. (Thus the inference from ⌜‘S’ is true⌝ to S is not truth-preserving.) Therefore, given that ascending truth fails to respect the instance of (T-elim) that takes us from ‘‘A is not true’ is true’ to ‘A is not true’, we can conclude that A is ascending true, but not true. (Again, here we are assuming that the occurrence of ‘true’ in A refers to ascending truth.)

The situation is quite symmetrical when we consider descending truth. Suppose that the occurrence of ‘true’ in A refers to descending truth, so that A expresses the negation of the Russellian proposition <A, descending truth>. To see the status of A in that case, consider the sentence δ:

(δ) δ is not descending true

Similarly to what we had with α, from the construction of δ and the fact that ‘descending true’ respects (T-elim), we can prove both ‘δ is not descending true’ and ‘‘δ is not descending true’ is not descending true’.[49] On pain of contradiction, then, we cannot infer from ‘δ is not ascending true’ to ‘‘δ is not ascending true’ is descending true’. Returning to the case of A, we can mimic the argument just given. Given that the occurrence of ‘true’ in A refers to descending truth, from

[49] See p.151.

the construction of A we can prove both ‘A is not true’ and ‘‘A is not true’ is not true’. And then on pain of contradiction, ascending truth cannot respect the instance of (T-intro) that takes us from ‘A is not true’ to ‘‘A is not true’ is true’.

Now note the following fact about respect. When a property P fails to respect an instance of (T-intro) involving a sentence S, that is because S is true, but S does not have P. (Thus the inference from S to ⌜‘S’ is true⌝ is not truth-preserving, if the occurrence of ‘true’ in the latter refers to P. In that case, the premise is true but the conclusion is not.) So, given that descending truth fails to respect the instance of (T-intro) that takes us from ‘A is not true’ to ‘‘A is not true’ is true’, we can conclude that A is true, but not descending true.

Summing up, if the occurrence of ‘true’ in A referred to ascending truth then A would be false, whereas if the occurrence of ‘true’ in A referred to descending truth then A would be true. Given that ascending truth and descending truth are candidate referents for the occurrence of ‘true’ in A, that is enough to show that A is indeterminate, on the second view about how lexical indeterminacy relates to truth values of sentences that I described. Since both such views yield the verdict that A is indeterminate, in what follows I will assume that A is indeterminate. The same goes for other Liar-like sentences.

3.3.4. The Advantages of Moderate Over Radical Aberrationism:

At this point, I have introduced my own preferred version of moderate aberrationism. According to it, Liar-like sentences fail to say what they appear to say, because the key occurrences of their alethic expressions fail to co-refer with those expressions. Rather, the aberrant occurrences of ‘true’ are indeterminate in reference as between ascending and descending truth; and similarly when it comes to other alethic expressions. Because of this, the

reasoning involved in Liar-like paradoxes does not go through. A crucial feature of this view is that it respects Semantic Regularity: whether we are dealing with ordinary theoretical terms or with the occurrences that feature centrally in generating a paradox, reference-determination proceeds according to the same Lewisian pattern. That is something one might reasonably call a "perfectly reliable, principled relationship" between facts about language use and facts about reference.

Even given all this, however, moderate views might not initially seem much better than radical ones. Both must, for example, deny, contrary to appearances, that A expresses the negation of the Russellian proposition <A, truth>. In denying this, both views depart from most people's initial, intuitive reactions to A. However, it should be clear that the radical views depart for a much more distant destination. The point is especially clear in the case of views that reject Semantic Supervenience. On such views, again, what A says needn't have anything to do with how we use or are disposed to use any of A's constituent expressions. As Smith puts it, what (if any) proposition A expresses is determined quite randomly, and similarly for its occurrences of 'A' and 'true'. Thus, for all we know, the occurrence of 'A' might refer to North Korea and the occurrence of 'true' to *being made of goat cheese*, so that A would say that North Korea is not made of goat cheese. One who denies Semantic Supervenience has no way to rule this sort of thing out.

In stark contrast, on my view and other moderate views like it, while A does not express the negation of the Russellian proposition <A, truth> or say of itself that it is not true, it nonetheless comes fairly close. On my own preferred view, in particular, A is indeterminate as between expressing the negation of the proposition <A, ascending truth> and that of <A, descending truth>. Likewise, it is indeterminate as between denying that it (namely, A) is

ascending true and denying that it is descending true. While neither ascending truth nor descending truth is identical with truth, each differs from truth only with respect to a quite narrow range of sentences. Recall, for example, that for all sentences S which are grounded in Scharp's sense, S's ascending truth value equals its descending truth value, which equals its truth value. With this in mind, compare a sentence which denies that North Korea is made of goat cheese with a sentence that is indeterminate as between denying that it (that sentence) is ascending true and denying that it is descending true. Whatever closeness amounts to in the end, it is clear that the latter sentence comes much closer to saying of itself that it is not true. Thus, my view comes closer to allowing that A says what it pre-theoretically seems to say. And, I take it, short of accepting contradictions, the closer one can get to allowing that A says what it pre-theoretically seems to say, the better.

Now admittedly, the denier of Semantic Supervenience could propose that it is simply a brute fact that the occurrence of 'true' in A refers to some very truth-like property. However, without appealing to any principles that link reference to facts about use, it is hard to see what could justify such a proposal. In particular, the availability of my view shows that such a proposal is not necessary for solving the paradox, and so cannot be justified solely on that basis. The brute fact idea solves the paradox while respecting our pre-theoretical impressions about what Liar sentences say, but my own claims—that is, (i)-(iv)—have these same advantages while in addition respecting the widely accepted, deeply entrenched thesis of Semantic Supervenience.

Similar things can be said concerning a slightly less radical theorist who rejects Semantic Regularity but retains Semantic Supervenience. To respect our pre-theoretical impressions about what Liar sentences say, a less radical theorist needs to hold that the occurrence of 'true' in A refers to something that is similar to truth. As compared to her more radical counterpart who

denies Semantic Supervenience, the only additional resource that this less radical theorist has are the non-semantic facts on which the reference of this occurrence of 'true' supervenes. But precisely what the less radical theorist claims is that the supervenience of the semantic on the non-semantic is messy, in some way that rules out a "perfectly reliable, principled" relationship. So precisely what she cannot do is appeal to the existence of such a relationship to argue that it determines that the referent of the occurrence of 'true' in any Liar sentence is something similar to truth.[50] In a way, the less radical theorist is in an even worse position than one who denies Semantic Supervenience, since the less radical theorist cannot claim that the semantic fact she needs is simply a brute fact. She holds that the reference of the occurrence of 'true' in every Liar sentence is determined by the non-semantic facts, but in a way that by her own lights admits of no straightforward explanation.

Like her more radical counterpart, the less radical theorist can claim that her view is justified because it solves the paradox while respecting our pre-theoretical impressions about what Liar sentences say. But again, moderate versions of aberrationism have these same virtues, while in addition respecting the plausible thesis of Semantic Regularity. In fact, even (Smith 2006) concedes it is better to retain Semantic Regularity than to reject it, other things being equal. It is just that Smith thinks other things are not equal, and that rejecting Semantic Regularity is the least painful way to extricate ourselves from contradictions in the face of the

[50] However, Vivek Matthew suggests to me that one who denies Semantic Regularity could appeal to an imperfect-but-still-somewhat-reliable relationship between matters semantic and matters non-semantic, in order to show that the occurrences of 'true' in Liar-like sentences refer to something truth-like. I have little to say against such a view, given how similar it is to my own. As I pointed out in Section 3.2, at the risk of misleading, my own position can be described in these terms. My point here is simply that there is tension between the claim that the relationship between matters semantic and matters non-semantic is unreliable and "unprincipled" and the hope of using some non-semantic facts to explain the reference of the occurrences of 'true' in Liar sentences. I do not insist that there is no way to resolve this tension, though for my own part I cannot see how to do it.

Liar paradox. However, if moderate aberrationism is indeed a viable option then Smith is wrong on this.

So much, then, for internal debates between different versions of aberrationism. In Sections 4-10, I will set these disputes aside, and instead contrast aberrationism with various competing approaches. Although I will discuss a wide variety of views, the main goal will be to showcase the distinctive advantages of aberrationism, not to conduct a fully exhaustive survey of the competition. I will thus take the liberty of omitting views about which I have little original to say. For example, while I have serious objections to the views articulated in (Tarski 1935), (Tarski 1944), and (Priest 1987), these objections are already articulated in (Smith 2006), among many other places.

4. The No-Proposition View

A close cousin of aberrationism is *the no-proposition view*, which rests on the claim that Liar-like sentences fail to express propositions.[51] The motivating idea of the no-proposition view is that if Liar-like sentences fail to express propositions, then they fail to say of themselves that they are not true, and in that case, there is no genuine inconsistency. As the reader may have guessed, the no-proposition view is compatible with several versions of aberrationism. My main point in this section will be that the best ways to develop the no-proposition view are the ways most congenial to aberrationism.

The classic problem for the no-proposition view, discussed by (Parsons 1974), is that once the notion of a sentence's expressing a proposition (or not) has been introduced, it then

[51] This view is discussed in (Parsons 1974), and is advocated in (Goldstein 2009). See Chapter 2 Section 5.2 for discussion of some prominent objections to the no-proposition view from (Scharp 2013).

seems we can formulate a sentence which says of itself that it does not express any true proposition:

(L_N) L_N does not express any true proposition

Applying the no-proposition view to L_N, we get the result that L_N fails to express any proposition. It follows that in particular, L_N fails to express any true proposition. However, the objection goes, that would make L_N true after all.

The obvious response is that while the objector has purported to assume the no-proposition view and then derive a contradiction, in fact she has simply reneged on the commitments of the no-proposition view, illegitimately acquiescing to our pre-theoretical impressions about what L_N says. But, the response stresses, if indeed L_N fails to express any proposition, then in particular—and despite whatever appearances suggest—L_N does not say of itself that it fails to express any true proposition. So, L_N's failure to express any true proposition does not render L_N true after all.[52]

Even if this response is good, however, advocates of the no-proposition view face a challenge. There are uncontroversial compositional rules that specify how the reference of a sentence's constituents, plus the way these constituents are arranged, combines to determine which proposition the sentence expresses. Let *general compositionality* be the claim that these rules generally apply to sentences in natural languages.[53] Given L_N's syntax and lexical constituents, and given our default views about the reference of its constituents, general compositionality gives us good reasons to believe that L_N says of itself that it is not true. Therefore, advocates of the no-proposition view must either reject our default views about the

[52] Compare the remarks on (Goldstein 2009) p.385.
[53] See Section 11 for some independent evidence that these rules do not apply to absolutely every sentence.

reference of some of L_N's constituents, argue that L_N constitutes an exception to general compositionality, or reject general compositionality.

The last of these three positions can be rejected out of hand, as it flies in the face of the myriad successes of compositional semantics. Turn, then, to the proposal to give up our pre-theoretical default views about the reference of some of L_N's constituents. E.g., one could claim that the word 'L_N' *simpliciter* fails to refer to anything, that 'express' fails to refer to anything, or that the word 'true'—again, the word *simpliciter*, not merely its occurrence in L_N—fails to refer to anything. If any of these claims were true, then that would arguably leave L_N without a proposition to express, since it is plausible that in general a sentence that contains a non-referring expression fails to express any proposition.

However, even if these diagnoses can succeed in staving off contradictions, they all posit implausible restrictions on the expressiveness of natural languages. For example, we can use the name 'L_N' in plenty of seemingly coherent, non-paradoxical sentences, such as 'L_N is a sentence of English'. This coherence would be hard to explain if the word 'L_N' *simpliciter* failed to refer as expected. Likewise, we often speak quite coherently of sentences' expressing propositions; this would be hard to explain if the word 'express' *simpliciter* failed to refer to the relation *expressing*. And lastly, in much of our talk involving 'true', we seem perfectly able to express definite propositions. That would be hard to explain if the word 'true' *simpliciter* failed to refer.[54]

In contrast to the extreme view just described, a more plausible way to develop the no-proposition view is more in the style of aberrationism. Instead of claiming that L_N's constituent

[54] (Brandom 1994) argues that 'true' is not a referring expression such as a normal predicate but instead something rather like a pronoun. Even if Brandom can explain the coherence in our use of 'true' without claiming that this word refers *simpliciter*, the view on offer here is one that takes 'true' to be a predicate that fails to refer. It is hard to see how this view can explain our ability to express propositions using the predicate 'is true'.

words fail to refer *simpliciter*, one can embrace the aberrationist idea that it is merely the occurrences of some of these expressions in L_N which fails to refer. Call this version of the no-proposition view *the occurrence theory*. As I explain in Section 10, I think the right expression to target here is 'true' rather than, for example, 'L_N' or 'expresses'; and any occurrence theorist who heeds this advice and focuses on 'true' (and generally, on alethic terms,) will, indeed, qualify as an aberrationist. She could qualify as either a moderate or a radical aberrationist, depending whether she claims (moderate) or denies (radical) that the aberrant occurrence's failure to refer can be explained by a theory of reference.[55]

Whatever expression an occurrence theorist chooses to target, her view has one of the distinctive advantages of aberrationism: it leaves unperturbed all the occurrences of the targeted expression in non-paradoxical sentences, and thus accords with the most plausible accounts of these sentences that we are afforded by casual inspection and by linguistics. E.g., instead of having to say, quite implausibly, that no sentence containing the name 'L_N' expresses a proposition, one can make this claim only about Liar-like sentences. This is all to the good, since it is the Liar paradox and not quite-general reflections on language that motivates the no-proposition view in the first place.

I just contrasted the occurrence theory with the approach that gives up our default views about L_N's constituents. That places it in the category of views that beg an exception to general compositionality when it comes to Liar-like sentences. And it is easy to see that, like aberrationism, the occurrence theory does this. When it comes to accounting for what

[55] Note, however, that the occurrence theory is inconsistent with my own favored version of moderate aberrationism, which claims not that aberrant occurrences of 'true' fail to refer whatsoever, but rather that their reference is indeterminate. Thus, I think it is indeterminate which proposition a Liar-like sentence expresses (see Section 3.3.3); but that is different from the claim that these sentences fail whatsoever to express propositions, as on the no-proposition view.

proposition a sentence expresses, general compositionality (as formulated above) looks to the reference of the sentence's constituent words *simpliciter* rather than to their occurrences; and here we have views that look to the reference of the occurrences. That said, however, occurrence theorists and aberrationists can still endorse general compositionality as a true generalization with some exceptions. Moreover, the position of aberrationists and occurrence theorists is from the beginning quite congenial to the spirit of general compositionality, if not the letter: they can claim that which proposition (if any) a sentence expresses is determined by the reference of the occurrences of its constituent expressions, plus the way they are combined. So, general compositionality need be rejected only lightly, if at all.

5. Aberrationism vs. Saul Kripke

Let us say that a diagnosis of the Liar is *incomplete* if it lacks the resources to account for all the semantic phenomena whose existence would follow from the diagnosis, were the diagnosis correct. In particular, a diagnosis of Liar sentences that posits a property P that can be had by sentences in any language must be able to account for reference to this property in the language in which the diagnosis is stated. If it cannot account for this, then it is incomplete. In this section, I will argue that the view defended in (Kripke 1975) is incomplete in this way. The same problem applies to more recent and elaborate views in the general family of (Kripke 1975). By contrast, aberrationism is not incomplete.

In (Kripke 1975) Saul Kripke defines a property of *groundedness* (recall Section 3.3.1 and see below), and argues that Liar sentences are ungrounded. Kripke develops a mathematical model that partitions the sentences of the object language—the language being modeled—into three disjoint and exhaustive sets: the set of true sentences, the set of false sentences, and the set

of ungrounded sentences. Roughly, a sentence S is *grounded* if, given the truth values of all the sentences that contain no uses of 'true', one can determine the truth value of S in finitely many steps. The point of introducing groundedness is that it gives us a way to diagnose Liar-like sentences: they turn out to be ungrounded.

Kripke's model is good as far as it goes, but it requires that the object language and metalanguage be distinct. For it is a time-honored observation that if English contained a predicate, 'grounded', that referred to the property of being a sentence of English that is grounded in Kripke's sense, then consideration of the sentence K below would lead quickly to contradictions:[56]

(K) K is either false or ungrounded.

Thus, on pain of contradiction, Kripke's model cannot be used to model any fragment of English that contains a word that refers to the property of being an ungrounded sentence of English. Since English contains such a word (thanks to Kripke's own work!), his view cannot straightforwardly be applied to English.[57] Of course, one can get out of this situation by claiming that while the word 'ungrounded' refers to ungroundedness, nonetheless its occurrences in Liar-like sentences such as K fail to do so. But clearly that claim would amount to a version of aberrationism.[58]

Now that I have accused (Kripke 1975) of incompleteness, the question arises whether aberrationist approaches to the Liar-like paradoxes are themselves incomplete. By the above remarks about incompleteness, any account of Liar phenomena which entails that there are propositions about the semantics of English that cannot be expressed in English would count as

[56] By now this is a time-honored observation.
[57] Kripke himself makes this observation about his own view.
[58] Compare similar remarks in (Smith 2006) pp.190-191.

incomplete. And *prima facie*, this is a concern one might reasonably have about views that endorse (Aberrations) and (Determined). In expressing semantic propositions about English, we often use the word 'true'. For instance, consider sentence *s* below:

(*s*) For any thing *x* and name *n*, if *n* is a name for *x* in English then ⌜*n* is a dog⌝ is true if and only if *x* is a dog.

However, if not all occurrences of the word 'true' refer to truth, then 'true' cannot always be used to state semantic propositions in this way. Notice, for instance, that if the occurrence of 'true' in *s* refers to the property *being made of goat cheese*, then *s* fails to express any semantic proposition. Rather, *s* says (falsely) that for any thing *x* and name *n*, if *n* is a name for *x* in English then ⌜*n* is a dog⌝ is made of goat cheese if and only if *x* is a dog.

An immediate response to this worry is that: (a) most sentences which we take to express semantic propositions, and *s* in particular, are not Liar-like; therefore, the occurrences of 'true' in these sentences refer to truth, as expected. Moreover, (b) the (semantic) things which Liar-like sentences cannot say about themselves can be said by other, non-Liar-like sentences. For instance, though A does not say of itself that it is not true, other sentences can surely say this about A. For instance, B does the job:

(B) 'A is not true' is not true.

(Throughout, I presuppose a scheme for individuating sentences on which A and B are distinct sentences. It helps to note that A and B contain distinct noun phrases: A contains 'A', whereas B contains ''A is not true''.) Note that, crucially, B is not a Liar sentence: B expresses the negation of the attribution of truth not to itself but rather to the (distinct) sentence A.

Now, a typical problem for approaches to the Liar paradox is that the apparatus used to state the semantic status of paradoxical sentences can be used to formulate a "revenge sentence,"

indescribable using that apparatus, which leads to contradictions.[59] (E.g., relative to the notion of groundedness, K is a revenge sentence.) And the introduction of B to articulate the semantic status of A raises an immediate question whether the method by which B was constructed—using quote marks—can be used to generate a revenge sentence. However, this cannot happen. The crucial point here is that all sentences have *finite depth*. That is, for any two syntactic locations in any sentence, the number of other syntactic locations between them is finite. Thus, on pain of having infinite depth, no sentence can have its own quote name as a lexical constituent. Now, the way that I constructed B was by using A's quote name. (By contrast, if I had simply used 'A', then B would have been identical with A, and therefore similarly defective.) So, to get a paradoxical sentence by using quote names instead of the usual non-meta-linguistic names (such as 'A'), one would have to construct a sentence that says of itself that it is not true, referring to itself by its own quote name. But on pain of having infinite depth, no sentence can contain its own quote name. So, constructing a paradoxical sentence by this method is impossible. Thus, the quote-name method for stating the defective status of Liar sentences is a safe one, and therefore it is possible to attribute that status to any such sentence without getting into any contradictions.[60]

[59] See Chapter 2 for a fuller discussion of revenge in connection with aberrationism.

[60] Still, one might reasonably be concerned about the fact that B can be converted into a Liar sentence by substituting the name 'A' for the co-referential quote name ''A is not true''. Should not B therefore count as Liar-like? And if indeed B is Liar-like, then doesn't its key occurrence of 'true' undergo an aberration, by aberrationists' lights? But then, if B witnesses such an aberration, then B fails to say of A that it is not true, as was intended. That failure would raise the question whether *any* sentence can say of A that it is not true. These reflections raise a deep question about what is the right way to define 'Liar-like sentence'. There is not enough space to explore this issue here; see Chapter 2 for a full discussion, including a definition of 'Liar-like sentence' which includes A but excludes B.

6. Contextualist Views

I have been arguing that the reference of an expression can depend on its component context. By contrast, many other approaches to the Liar paradox turn on the claim that some of the expressions implicated in the paradox are sensitive to contexts of utterance[61] or contexts of evaluation. Let us call any such approach a *contextualist* one. In this section, I will discuss several prominent contextualist views.

I'll begin with that of (Glanzberg 2004a), because on the one hand I have a particular, rather intricate objection to his view, and on the other hand he gives a plausible criticism of other contextualist views that I will echo later on. So, here is the plan for Section 6. In Sections 6.1 and 6.2 respectively, I'll present and criticize Glanzberg's view. Next, in Section 6.3 I'll present Glanzberg's criticism of other contextualist views. I'll show that whereas aberrationism is immune to this criticism, one can apply it with great force to Tyler Burge's view from (Burge 1979). Then in Section 6.4 I'll discuss how Burge might respond to Glanzberg's criticism. Finally, in Section 6.5 I'll discuss (Simmons 1993).

6.1. Glanzberg's View

In (Glanzberg 2004a), Michael Glanzberg offers a contextualist diagnosis of Liar sentences. However, rather than claiming that the truth-predicate is context-sensitive, Glanzberg argues that the logical form of each Liar sentence involves a tacit quantifier which ranges over propositions, and whose domain—despite initial appearances—is contextually sensitive. To arrive at this view, Glanzberg begins with the assumption that truth is fundamentally a property of propositions. Thus, he thinks, when someone says that any given sentence is not true, what she

[61] Throughout, I use 'context of utterance' in a way that is synonymous with 'discourse contexts' as used by linguists. I take this usage to be standard.

is really saying—or, as Glanzberg puts it, the logical form of the sentence she is uttering—is that there is no true proposition which is expressed by that sentence (presumably: in that context). Therefore, according to Glanzberg, any paradigmatic Liar sentence such as A above has a logical form closely analogous to that of *l*:

(*l*) It is not the case that there is some proposition p such that *l* expresses p and p is true.

Glanzberg then asserts that the only plausible source of context-dependence in *l* is in the existential quantifier.[62] (In what follows I'll have more to say about this assertion. For now, just note that he makes it.) On Glanzberg's view, one of the steps in the reasoning that moves from consideration of *l* to a contradiction causes a context-shift that changes the domain of this existential quantifier. With a few cosmetic changes, here is the reasoning that Glanzberg presents:

1. Suppose *l* expresses proposition *q*.
2. Suppose *q* is true.
    a) so, *l* is true.
    b) thus, there is no true proposition that *l* expresses.
    c) thus, *q* is false.
3. On the other hand, suppose *q* is false.
    a) so, *l* is false.
    b) then there is a true proposition, *p*, expressed by *l*.
    c) since *l* expresses at most one proposition, $p = q$.
    d) thus, *q* is true.
4. Thus *q* is true if and only if *q* is not true, which leads to contradiction.
5. Therefore, by *reductio ad absurdum*: there is no proposition which *l* expresses.
6. Therefore: there is no true proposition which *l* expresses.
7. Now we have proven *l*.
8. So, *l* is true.
9. But then *l* expresses a true proposition, contradiction.[63]

[62] See p.33 and p.34 final paragraph.
[63] See pp.33-34 for Glanzberg's presentation.

Glanzberg claims that between steps (5) and (6) there is a context shift, and that the domain of the existential quantifier is sensitive to this shift. Thus, he alleges, while the context-shift renders the reasoning invalid, both (5) and (6) are true: the context shift effected by (5) creates a context in which a certain proposition is newly available to be quantified over by the existential-quantifier-occurrence in (6).

Here is the view in greater detail. For Glanzberg, propositions are sets of truth-conditions. (Glanzberg is neutral on what exactly truth-conditions are. He suggests that they might be sets of possible worlds, "or whatever else we use to model individual truth-supporting circumstances" (p.35). It is worth noting the difference between the familiar view that propositions are sets of possible worlds and Glanzberg's suggestion that propositions are sets of *sets of* possible worlds.) His idea is then that for each context X, there is a "background domain" of truth-conditions which are available for forming the propositions that one can quantify over in X (p.35). Thus, cross-contextual shifts in this background domain induce shifts in the ranges of quantifiers that quantify over propositions. Glanzberg's proposal is that between steps (5) and (6) in the reasoning above, there is a shift in the background domain of truth-conditions, and that this shift adds a proposition to the range of the existential quantifier, making (6) true. In particular, he thinks, the mention of the expression relation (viz., *expressing*,) in (5) makes this relation contextually salient, and—in a way that is fairly intricate—this changes the background domain of truth-conditions.

More specifically, Glanzberg's idea is that which truth-conditions are in the background domain of a given context depends on which "expressive resources" are available for making distinctions among sets of possible worlds in that context (p.44).[64] In the above Liar reasoning,

[64] See pp.42-44 for this claim.

Glanzberg holds, making the expression relation salient at step (5) greatly increases the contextually available expressive resources, which in turn expands the background domain of truth-conditions.

The overall point is that, relative to the context thus created—the one in which (6) enters the picture—there is a proposition that *l* expresses, and it is false. Thus (6) is true.[65] (Glanzberg's account of precisely which proposition *l* expresses relative to the post-(5) context is quite technical and involved, but for present purposes it does not matter exactly what proposition that is.)

Now that I have presented Glanzberg's view, I'll articulate two criticisms of it, and then address his criticisms of views that, like aberrationism, target the truth predicate rather than posit a tacit quantifier.

6.1.1. Problems for Glanzberg

6.1.1.1. The Existential Quantifier

Glanzberg's approach rests on the assumption that the logical form of every Liar sentence includes an existential quantifier. While it is plausible that if propositions exist then sentences can be true only when they express propositions, it need not follow that the logical form of any sentence which attributes truth to a sentence involves an existential quantifier over propositions. In particular, the claim that propositions exist is, though plausible, metaphysically substantial. While common ways of speaking can turn out to come with tacit, substantial metaphysical commitments, still a philosopher who claims this about any particular way of speaking needs to support that claim with an argument. To be fair, Glanzberg is not alone in needing an argument

[65] See p.77 for Glanzberg's explanation of this point.

for the claim that attributions of truth to sentences involve existential quantification over propositions. Many other philosophers make that claim. But on the other hand, there are also plenty of philosophers who deny that propositions exist and are nonetheless happy to attribute truth to various sentences, without, of course, taking themselves thereby to be committed to any contradictions.[66]

Another *prima facie* concern about the claim that the logical form of truth-attributions involves a quantifier over propositions is that if the logical form of a sentence is something the grasp of which is necessary for understanding the sentence, then it is implausible that the logical form of every sentence that attributes truth to a sentence involves an existential quantifier over propositions. That is because people who do not know what propositions are can understand attributions of truth to sentences. Glanzberg might respond by suggesting that such people could in fact be said to have only *tacit* grasp of the logical forms of these sentences, where that grasp would consist in how these people would think and speak about attributions of truth to sentences once propositions were explained to them, or how they are disposed to defer to the assertions of people who do understand propositions. But, *prima facie*, not everyone has those dispositions. Some people, including the philosophers mentioned above, respond to learning about propositions by denying that propositions exist.

[66] Here I am thinking of Ted Sider, Vann McGee, Mark Balaguer, J.C. Beall, Bradley Armour-Garb, James Woodbridge, and W.V.O. Quine. As far as Quine is concerned, Glanzberg may have a good response here, based on the fact that Quine was primarily interested in formal languages, not in natural language. While the truth predicates in Quine's formal languages were defined so as to apply only to sentences, Glanzberg could insist that if Quine had paid closer attention to natural language he would have seen that in that setting, any attribution of truth to sentences involves existential quantification over propositions. Now, since Quine refused to countenance propositions, Quine would then have had to claim that in any natural language, any attribution of truth to any sentence is false. And he would have had to relinquish his claim that 'true' is nothing but a device of disquotation, at least as far as natural languages are concerned. But, Glanzberg could point out, since Quine's concern was with formal languages, he might have been perfectly happy to say these things; so much the worse for natural languages. Still, it is less plausible that this argument can be applied to other philosophers who take natural language more seriously.

Now, Glanzberg could insist that these philosophers do in fact have the relevant dispositions, despite what they explicitly say when discussing ontology. To that end, he could stress that linguists often propose hidden mechanisms such that other linguists—competent speakers—explicitly deny that there are such things. Surely it would be premature to conclude just from the presence of these denials that the mechanisms in question do not exist. Given this, I do not insist that Glanzberg has no way to develop his view about the logical form of truth attributions to sentences, and accordingly I do not rest my case against Glanzberg's view on this objection. The point is merely that more needs to be said. If indeed logical form is something of which speakers must have tacit grasp, then Glanzberg needs to show that people who understand attributions of truth to sentences have tacit grasp of the logical forms of these attributions, whatever that tacit grasp consists in. And the explicit statements of skeptics of propositions who are happy to ascribe truth to sentences will need to be explained away in one way or another.

6.1.1.2. Contextual Salience

My second and more important concern about Glanzberg's view[67] is as follows. It is central to his view that the expression relation only becomes salient between steps (5) and (6), for it is this change in salience that induces the context shift that enables (5) and (6) both to be (non-paradoxically) true. However, the argument starts out as a proof that there's no proposition that *l expresses*—and indeed the expression relation is mentioned already in step (1).[68] For whoever carries out this argument, the expression relation will be salient from the beginning. But

[67] As (Scharp 2013) explains (and as Glanzberg himself admits in (Glanzberg 2004b), p.289), Glanzberg cannot allow for unrestricted quantification, and, relatedly, must deny that the concept of truth can ever be fully articulated. (See Scharp, pp.117-119). I am quite sympathetic with Scharp's criticisms, but I won't reproduce them here because I have little to add. Note, though, that both these criticisms, if correct, would classify Glanzberg's account as incomplete.

[68] Moreover, Glanzberg's own presentation of the argument relies on some further premises, "(T-Exp)" and "(U-Exp)", which also make reference to *expressing*. See p.33 for these.

if *expressing* is salient already in step (1), then the context shift that Glanzberg argues for does not occur where he needs it to occur.

Glanzberg does provide a reason for thinking that the shift occurs between (5) and (6). He explains that (5) "corresponds to the first point in the proof...where there are no undischarged premises" (p.39). Then he writes, "Hence, I suggest, it is the point where the [expression] relation is accepted as salient in the discourse." So, Glanzberg is claiming that the expression relation is salient when it appears in an undischarged premise, but not before. But this is a substantive claim about contextual salience, one which is far from being obviously true. At least, it is not generally the case that items occurring in undischarged premises fail to be salient. To see the point, consider the following argument, in which any step prefixed by '|' is or occurs under an undischarged assumption.

1. | Suppose Jacob ate your cookies.
2. | If he ate your cookies, then there would be crumbs on his shirt.
3. | Therefore, there are crumbs on his shirt.
4. | There are no crumbs on Jacob's shirt. Contradiction.
5. Therefore, he did not do it.

One mark of an item's contextual salience is its ability to serve as a referent of subsequent anaphors.[69] And it is clear that the occurrences of 'he' and 'his' in (2), (3), and (5) anaphorically refer to Jacob, and that the occurrence of 'it' in (5) anaphorically refers to eating the interlocutor's cookies. So, although they are introduced in premises that occur under an undischarged assumption, it can't be that Jacob and *eating the interlocutor's cookies* become salient only with the advent of step (5).

[69] See, e.g., (Karttunen 1976) for more on markers of salience.

6.2. Glanzberg's Criticisms of Other Contextualist Views

Glanzberg briefly addresses diagnoses of the Liar paradox that target the truth-predicate rather than posit a tacit quantifier. In particular, he discusses views according to which the truth-predicate is sensitive to contexts of use. He writes:

> A more common idea [than Glanzberg's own] is to suppose that the truth predicate itself contains a hidden indexical component. Let me briefly note that I do not think this is a promising option. It is a commonly voiced objection to it that we simply do not intuitively see such an indexical element in our ordinary truth predicate, expressed by the ordinary term 'true'. I believe this line of argument can be bolstered. If there were such a hidden indexical, it would behave as other implicit parameters do. In the case of a gradable adjective, for instance, we can see the hidden comparison class at work when we bind the hidden variable, as in:
>
> Most species S have members that are small for S.
>
> We see no such behavior with the truth predicate (pp.30-31).

In an attached footnote, Glanzberg refers his readers to some further diagnostic tests for context-sensitivity, though he does not explicitly spell out why the truth predicate fails these tests.[70] I will not go through the details here, since what Glanzberg says in the above passage is already persuasive.

6.2.1. *Ad hoc*-ness and Glanzberg's Own View

Prima facie, Glanzberg's argument here seems odd. On the one hand, the essence of his objection to views that posit a context-sensitive truth predicate is to stress that neither casual inspection nor certain standard tests show this expression to be context-sensitive. Yet at the same time, Glanzberg's own approach involves positing a kind of context-sensitivity that he considers extraordinary. He writes:

> We now have some idea what the extraordinary context dependence involved in the Liar is. It is the dependence of the background domain of truth conditions upon context. This is not ordinary context dependence, as it is the context dependence of the *background*

[70] Here he cites (Larson 1988) and (Ludlow 1996).

> domain, the maximal domain of truth conditions with which speakers may form propositions. It markedly does not behave like other forms of context dependence. But it is a form of context dependence nonetheless (p.43).

A sensible initial reaction to Glanzberg's view is: if Glanzberg is allowed to posit an unusual variety of context-sensitivity, then why cannot his competitors do the same?

However, Glanzberg has a ready answer to this question:

> though extraordinary, the context dependence involved [in Glanzberg's view] does find motivation in the linguistics of context dependence….it is not an *ad hoc* posit (p.28).

Thus, the idea must be, it is generally acknowledged by linguists that the expressive resources available to interlocutors in a context influences the variety of possible worlds they can distinguish in that context, and thereby influences the background domain of truth-conditions for that context. Thus, given the assumption that propositions are sets of truth-conditions, it makes sense that a conversational move which increases the expressive resources could induce an expansion in the range of propositional quantifiers—although, Glanzberg grants, this occurs quite rarely. By contrast, Glanzberg holds, since 'true' does not pass the standard test(s?) for having a hidden indexical, anyone who targets it as a source of context-sensitivity is thereby making an *ad hoc* posit.

So, Glanzberg's strategy is to charge that diagnoses of the Liar paradox that contradict the linguistic data (e.g., by positing context-sensitivity in expressions that do not pass standard tests for context-sensitivity) are *ad hoc*. By contrast, diagnoses whose posits are sanctioned by the dominant views in linguistics are not vulnerable to this charge, however unusual might be the phenomena that they posit, such as "extraordinary" varieties of context-dependence.

In fact, for the reasons I spelled out in the introduction, it might even be a virtue that a diagnosis of the Liar paradox relies on positing some unusual phenomena. For any piece of reasoning involved in any version of the Liar paradox, eminently plausible premises seem

compellingly to lead to an utterly unacceptable conclusion. This gives us reason to expect that something unusual is going on, either with the sentences involved or with the steps leading from one to the next. Indeed, any view which purports to avoid contradictions by appealing to the normal operation of familiar semantic rules runs a risk of not being able to explain why the premises and inferential steps so strongly seem to be acceptable. (Of course, that is not to say that the wilder the malfunction posited by a diagnosis, the more plausible the diagnosis.) Glanzberg can claim to avoid this risk by positing a kind of context-sensitivity that, while sanctioned by the dominant views in linguistics, is highly unusual.

All of this defense of Glanzberg is fine as far as it goes. But it is worth keeping in mind that even if sensitivity of the ranges of tacit quantifiers to the background domain of truth-conditions is sanctioned by contemporary linguistics, I emphasized in Section 6.1.1.2 that this is less obviously so for the claim about salience that Glanzberg needs (namely, that in his version of the Liar reasoning, the expression relation is not salient until it appears in a premise with no undischarged assumptions). Unless this claim is also sanctioned, Glanzberg's own view is vulnerable to the *ad hoc*-ness objection that he raises.

6.2.2. *Ad hoc*-ness, (Burge 1979), and Aberrationism

Whatever may be the applicability of the *ad hoc*-ness charge to Glanzberg's view, (Burge 1979)'s view is a ripe target for this criticism, since it indeed holds that "the truth predicate…contains a hidden indexical component" (Glanzberg p.30). Burge starts with the idea that instead of there being a single property, truth, there are infinitely many truth-like properties, arranged hierarchically as on the view of (Tarski 1935) and (Tarski 1944). For instance, at level 1 there is a property, truth_1, which applies to all and only the pre-theoretically-true sentences that

contain no semantic terms. So, sentences like 'Snow is white' are true_1. Then, at level 2, we find all and only the pre-theoretically-true sentences that make semantic claims about sentences that contain no semantic terms. Thus, the sentence ''Snow is white' is true' is true_2. And so on, infinitely far up. Then, Burge's thought is, in each context in which an application of 'true' is evaluated, there is some *i* such that truth_i is maximal for that context, in the following sense: where *n* is the highest number of nested semantic attributions which are up for evaluation in that context, $i = n + 1$. Truth_i is then the referent of 'true' relative to that context.[71] Since it is obvious that this view posits indexicality in the truth predicate, it is plain that the view falls prey to Glanzberg's *ad hoc*-ness charge.

By contrast, according to aberrationism, the truth-predicate has no indexical component, hidden or otherwise. In every non-paradoxical sentence, 'true' behaves as a context-insensitive predicate that refers to truth.[72] What 'true' does in otherwise-paradoxical sentences—namely,

---

[71] Here I omit many complexities of Burge's view that are irrelevant for present purposes. Despite my simplification, however, there is reason to think that in its relevant aspects the view is as I present it. Here is a passage that provides some evidence for this. Burge writes:

> Suppose Dean says,
>
> (i) All Nixon's utterances about Watergate are untrue
>
> and Nixon asserts
>
> (ii) Everything Dean utters about Watergate is untrue.
>
> Each wishes to include the other's assertion within the scope of his own assertion….each person's truth predicate should be assigned the same subscript, *i*….*we assume i is high enough to interpret any statement by Dean or Nixon other than (i) or (ii)*….[I]n evaluating (i) and (ii), *we* use 'true_{i+1}', since on this approach sound semantical evaluation will be forced to a higher level….Suppose Dean has uttered at least one truth_i about Watergate. It follows from the semantical rules for the quantifier…that Nixon's assertion (ii) is…not true_{i+1}….If none of Nixon's other Watergate utterances besides (ii) are true_i, then since (ii) itself is not true_i…Dean's (i) is true_{i+1}. On the other hand, if Nixon eked out at least one true_i statement, then Dean's (i) is not true_{i+1}….By erasing the subscripts…we have a piece of reasoning that is intuitive. Our theory accounts for the reasoning (p.194, italics mine).

Burge's argument here is that, by replacing each subscripted term with 'true' but taking the resulting occurrence of 'true' to attribute the truth-like property attributed by the original subscripted term, we accurately model the behavior of 'true'. In the sense of 'minimal' defined in the main text, $i + 1$ is minimal with respect to any context of evaluating (just) (i) and (ii); and sure enough, here we see Burge suggesting that the referent of 'true' in any such context is true_{i+1}. So, this evidence suggests that, as I claimed, Burge holds that 'true' is sensitive to contexts of evaluation, and that in any context of evaluation X it refers to the truth-like property that is minimal with respect to X.

[72] Strictly speaking there may be an exception to this, since whether a sentence is a Liar sentence can depend on the context. Thus, on my view whether an occurrence of 'true' refers to truth must also depend on the context. See below for further discussion.

fail to refer to truth—is then a one-off aberration from its behavior everywhere else. (Thus, aberrationism does not predict that 'true' will pass the standard tests for context-sensitivity, or even for sensitivity to component-contexts.) This is precisely how advocates of aberrationism are able to explain why Liar sentences so strongly appear to lead to contradictions, even though none in fact follow. So, it is hardly an objection to point out that on their view what happens in Liar sentences is different from what happens everywhere else. But this is what Glanzberg's charge of *ad hoc*-ness would come to, as applied to aberrationism.

6.3. (Burge 1979) in Earnest

Since I have applied Glanzberg's *ad hoc*-ness charge to Burge, I should mention that Burge's discussion of a similar problem supplies him with a potential response to this charge. This similar problem has to do with our intuitive judgments about the univocality of 'true'. As Burge observes, Tarski has been charged with violating these intuitive judgments:

> Criticisms of Tarski's construction as a resolution of the natural language paradoxes have taken several forms. It has been held…that 'true' is univocal, whereas Tarski fragments the notion of truth into infinitely many predicate constants (p.171).

Burge thinks that he has a way to avoid this problem, however:

> What of the univocality criticism of Tarski?....Unlike Tarski, we[, the author,] do not interpret our systems as involving constant truth predicates. In natural language there is a single indexical predicate. We represent this predicate [in our formal model of English] by the schematic predicate expression 'true_i'. This expression may in particular contexts be filled out by any of an unlimited number of numerical subscripts. Any one of the resulting predicates (formally, there are infinitely many) may represent a particular occurrence of 'true' in a context in which its application is fixed. Thus numerals substituted for '*i*' mark not new predicate constants, but contextual applications of the indexical 'true'. We have a general method for using this predicate. The existence of this method…provides considerable substance to the notion that 'true' has a single meaning (p.191).

It is indeed a virtue of Burge's view that it allows 'true' to have a single meaning; and in that sense, he does succeed in allowing that this expression is univocal. However, the pre-theoretical impression that 'true' is univocal goes even further. Prior to consideration of the Liar, we are also inclined to judge that 'true' has a single referent, which is attributed in assertive utterances of 'true'. Burge's view violates this intuitive judgment. At this point in the discussion—the point where we are talking about intuitive judgments of univocality—the pressure of Glanzberg's observation that 'true' fails the standard tests for context-sensitivity should be keenly felt.

Burge responds to such concerns by claiming that we have some further pre-theoretical impressions which his view respects and the single-referent view does not. "[T]he view that 'true' has a single extension is in conflict," he tells us, "with intuitions about…the moves from (b) to (c) in our examples" (p.191). Here is one of the examples to which Burge is referring:

> (C) Suppose a student, thinking that he is in room 10 and that the teacher in room 9 is a fraud, writes on the board at noon 8/13/76: (a) 'There is no sentence written on the board in room 9 at noon 8/13/76 which is true as standardly construed'. Unfortunately, it being Friday the 13th, the student himself is in room 9, and the sentence he writes is the only one on the board there-then. The usual reasoning shows that it cannot have truth-conditions. From this, we conclude that it is not true. But this leads to the observation that (b) there *is* no sentence written on the board in room 9 at noon 8/13/76 which is true as standardly construed. But then we have just asserted the sentence in question. So we reason (c) that it is true (p.179).

Burge interprets this example as follows.

> In the moves from (a) to (b) and (b) to (c) in example (C), there seems to be no change in the grammar or linguistic meaning of the expressions involved….Since there is a shift from saying that the relevant sentence is not true to saying that the sentence is true ((b) to (c))—a shift in truth value without change in meaning—there is an indexical element at work. The indexicality is most plausibly attributed to the truth predicate (p.179).

Burge's claim, then, is that we find the shift from (b) to (c) intuitive, and therefore a good semantic account of the reasoning will allow that both (b) and (c) are true. But, his view is, the best way to allow this is to hold that the property which (b) denies of the sentence in question

(namely, the token of ‘There is no sentence written on the board in room 9 at noon 8/13/76 which is true as standardly construed’ which is written on the board at noon 8/13/76) is distinct from the property which (c) affirms of the sentence. This can only occur if ‘true’ undergoes a reference shift between (b) and (c). If Burge is right, then his observations here help him respond to Glanzberg’s criticism. If correct, Burge’s observations count as positive evidence that ‘true’ experiences contextual reference shifts, in a way that allows both (b) and (c) to be true. Thus, even if ‘true’ fails some tests for context-sensitivity, if Burge is right then there is also some positive evidence that it exhibits some context-sensitive behavior.

Here is my response. As I emphasized in the introduction, given the eminent plausibility of the inferences in the Liar reasoning and the unacceptability of the conclusion, we should be prepared to give up some of our pre-theoretical impressions about Liar sentences, or about the reasoning that leads to paradox. So, to provide the strongest support for Burge’s view, the evidence of context-sensitivity should be independent of our impressions about what happens in contexts in which Liar sentences are uttered. Burge is proposing not only that ‘true’ exhibits context-sensitivity in contexts in which Liar sentences are uttered, but, more strongly, that it is context-sensitive more generally (though, perhaps, sensitive only to context-shifts that involve changes in semantic attributions). For evidence of this, we need to see evidence of context-sensitivity outside of paradoxical contexts. But Glanzberg’s objection is that such evidence cannot be found. If the best Burge can do is point to context-sensitive behavior in paradoxical contexts, we have no reason to believe that ‘true’ exhibits this behavior anywhere else, and so no reason to accept his description of the general behavior of ‘true’. Burge is urging us to sacrifice a more trustworthy pre-theoretical impression (one about the behavior of the word ‘true’ in a wide

range of different contexts) in favor of a less trustworthy one (one about the status of certain sentences which are crucially involved in the Liar reasoning).

These reflections raise the question of whether a friend of Burge should instead claim that 'true' shifts in reference in, and only in, contexts of utterance that involve otherwise-paradox-inducing sentences. Indeed, one might reasonably wonder why consideration of the Liar reasoning motivates my view, rather than one that posits aberrations that occur in certain very specific contexts of utterance rather than in certain very specific component contexts.

However, this play on Burge's view has a problem of its own: it entails that even occurrences of 'true' in straightforwardly non-paradoxical sentences shift in reference, relative to any context in which a seemingly-paradoxical sentence has been uttered. Thus, for example, if someone utters sentence A, then relative to that context of utterance even the occurrence of 'true' in the sentence

(Snow) 'Snow is white' is true.

has shifted in reference. But this consequence seems false; uttering a paradoxical sentence does not change what references we make when we then utter a non-paradoxical sentence. Moreover, in addition to being implausible this consequence is unmotivated, as it does nothing to enhance the effectiveness of the view that it serves in avoiding paradox. All the work is done by the shift that occurs in the paradoxical sentence.

This last remark brings to mind a yet further view. Notice that *contingent Liar sentences*, such as the sentences described in Burge's case (C), show that context can affect whether or not a sentence is a Liar sentence. On the view now being canvassed, an occurrence of 'true'

undergoes a referential shift when, and only when, the context renders the sentence in which it occurs a Liar-like sentence.[73] More precisely:

(Context Sensitive Occurrences)	Occurrences refer only relative to context, and
(Context Sensitive Aberrations)	For any sentence S and any context C, if S is Liar-like relative to C then, relative to C, S's key occurrences of its alethic expressions differ in reference from the expressions of which they are occurrences, and
(Context Sensitive Determined)	For any sentence S and any context C, if S is Liar-like relative to C then what S says relative to C is determined by the reference-in-C of S's key occurrences of its alethic expressions, rather than by the reference-in-C of the expressions of which they are occurrences.

Clearly this view is just a version of aberrationism that has incorporated the general relativization of reference to contexts. So, let's call it *CSaberrationism*, for "context-sensitive aberrationism". CSaberrationism is a version of aberrationism in that it adopts versions of (Aberrations) and (Determined), but it is contextualist in the sense that it claims that the reference of occurrences of 'true' can change with the context. Since CSaberrationism is a version of aberrationism, I have little to say against it. Indeed, I think adopting CSaberrationism is precisely how aberrationists should react to contingent Liar sentences. It is only for simplicity of presentation that I have suppressed issues of context-sensitivity throughout. Returning now to the dialectical context of my criticism of Burge, my point is this: it is hardly an objection to aberrationism to point out that contextualists can solve the problems that arise for their views by adopting a view that is a version of aberrationism.

Burge's approach serves as just one example of a more general trend: views that posit the same mechanisms at work both within Liar sentences and in non-Liar sentences overwhelmingly

[73] I am grateful to Yuna Won, David Fielding, and Brandon Conley for helpful discussion of this issue.

misrepresent the behavior of the latter. Accordingly, we should question the common wisdom that it is *ad hoc* to solve the paradox by claiming that something exceptional goes on in paradoxical sentences. Rather, we should be open to the idea that the Liar Paradox reveals something genuinely distinctive and new, a kind of linguistic behavior that is peculiar to the words and sentences involved. This lesson lies at the heart of aberrationism.

6.4. (Simmons 1993)

In (Simmons 1993), Keith Simmons develops a context-sensitivity approach to the Liar that differs significantly from Burge's. For Simmons, "a given use of 'true' applies to all the truths, except for certain *singularities*—sentences to which the given use does not apply truly or falsely" (preface p.*x*). Moreover, a sentence which is a singularity relative to one use of 'true' can be in the extension of a different such use.

The motivation for Simmons' view as an approach to the Liar paradox is that it tries to honor certain pre-theoretical impressions about certain evaluations of utterances of Liar sentences. On Simmons' story, reflection on a Liar sentence goes like this. First, someone utters a strengthened Liar sentence—one that seems to say of itself that it is not true. A few brief reflections on this sentence lead us to the conclusion that it is paradoxical. This brings about a context in which we can correctly evaluate the sentence as being neither true nor false. Then, bearing this evaluation in mind, we reflect on what the sentence says in the first place, namely, that it is not true. This last reflection brings us into yet a third context, in which we can correctly evaluate the sentence as being true.

(Grim 1995) offers some compelling criticisms of Simmons' view. The most impressive of these points to some inconsistencies in Simmons' claims about the context-sensitivity of

'true'. As we saw, according to Simmons the referent of 'true' shifts relative to each context of evaluation. But, Grim asks, what Simmons can say about the sentence X:

(X) X is not true in any context (Grim 1995, p.468).

(Here it is understood that 'context' refers to contexts of evaluation.) Grim explains that "In reflection upon X…Simmons accepts the reasoning that X cannot be true in any context" (p.468). However, as Simmons would surely grant,

> This seems to lead to the reflection given what X says, that
> (Y) X is true after all.

But then, Grim explains,

> Simmons accepts the validity of the argument and the truth of Y. But "if we take the use of 'true' in our final evaluation Y as a context-sensitive use, then the intuitively valid argument is invalidated. So we should regard the use of 'true' in Y as context-independent" [(Simmons 1993, p.173)]. In direct contrast to the initial claim that "there is in English a single, context-sensitive truth predicate" [(Simmons 1993, p.x)], Simmons concludes that "there are uses of 'true' that are context-independent" [(Simmons 1993, p.174)] (Grim 1995, pp.468-469).

Admittedly, it is an appealing feature of Simmons' view that it allows us to truly assert Y after reflecting on X, as we are naturally inclined to do. But Grim's observations—which I fully endorse—show that Simmons' way of accommodating this inclination comes with a heavy cost. I mention Simmons' view only in order to emphasize that aberrationism does not fall afoul of Grim's criticism. Like Simmons' view, aberrationism could be said to posit some sort of context-sensitivity in 'true': whereas Simmons holds that 'true' is sensitive to contexts of evaluation, aberrationists hold that it experiences referential shifts in a very limited range of component contexts. Accordingly, one might wonder whether X above causes problems for aberrationism as well. This, however, is not the case. According to aberrationism, relative to every context of evaluation, the occurrence of 'true' in X fails to refer to truth; it is indeterminate between ascending and descending truth. Suppose we are given a context of evaluation, C. If the

occurrence of ‘true’ in X refers to truth in C, then X’s truth relative to C entails its non-truth relative to C, and vice versa, and we have a paradox. Thus, the occurrence of ‘true’ in X fails to refer to truth relative to C. By my reasoning in Section 3.3.2, this occurrence is indeterminate in reference between ascending and descending truth.

As for the semantic value of X itself, it follows from the above that X is indeterminate relative to every context of evaluation. To see why, recall from Section 3.3.2 the two views about sentences that contain indeterminate expressions. If every such sentence is indeterminate, then in particular X is indeterminate in every context of evaluation, since, as just argued, its occurrence of ‘true’ is indeterminate relative to every such context. On the other view, a sentence containing an indeterminate expression is true if every way of resolving the indeterminacy would render it true, false if every such resolution would render the sentence false, and indeterminate otherwise. Recall also that every sentence that denies of itself that it is ascending true is false, and every sentence that denies of itself that it is descending true is true. So, suppose we are given a context C. If the occurrence of ‘true’ in X were to refer to ascending truth relative to C then X would be false in C; whereas if the occurrence of ‘true’ in X were to refer to descending truth relative to C then X would be true relative to C. So, on the view under consideration, if ‘true’ is indeterminate relative to C then X is indeterminate in C. Thus, X is indeterminate relative to every context of evaluation. Unlike Simmons, however, I do not conclude from all this that X is true after all. Since X does not say what it appears to say, that conclusion must be resisted. This enables me to claim that Y is false, and so I am not forced to conclude that the occurrence of ‘true’ in X was “context-independent”, referring to truth after all.

On the other hand, it is no problem for aberrationism if the reference of some occurrences of ‘true’ is independent of contexts of evaluation. Indeed, contexts of evaluation are not even the

place to look for trouble, since the main mechanism of aberrationism is to posit reference shifts in response to component contexts. Still, one might try to formulate a problematic sentence by quantifying over component contexts:

(X*) In no component context does the occurrence of ‘true’ in X* refer to anything which would make the resulting sentence true.

However, X* makes no sense. Each occurrence has a component context as an essential ingredient, so it makes no sense to speak of reference by occurrences as being relativized to component contexts.

Still, one might think that the following related sentence causes problems:

(X**) The occurrence of ‘true’ in X** fails to refer to anything that would make X** true.

However, it is straightforward what aberrationism says about sentences like this: they fail to say what they appear to say, because their occurrences of ‘true’ fail to refer to truth. Thus X** fails to say anything that entails that X** is not true; so, there is no paradox. For the sake of explicitness, I’ll now take a moment to spell out how I arrive at these conclusions, given my claims in earlier sections.

Assume that the occurrence of ‘true’ in X** refers to truth. If that makes X** true, then X** is not true. (That is because, on the assumption given, X** denies that its occurrence of ‘true’ refers to anything that makes X** true.) Thus, it must be that the occurrence’s referring to truth fails to make X** not true. But then the occurrence of ‘true’ in X** fails to refer to anything that makes X** true; and so, X** is true after all. We arrived at this contradiction by assuming that the occurrence of ‘true’ in X** refers to truth, so we should give up that assumption. By my reasoning in Section 3.3.2, this occurrence is indeterminate in reference between ascending and descending truth.

7. Token-Based Views

In (Gaifman 1992), Haim Gaifman develops an approach to some Liar-like phenomena that focuses on sentence-tokens. His discussion takes off from examples such as the following:[74]

> line 1: The sentence on line 1 is not true.
> line 2: The sentence on line 1 is not true.

Reflecting on this example and other similar ones, Gaifman writes:

> The moral of all these puzzles is simple. In situations of this nature we should assign truth values not to sentence types but to their tokens. Having concluded that the line-1 sentence [token] is not true, we state the conclusion by displaying another token of the very same sentence [type]. The second token, on line 2, expresses something altogether different from what is expressed, if anything, by the token on line 1 (pp.224-225).[75]

Gaifman's idea, then, is as follows. Assume that in the above sentence tokens, the tokens of the phrase 'The sentence on line 1' refer to the sentence-token displayed on line 1. (This is the most charitable interpretation of the above passage. See below for what happens if we do not assume that reference is to the sentence-token.) According to Gaifman, the sentence token on line 1 is neither true nor false, but the token on line 2 is true. On his view, the token on line 1 has the truth value "GAP"—neither true nor false.[76] But since this token fails to express the proposition that it (that very token) is not true, the token on line 1 is not (paradoxically) made true by the fact that it is not true. Nonetheless, the token on line 2 expresses the (true) proposition that the token on line 1 is not true. In its general form, Gaifman's view is that when assigning a value of either truth or

[74] See pp.224-225 for Gaifman's statement of his view. Contingent liars—sentences that are only contingently paradoxical—pose a problem for Gaifman as presented in the main text. For an example of such a sentence, see (Simmons 1993) pp.101-102. However, Gaifman deals with this difficulty by using the abstract notion of a *pointer* instead of the notion of a token. For brevity, I omit these complexities from the main text.
[75] Here I have disambiguated Gaifman's uses of 'sentence' by inserting 'token' and 'type'. What results is the most charitable interpretation of his view. But this interpretation masks an important question that I will take up shortly in the main text: even if when paradox threatens we ought only to assign truth values to sentence tokens, how should we evaluate seemingly-paradoxical tokens that attribute truth values to sentence types?
[76] See p.225.

falsehood to a sentence-token would result in paradox, this token a) is neither true nor false, and b) fails to say this about itself. Nonetheless, however, other, unproblematic tokens of the same sentence-type (such as the token on line 2) can correctly say of this token that it is not true.

Gaifman spends much of his essay developing a procedure for consistently assigning truth values to sentence-tokens in light of examples such as the one just given. And for all that I will say here, Gaifman's claims are both true and effective for avoiding the problems that these examples raise. My point is that his remarks do not suffice to solve the problem that centrally concerns me. Despite his recommendation that "In situations of this nature we should assign truth values not to sentence types but to their tokens," many sentences—both types and tokens—do attribute truth values to sentence types. Thus, the question arises what Gaifman should say about such cases. In particular, one wonders what he should say about sentence-types S which appear to express the negation of the Russellian proposition <S, truth>, and about sentence-tokens *s* that appear to express the negation of the proposition <S, truth>, where S is the type of which *s* is a token. Presumably this is something that any token of a Liar sentence would appear to do. I will consider these cases in turn.

Let us begin with sentence types, and A in particular. Is the case of sentence A a "[situation] of this nature," in which "we should assign truth values not to sentence types but to their tokens"? That is, on Gaifman's view, does A lack a truth-value? If so, then he must explain how exactly that comes about, given that many other sentence-types do have truth values. One option, of course, is simply to take a step toward embracing aberrationism: he could hold that Liar sentence-types S lack truth values because they fail to express the negation of the Russellian proposition <S, truth>. On the other hand, Gaifman could hold that sentence types such as A do have truth values; e.g., he could say that they are true, that they are false, or that they have the

value GAP. In any of these cases, he would have to spell out how paradox can be avoided.[77] Again an option for him would be simply to move in the direction of aberrationism, and say that such sentences S are indeterminate, because they fail to express the negation of the Russellian proposition <S, truth>. The point is that whatever he might say, Gaifman's claims about the tokens on lines 1 and 2 provide little guidance here, since unlike A, those tokens concern a sentence-token rather than a type. To qualify as a general solution to the Liar paradox, Gaifman's view needs to address the status of sentence types that appear to comment on the semantic values of sentence-types.

Gaifman's extant remarks are similarly unhelpful when it comes to sentence tokens that appear to comment on the semantic status of seemingly-paradoxical sentence types. Consider, in particular, the following token of A:

(A_t) A is not true.

Does the token A_t express the negation of the Russellian proposition <A, truth>? Given what we have seen him say already, it might be natural for Gaifman to answer affirmatively. Gaifman already holds that some tokens of a sentence-type *t* can without any trouble comment on the semantic status of some other, problematic tokens of type *t*. So perhaps he would also hold that some tokens of type *t* can also comment on the semantic status of the problematic type *t*. The general idea would be that some tokens of any paradoxical sentence (type or token) can successfully state the status of that sentence. If he took this line, Gaifman would have two options, depending on what he would say about the sentence-type A. He could hold that A has no truth value, but that nonetheless the token A_t (correctly) expresses the negation of <A, truth>, or

[77] There is a third option: one says that A has a truth value, but that it is indeterminate which value A has. However, I doubt that Gaifman would take this option. The motivation for taking this option is the threat that A generates paradox. But because "in situations of this nature we should assign truth values not to sentence types but to their tokens," that threat would motivate Gaifman to deny that A has a truth value.

he could say that A has the truth value GAP, and so because A_t expresses the negation of <A, truth> A_t itself has the value GAP.

However, it is not clear that A_t can express the negation of <A, truth> unless A itself does this. Indeed, some of Gaifman's own remarks emphasize this point:

> in general what a token expresses depends on (1) what it says, i.e., on the sentence type and (2) on the whole network: on the tokens to which the sentence refers and on the tokens to which they in their turn refer, etc. (p.225).

While it is plausible that (1) plays a significant role quite generally—that is, in general what proposition (if any) a sentence token expresses depends partly on of what type it is a token—factor (2) is present only for sentence tokens that comment on the status of other sentence tokens. When it comes to A_t this factor is absent, since A_t concerns a sentence type. Thus, A_t's truth value (if any) cannot depend on those of "the tokens to which [A_t] refers." So, if Gaifman's remarks in the above passage are correct, then what proposition A_t expresses is determined solely by "what it says, i.e., on the sentence type." But if what proposition A_t expresses depends on its being a token of A, then it is hard to see how A_t could express the negation of the Russellian proposition <A, truth>, unless A itself expresses the negation of that proposition. And, of course, if A itself expresses the negation of that proposition, then it needs to be explained how paradox is to be avoided. The point is that if Gaifman wants to say that A_t expresses the negation of <A truth>, then he has to say more about how that comes about. If the passage quoted above is any indication, his answer would likely involve claiming that A expresses the negation of <A, truth>. But this claim raises problems that are not obviously solved by Gaifman's view, given that that view only concerns sentence tokens that comment on the status of other tokens.

Alternatively, Gaifman could claim that neither A nor A_t expresses the negation of <A, truth>. That would sit well with his claim that "what a token expresses depends on…the sentence

type," and would free him of the problems associated with allowing that A expresses the negation of <A, truth> (p.225). However, if Gaifman were to take this route, then his view about tokens that concern sentences such as A would not even come close to involving a generalization of the approach that he develops in connection with sentence tokens that concern other tokens. The crux of that approach is that its allowance that certain tokens can (correctly) state the semantic status of items (tokens, in Gaifman's original example) that are paradoxical. But here we do not have a case of A_t stating the semantic status of the paradoxical entity in question (namely, the sentence-type A).

Again, my point is not to criticize Gaifman's approach to the examples that he discusses. It is merely to emphasize that this approach alone is insufficient to solve the problem that concerns me—that is, the problem that appears to arise in connection with sentences (types and tokens) that attribute truth values to sentence types.

7.1. Another View that Focuses on Tokens

In the above discussion of Gaifman, I left out the view that A expresses the negation of the Russellian proposition <A, truth>, but no token of A expresses the negation of that proposition. I left this view out because it is antithetical to the idea that seems to motivate Gaifman, namely, that what cannot be said by a sentence type (without inviting paradox) can be said safely by some of its tokens.

It is worth noting that quite independently of its unavailability to Gaifman, this view faces a serious problem. Whether or not any token of A can express the negation of <A, truth>, the claim that A itself does is enough to generate a paradox. Consider the argument that I presented in Section 2.4, which begins from consideration of A and ends with a contradiction.

Even if that argument can never successfully be tokened in order to explicitly derive a paradox (since the token of A that it would contain would fail to express the negation of <A, truth>), as long as A expresses the negation of <A, truth>, the argument itself, considered as a sequence of sentence types, successfully proves a contradiction. (Pending the success of some other solution to the Liar paradox, that is.)

8. Prosententialism

Another approach to Liar phenomena is provided by *prosentential theories* about 'is true' and its kin. Prosentential theories about 'is true' hold that sentences involving this expression behave like anaphora. Anaphora are expressions, such as pronouns, which inherit their reference from other expressions that occur earlier in the discourse. Consider, e.g., the following sentence:

(M) When Martha opened the car door, she could hardly believe what she saw.

Both occurrences of 'she' in M inherit their reference from the occurrence of 'Martha' at the beginning.

According to prosentential theories, any sentence involving the word 'true' is a *prosentence*—a sentence that inherits all the content it has from some antecedent sentence. To see the view, consider the following dialogue:

Chrissy: Mousavi became president of Iran in 2009.
Angela: That's true.

On any prosentential theory, Angela's sentence has exactly the same content as Chrissy's; it is just as if Angela had uttered the same sentence as Chrissy did. Prosentential theories have similar ways of handling 'is not true'. Suppose instead that the following dialogue takes place between Chrissy and Angela:

Chrissy: Mousavi became president of Iran in 2009.

Angela: That's not true.

In this case, Angela's sentence still inherits its content from Chrissy's; the only difference is that now it is just as if Angela had said 'Mousavi did not become president of Iran in 2009'.

Prosentential views about truth have a convenient response to the Liar paradox that is somewhat like my own. According to any prosententialist, sentence A is a prosentence that is its own antecedent. Thus, A inherits whatever content it has from itself. But since prosentences do not have any content independently of what they inherit from other sentences, when it comes to the inheritance of content, A never gets off the ground. So, A has no content; it fails to express any proposition. Thus, prosententialists conclude, there is no paradox.

There are two prominent prosentential theories about truth, that of (Grover, Camp, and Belnap 1975) and that of (Brandom 1994). For brevity, I will consider only Brandom's view, which improves on Grover, Camp, and Belnap's in an important respect.[78] According to Brandom, 'is true' is a prosentence-forming operator. That is, when it is appended to an expression that refers to a sentence, the result is a prosentence that inherits its content from the sentence to which the referring expression refers. Consider again the sentence (Snow):

(Snow) 'Snow is white' is true.

On Brandom's view, (Snow) is a prosentence whose content is identical to that of the sentence named by the expression ''Snow is white''; that is, the sentence 'Snow is white'.

However, Brandom's view faces some serious problems. James Beebe gives a particularly nice explanation of one of these, which he attributes to (Wilson 1990):

[78] (Kirkham 1992) pp.325-329 raises an important objection to Grover, Camp, and Belnap's view. See also (Beebe 2015) Section 7 for a helpful exposition of this objection, and an explanation of the way in which Brandom's improved version of prosententialism addresses the concern.

> Consider the following example inspired by [(Wilson 1990)'s] criticisms of the prosentential theory.
>
> (44) Steve: Boudreaux won the mayoral election.
> Kate: What that conniving, good-for-nothing bum said was true.
>
> If Brandom's version of the prosentential theory is correct, Kate's utterance should have no more content than Steve's. Clearly, however, Kate's remark does more than simply reassert the content of Steve's remark. It casts aspersions on Steve's character. According to Brandom's seemingly more defensible version of the prosentential theory, a referring expression used at the head of a prosentence serves only to pick out an antecedent from which the prosentence can inherit its content. But referring expressions can be naughty or nice, informative or dull. Once Brandom opens the door for prosentences to be formed by conjoining *any* referring expression to the prosentence-forming operator '…is true,' it seems that he can no longer maintain that prosentences never have any more content than their anaphoric antecedents. [Many] referring expressions are not all like proper names. Very often they bring with them a great deal more content than is strictly necessary for them to succeed in referring. A proper interpretation of prosentences cannot ignore this extra content. (Beebe 2015, Section 7)

(Wilson 1990) raises another objection, connected with sentences such as

(4) That Bleda is vicious is true.

and

(5) It is true that Bleda is vicious.

On any prosententialist view, including Brandom's, the expressions 'that…is true' and 'it is true that…' are what Wilson calls "sentential connectives": when they are applied to a sentence, a sentence results. Therefore, Brandom must analyze (4) and (5) in such a way that the expressions 'that…is true' and 'it is true that…' turn out to be grammatical constituents. However, "an alternative hypothesis is that 'is true' functions as a predicate." This hypothesis, Wilson writes, "is better supported by the data of English than is the [prosententialist] hypothesis that 'it is true that' is a sentential connective in English" (Wilson p.23). According to Wilson, the linguistic evidence about English supports taking the constituents of (5) to be 'it is true' and 'that Bleda is

vicious', rather than, as the prosententialist would have it, 'it is true that' and 'Bleda is vicious'. Here is Wilson's explanation.

> Sentences of certain forms are widely recognized as grammatically related. An explanation of this is that they are related by rules which move an element in one sentence from one position to another to obtain the other sentence. It is a fundamental assumption of grammatical theory that these rules apply only to constituents. Some of these relations follow….
>
> [F]rom (5) we can obtain the *cleft*
>
> It is that Bleda is vicious that is true.
> but not
> *It is Bleda is vicious that that is true.
>
> From (5) we can obtain the *pseudo-cleft*
>
> What is true is that Bleda is vicious.
> but not
> *What is true that is Bleda is vicious.
> ….
>
> Additional evidence comes from the possibility of *inserting parenthetical expressions* between 'true' and 'that', possible usually only at breaks between major constituents:
>
> It is true—as you should know better than anyone—that crimes have been committed.
>
> *It is true that—as you should know better than anyone—crimes have been committed.
>
> Notice also that we have
>
> A: Bill said that Susan is ill.
> B: Is it true? / That Susan is ill? / *Is it true that? (Wilson 1990 pp.23-24).

For my purposes, the upshot of Wilson's observations is that while prosententialism supplies a convenient solution to the Liar paradox, the view conflicts with the linguistic data about English. It does this because it is a quite general claim about the behavior of the English expression 'is true' and its kin. By contrast, aberrationism concerns only the behavior of occurrences of 'is true' in Liar-like sentences, and so it does not challenge any of the data

concerning the behavior of this expression anywhere else. As a second point, whereas the theory of the relation of grammatical constituency is on comparatively solid footing, reference remains a topic of active research and competing theories. Accordingly, then, it is preferable to take the Liar paradox to teach us something surprising about reference than to take it to refute linguists' canonical views about grammatical constituency.

Still, however, it is important to note that strictly speaking, prosententialists do not need to claim that 'it is true that' acts as a separate grammatical constituent in the sentences in which it occurs. Rather, all that they really need to endorse is a claim about the logical form of sentences that involve this expression. As Wilson concedes,

> Some of [the above] relationships show at most that 'that S' is a constituent of surface structure. It might be argued that at the level of representation of logical form, 'it is true that' is nevertheless a constituent. (Wilson 1990 p.24).

Indeed, prosententialists can point out, it is widely acknowledged that expressions with different semantic behavior can occupy some of the same grammatical roles. E.g., in the sentence 'I saw Jolene', 'Jolene' refers; whereas in the sentence 'I saw a woman', the phrase 'a woman' serves as a noun phrase, like 'Jolene', despite the fact that at the level of logical form it expresses existential quantification, not reference to an individual.

All that said, while there are known mismatches between occupying the grammatical role of a referring expression and being a referring expression, there is a general presumption against positing such mismatches. This presumption underlies Wilson's final observation concerning his examples displayed above, that "it remains for the advocates of [prosententialism] to find the syntactical theory that would support their analysis" of these examples (p.24).

8.1. Prosententialism and the Existence of Truth

One motivation shared by all prominent prosententialists is the desire to "undermine the idea that 'is true' is a property-ascribing locution" (Beebe 2015, Section 7). For Brandom, this motivation in turn stems from a desire to deny that there is such a property as truth. Prosententialism facilitates this denial by removing the need to appeal to truth as the property attributed by 'is true'. (Thus, while prosententialism does not in itself require one to relinquish the claim that truth exists, it does remove an obvious reason for making this claim.) Thus, prosententialism tends to appeal to philosophers who regard truth as being suspiciously mysterious, metaphysical, or both.

On the other hand, pre-theoretically the claim that truth exists is uncontroversial; surely the burden of proof is on those who would deny this claim. Of course, someone who recognizes this might nonetheless be impressed by the prosententialist approach to the Liar paradox. But then any such theorist should be drawn to the aberrationist approach, which has similar virtues but allows one to say that truth exists. Like prosententialists, aberrationists also hold that Liar sentences fail to say of themselves that they are not true, and trace this failure to the behavior of the occurrence of 'is true' in these sentences. (Again, the prosententialist view is that (a) the presence of an occurrence of 'true' makes the sentence a prosentence, and that (b) from the way the sentence is constructed, this occurrence is its own antecedent, preventing the sentence from acquiring any content.) However, unlike prosententialism, my view is compatible with the claim—and the linguistic evidence—that 'is true' is a normal predicate whose job is to attribute a property. Thus it is also more congenial to the intuitive, pre-theoretical view that there is a property to which 'true' refers, namely, truth.

9. Wholesale Indeterminism

I will now examine an approach to the Liar-like paradoxes that is closely related to aberrationism, but differs from it in an important way. According to *wholesale indeterminism*,

a) Every occurrence of the word 'true' is indeterminate in reference, as between two or more distinct properties, such as ascending truth and descending truth.
b) However, the extensions of these properties diverge from the class of sentences which we pre-theoretically judge to be true, and from one another, only when it comes to Liar-like sentences.
c) None of these properties is determinately identical with truth.

By (a), the indeterminacy of 'true' is *wholesale* in the sense that it affects all occurrences of 'true' rather than just those in Liar-like sentences. On the other hand, (b) guarantees that, and helps to explain why, nearly all of our informal uses of 'true' are unproblematic.[79]

In broadest strokes, the difference between aberrationism and wholesale indeterminism is as follows. Aberrationism preserves the bulk of our informal reasoning with 'true' by restricting its aberrant behavior to Liar-like sentences, whereas wholesale indeterminism posits widespread aberrant behavior—insofar as indeterminacy is considered aberrant behavior—but insists that the aberrance only affects matters when we are dealing with Liar-like sentences. That said, both views preserve the vast majority of our informal uses of 'true'. Moreover, like moderate aberrationism, wholesale indeterminism makes an effort to accommodate our pre-theoretical impressions about what Liar-like sentences say, allowing for a sense in which these sentences come close to saying what they appear to say.

Before examining some important differences between these two views, I should clarify an important detail about wholesale indeterminism. On first blush, wholesale indeterminism can seem to deny that truth exists. And whatever the other advantages might be, on first blush, that

[79] I take wholesale indeterminism to be in the same general family as the views of (Eklund 2002) and (Burgess 2014). I am grateful to Eklund for bringing wholesale indeterminism to my attention.

denial would certainly make the view less appealing. However, this reaction conflates the denial that truth exists with the denial there is any single property that is determinately identical with truth. As it turns out (see below), the wholesale indeterminist can assertively utter the sentence 'There is a property P such that P = truth'. Indeed, she can even assertively utter the sentence 'There is a unique property P such that P = truth'. Rather, what she cannot assertively utter are things like 'There is a unique property P such that it's determinate that P = truth'.

More precisely, the move I am describing here—and it is a standard one among advocates of views of this kind—is to adopt something like a supervaluational semantics for sentences involving the word 'true'. If *w* is a word that is indeterminate in reference, let a *precisification* of *w* be a model M = <D, V>, where D is a domain and V is a valuation function such that V(*w*) = P, where P is a candidate referent for *w*. For simplicity, assume that the referents of all the other expressions in the language are fixed across all precisifications, so that for each candidate referent P of *w* there is a unique precisification, M_P. *Truth-under-a-precisification of w* is then simply truth-in-M_P for some P that is a candidate referent for *w*. With these notions in place, the idea is that when it comes to sentences that involve the word 'true', the wholesale indeterminist is to endorse all and only those sentences which come out true-under-all-precisifications-of-'true'. Thus, in effect, wholesale indeterminists treat 'true' not as a normal predicate but rather as an expression that involves universal quantification over candidate referents. However—and this is the key point—because the candidate referents differ from our informal truth-attributions only when it comes to Liar-like sentences, our informal truth-talk fails only in connection with Liar-like sentences. For most practical purposes it is harmless to treat 'true' as a normal predicate.

Now to explain the wholesale indeterminist's claims concerning truth. On the one hand, one can readily check that the sentences

(Identity) Truth = truth

and

(Unique) There is a unique property Q such that Q = truth

are true-under-all-precisifications of 'true'. For each property P that is a candidate referent for 'true', 'truth' refers-in-M_P to P. Thus, for each such P, 'Truth = truth' says-in-M_P that P = P and thus is true-in-M_P. Similarly, for each such P, 'P is unique and P = truth' is true-in-M_P, making Unique also true-in-M_P, since the rule of existential generalization is valid in M_P. Since these conditions hold for each candidate referent P and thus for each precisification M, the wholesale indeterminist can endorse Identity and Unique.

On the other hand, consider what happens when we add the determinacy operator. For any sentence φ, ⌜determinately φ⌝ means that φ is true-in-all-precisifications. Now, since wholesale indeterminists take the reference of 'true' to be indeterminate as between two or more distinct candidate referents, for any candidate referent Q there exists a distinct candidate referent P such that 'Q = truth' is false-in-M_P. So, for no candidate Q is the sentence 'Q = truth' true-under-all-precisifications. Thus, the wholesale indeterminist cannot endorse the following:

(Determinate) There is a unique property Q such that determinately Q = truth

As the wholesale indeterminist reads Determinate, it says that there's a unique property Q such that 'Q = truth' is true-in-all-precisifications. But it is a constitutive part of her view that that is false.

9.1. Pros and Cons of Wholesale Indeterminism

9.1.1. A Methodological Point

My first point about wholesale indeterminism is a methodological one. Though it is certainly not beyond question that there is a unique property determinately identical with truth, still this is a claim that many people are pre-theoretically disposed to accept. So, there is motivation for developing approaches to the Liar paradox that enable us to keep this claim. Of course, if those approaches prove too costly, then one will be forced to choose some other alternative, and in that event wholesale indeterminism may be the best option. However, one cannot judge the case until views that posit a property determinately identical with truth have been fully developed. I take this essay to be a step in that direction. On that note, even if all my discussion shows is that aberrationism is the best of the views which insist that there is a property determinately identical with truth, I take that to be an interesting result.

9.1.2. In Practice, but not in Spirit

My second point is that in positing widespread indeterminacy, one commits oneself to some strong claims about the behavior of 'true' in non-Liar-like sentences, claims that are largely unmotivated except for their role in solving Liar-like paradoxes. (Here I have in mind the claim that for each non-Liar-like sentence S that contains 'true', the occurrence of 'true' in S is indeterminate in reference.[80]) While I have stressed that some measure of *ad hoc*-ness is inevitable in any solution, the availability of aberrationism demonstrates that one can restrict the

[80] However, it is worth noting that some might take this view to be motivated by considerations of vagueness. If 'Harry is bald' is indeterminate in truth value, then perhaps ''Harry is bald' is true' is as well. One way to diagnose the latter indeterminacy is by taking 'true' to be indeterminate in reference as between a number of different candidate properties, some of which the sentence 'Harry is true' has and others of which it lacks. If this view is right, then the idea that 'true' is indeterminate can be motivated independently of the Liar paradox. Still, it is worth noting that this diagnosis of sentences such as ''Harry is bald' is true' is controversial. A different approach would be to say that 'true' refers to a single property, such that it is indeterminate whether 'Harry is bald' has that property.

ad hoc-ness to one's diagnoses of the occurrences that are directly implicated, a desirable result. This is something that wholesale indeterminism manifestly fails to do, since it posits a wholesale indeterminacy in the reference of 'true'. In that respect, it has the same flaw as the classic contextualist views that I criticized above.

On the other hand, wholesale indeterminism does have an advantage over those views: if indeed the candidate referents agree with our intuitive judgments of truth everywhere except on Liar-like sentences (as specified in part (b) of wholesale indeterminism), then in practical terms the indeterminacy affects nothing except when we are dealing with such sentences. So, unlike classic contextualists, the wholesale indeterminist can endorse nearly all of our pre-theoretical reasoning involving 'true'; it comes out correct on her supervaluational reading. In this sense, one might claim, wholesale indeterminism respects the linguistic data about 'true' much more effectively than the vast majority of other approaches that make claims about its behavior outside the realm of paradox.

Still, part (b) of wholesale indeterminism is a substantive claim, one that needs to be verified by careful attention to the candidate referents being attested. As I mentioned in passing in Section 3.3.3, Scharp endorses this claim when it comes to ascending truth and descending truth:[81] that is, he claims that for all non-Liar-like sentences S, each of ascending truth and descending truth respects the instances of (T-intro) and (T-elim) that involve S. However, Scharp does not provide a proof, and settling the matter is too big a job to undertake here.[82] It is worth noting that wholesale indeterminism relies on this strong, as-yet-unproven claim.

[81] See his p.186.

[82] For one thing, in general it is hard to prove negative existential claims. For another, most consistent approaches to the Liar paradox end up excluding some pre-theoretically unproblematic sentences from participation in (T-intro), (T-elim), or both. A third difficulty is that it is arguably vague which sentences are Liar-like and which are not.

Moreover, even if the claim turns out to be true, I see a problem in the fact that so much hangs on it. As a methodological principle, it is best for one's accounts of non-paradoxical sentences not to be driven by one's diagnosis of the Liar paradox. Because the thesis of wholesale indeterminacy in the reference of 'true' is motivated by the Liar paradox, it is therefore best that this thesis' consequences be necessarily restricted to paradoxical sentences. One's ability to honor our informal observations and linguistic data about non-paradoxical sentences should not depend on the extent to which the candidate referents for occurrences of 'true' turn out to be truth-like, even if in fact they do turn out to be quite truth-like. In spirit if not in practice, wholesale indeterminism disregards the lessons brought out in my criticisms of contextualist approaches.

9.1.2.1. *Tu Quoque*

Given that I myself introduced two properties (ascending truth and descending truth) that I held to be quite truth-like, a wholesale indeterminist might attempt a dialectical response. If the candidate referents for occurrences of 'true' (whether all of them or just those in Liar-like sentences) turn out not to be very truth-like, then isn't that also bad for my own preferred version of moderate aberrationism, described in Section 3.3? After all, the less truth-like these candidates are, the less plausible it is that Liar-like sentences come close to saying what they appear to say. While this is a reasonable question, a few important observations reveal that the situation just described is much more tolerable for moderate aberrationism than it is for the wholesale indeterminist.

Firstly, and most importantly, even in the situation just described, one could maintain that occurrences of 'true' in non-Liar-like sentences refer to truth, and so behave exactly as one

expects prior to exposure to the Liar paradox. In that important respect, the situation would be much better for aberrationism than for wholesale indeterminism. Secondly, the other central claim of moderate aberrationism—that reference is powerfully influenced by language use—would remain untouched. So, there would be no general problem for moderate aberrationists, only for those that claim that Liar-like sentences come close to saying what they appear to say.

A third observation is that wholesale indeterminism requires a higher degree of truth-likeness than my own preferred view requires. That is because for my view, failures of truth-likeness only affect the diagnosis of Liar-like sentences: if the "second in line" property is truth-like but not dramatically so, then I can still claim that Liar-like sentences come close to saying what they appear to say; I just have to admit that they do not come dramatically close. By contrast, for wholesale indeterminists, failures of truth-likeness affect all of the sentences with respect to which the property in question diverges from our intuitive truth-attributions.

The property *truth*$_{mp}$, defined in (Kripke 1975), serves as an illustrative example here.[83] Kripke's idea is that for any language, there are stages in the determination of truth values for its sentences. Sentences which do not address the semantic status of other sentences are assigned truth values at stage 0. At stage 1, sentences which attribute truth values to stage-0-sentences acquire truth values. And so on, to infinity. Thus 'Snow is white' gets the value *true* at stage 0, and 'The sentence 'Snow is white' is true' gets the truth value *true* at stage 1. By contrast, Liar-like sentences such as A are not assigned any truth values at any stage in this process. Kripke proves that there is a unique (infinite) least ordinal number κ such that all the sentences which will ever acquire truth values via the process just described have received them by stage κ. This

[83] The term 'truth$_{mp}$' stands for *truth-in-the-minimal-fixed-point*. The minimal fixed point is also the least fixed point. (See below in the main text.)

ordinal κ is *the minimal fixed point*. A sentence is then true_{mp} if and only if it receives the value *true* at some stage less than or equal to κ.

For present purposes, the point is this. English speakers generally feel at least some inclination to say that the sentence COND is true:

(COND) If COND is true then COND is true

However, COND is not true_{mp}. So truth_{mp} differs at least that much from truth as pre-theoretically understood. Now suppose that ascending truth and descending truth do not exist, and instead it is truth_{mp} which comes the closest, after truth itself, to respecting (T-intro) and (T-elim). Then, according to versions of moderate aberrationism that endorse the Lewisian views about reference described in Section 2.3, the key occurrences of 'true' in Liar-like sentences refer to truth_{mp}. For aberrationist purposes, this situation is sub-optimal, but still acceptable. For there would still be a reasonably robust sense in which Liar-like sentences come close to saying what they appear to say—though they would not come as close as we would have had with ascending truth and descending truth in place of truth_{mp}. And, all the other, non-paradoxical sentences of English, including COND, would be exactly as we pre-theoretically expect. By contrast, this situation would be much worse for a wholesale indeterminist. Such a theorist would have to reject the intuitive (for many) judgement that COND is true, together with all our other pre-theoretically-acceptable attributions of truth to sentences that are not true_{mp}.

This concludes my discussion of alternative approaches to the Liar paradox that do not embrace (Aberrations) and (Determined). Now I will discuss some approaches that embrace these views, but which target expressions other than those that refer to truth.

10. Alternative Culprits

As I have been stressing throughout, the Liar paradox gives us good reasons to suspect that either within Liar sentences or somewhere in the reasoning that leads to paradox, at least one of the expressions involved must fail to behave as it pre-theoretically seems to behave. Borrowing a term from (Eklund 2002), let us call this expression(s), whatever it is, *the culprit(s)*.[84] And if an expression is under consideration for being a culprit, let's call it a *suspect*. Throughout, I have been advocating that in any Liar-like sentence, the (unique) culprits are the alethic expressions—that is, the expressions that are governed by (Sat-intro) and (Sat-elim), either by definition or via some auxiliary theses. But, why not instead blame the expression that refers to the sentence itself (when indeed there is such an expression), or the expression that refers[85] to negation (when there is one), or all of the above?[86] Before examining any such views in detail, I will present an argument for targeting only alethic expressions, and not anything else, as the culprits. The argument appeals to a methodological principle and then an important observation about seemingly Liar-ish paradoxes.[87]

The methodological principle is this:

(Method)

a) if one cannot set up a seemingly Liar-ish paradox without employing an expression of a given sort, then one is justified in taking expressions of that sort to be culprits in all seemingly Liar-ish paradoxes in which they feature.

[84] An important difference between my use of 'culprit' and Eklund's is that for Eklund, the culprits are false assumptions, whereas for me, they are those expressions whose behavior is not as it pre-theoretically seems to be, on pain of paradox.

[85] The word 'express' is more familiar here. But I reserve 'express' for the relation between expressions and concepts, and between sentences and propositions. Here when I talk of negation I am talking about a certain truth-function, not a concept. So I use 'refer' rather than 'express'.

[86] In fact, (Smith 2006) faces a similar worry. Smith claims that either the name 'A' or the predicate 'true' is the culprit. But, one might ask, why not 'not' instead?

[87] In this section, one of the questions at issue is that of exactly which paradoxes are Liar-like. To avoid begging the question against such philosophers, I use 'seemingly Liar-ish' in place of the contested term 'Liar-like'. See below in the main text for further discussion.

b) And on the other hand, if one can readily set up a seemingly Liar-ish paradox without employing expressions of a given sort, then that counts against one's justification for holding that expressions of that sort are culprits in any seemingly Liar-ish paradox in which they feature.

One thing that makes (Method) appealing is its likelihood of leading to satisfying solutions of the paradox. The satisfactoriness of a solution to one seemingly Liar-ish paradox lies in part in its usefulness against a wide variety of other seemingly Liar-ish paradoxes.[88] Solutions which target expressions that are essential to many different such paradoxes are, if good at all, good also against those many paradoxes. By contrast, a solution that targets an expression that is only essential for formulating one seemingly Liar-ish paradox will not straightforwardly apply to others that do not involve this expression.

In addition to leading to satisfying solutions, (Method) is commonly accepted. For instance, it is common to object to a purported solution to a seemingly Liar-ish paradox by emphasizing the failure of this approach to address other seemingly Liar-ish paradoxes. But the failure of a solution to generalize to various other paradoxes poses no problem for it, unless one assumes that a single phenomenon is responsible in all the cases one has in mind. And on most proposals which take all of various different paradoxes to be due to a single phenomenon, that phenomenon involves a single expression's being the unique culprit.

Others before me have used (Method) to justify identification of the truth predicate as a culprit. (Scharp 2013)'s discussion serves as an excellent example:

> We can construct artificial languages that contain the [other suspects besides the truth predicate], and they are perfectly well-behaved as long as they do not contain truth predicates (or related semantic terms). Of course, we can also construct artificial languages with truth predicates that are perfectly well-behaved as long as they do not contain the [other suspects]. However, the difference is that there are many different ways to construct revenge paradoxes; one involves *truth* and exclusion negation, one involves *truth* and another non-monotonic sentential operator, one involves *truth* and the

[88] Here I leave open whether these other paradoxes are, in some ultimate sense, versions of the Liar paradox.

> conditional, one involves *truth* and an idempotent determinacy operator, etc. Exclusion negation is not involved in each case, nor are any of the other outlaw linguistic expressions. However, truth is involved every time (p.120).

Here Scharp makes tacit appeal to (Method). Because expressions that refer to truth are involved in every known seemingly Liar-ish paradox, he argues, it is these expressions that should properly be targeted as responsible in all cases.[89] Indeed, Scharp also articulates something close to the observation that is central to my argument: he claims that the truth predicate is the only expression that is necessary for setting up all seemingly Liar-ish paradoxes. Thus, by (Method), one can conclude that the truth predicate is the sole culprit in any such paradoxes in which it features.

My only modification to Scharp's claims here is that instead of just the truth predicate, the focus should instead be on all alethic expressions—that is, all expressions that are, either by explicit definition or via intersubstitutability, governed by (Sat-intro) and (Sat-elim), either in general or for some particular natural number n. The reason for this modification is that Grelling's paradox is a palpably Liar-ish paradox that employs 'satisfies' rather than 'true'.[90]

One might reasonably feel concerned about (Method)'s appeal to the vague and obscure property *being seemingly Liar-ish*. Unless this property is defined in some precise way, it will be unclear exactly which paradoxes are eligible to falsify hypotheses as to which expression is the culprit. My own preferred way to help clarify this property is to identify it with *being Liar-like* as

---

[89] Strictly speaking, Scharp's concern is with the everyday concept of truth, assuming there is a unique one. But for my purposes it is safe to assume that he is also concerned with the predicate 'is true' (as ordinarily used).

[90] Here is Grelling's paradox:

1.	|satisfies ('~satisfies (x, x)', '~satisfies (x, x)')	(assume for *reductio*)
2.	|~satisfies ('~satisfies (x, x)', '~satisfies (x, x)')	(Sat-elim), (1)
3.	|Contradiction	Contradiction Intro, (1), (2)
4.	~satisfies ('~satisfies (x, x)', '~satisfies (x, x)')	*Reductio ad absurdum*, (1)-(3)
5.	satisfies ('~satisfies (x, x)', '~satisfies (x, x)')	(Sat-intro), (4)
6.	Contradiction	Contradiction Intro, (4), (5)

Most commentators who consider Grelling's paradox find it to be seemingly Liar-ish.

defined in Section 2. (Recall: a paradox is *Liar-like* if it makes ineliminable use, for some natural number n, of (Sat-intro) and (Sat-elim), possibly modulo some minor cosmetic differences. So, given that (T-intro) and (T-elim) are essentially the special case of (Sat-intro) and (Sat-elim) for n = 0, any paradox that makes ineliminable use of (T-intro) and (T-elim) will count as Liar-like.) One can readily check that this definition agrees with most of our informal judgments about which paradoxes are Liar-ish, and it is hard to come up with counterexamples. However, this way of making 'Liar-ish' precise puts my claim that 'true' is the culprit at a serious advantage by legislating paradoxes that don't involve expressions that refer to truth out of the definition. Therefore in the rest of this section I'll leave 'Liar-ish' unspecified, to be understood in an informal way. It won't matter in what follows, since most people would be willing to count all of the particular paradoxes that I'll discuss as being seemingly Liar-ish.

10.1. Negation

Concerning most views that target 'not' and its kin, I have already said most of what I want to say. For one thing, in the passage quoted above, Scharp already mentions that there are seemingly Liar-ish paradoxes in which negation does not feature, such as Curry's paradox (see below). So, the view that negation is the culprit in the seemingly Liar-ish paradoxes in which it features falls afoul of (Method). Moreover, most views that target negation involve allowing that some sentences that express genuinely contradictory propositions can nonetheless be jointly true. However, this concession violates a belief that is much more deeply entrenched than any that aberrationism requires us to relinquish.[91]

[91] See (Smith 2006) for a related criticism of (Priest 1987).

All that said, there is one particular view that targets negation which I ought to discuss, given its similarity to aberrationism in a certain important respect. Like aberrationists, in (Kearns 2007) John Kearns posits a one-off aberration in the culprit that he identifies; but unlike aberrationists he identifies negation as the culprit. Kearns' view concerns not sentences but rather entities that he calls "*statements*"—speech acts that are performed by uttering sentences and which are either true or false.[92] The core of the view is the claim that when we try to make a "Liar statement"—a statement that asserts of itself that it is not true, we fail. He considers the following example:

(a) Statement (a) is not true

Here, following Kearns, one should take '(a)' to refer not to the sentence displayed, namely, 'Statement (a) is not true', but rather to the (putative) statement that one would make by assertively uttering that sentence. According to Kearns, "In trying to make a statement that is true or false [by uttering the displayed sentence], we end up with a 'statement' that is true if, and only if, it is not true. But this is not possible. Nothing has a property if, and only if, it does not have that property" (p.53). Thus, he concludes, (a) is not in fact a statement, but rather an "attempted statement."

Now, Kearns does think that there is such a thing as the statement

(b) Statement (a) is true

(Similarly to '(a)', '(b)' here should be understood as referring to the statement one would make by assertively uttering the sentence 'Statement (a) is true'.) One initial complication that Kearns acknowledges is that since (a) is merely an attempted statement, the term 'Statement (a)' may fail to refer, thereby robbing (b) of any subject matter. That would, it seems, interfere with (b)'s

[92] See p.32 for Kearns' definition of 'statement'.

being a statement. However, Kearns suggests instead that 'Statement (a)' refers to the "attempted statement" (a), despite the fact that (a) is not a genuine statement (p.53). Let us grant this for the sake of argument. (Note also that the concern is mollified somewhat if one changes (b) by deleting 'Statement'. Even if (a) isn't strictly a statement, whatever it is, as long as it exists one can surely use '(a)' to refer to it.)

Kearns claims that (b) is simply false.[93] At the same time, he denies that the result of applying negation to (b)—namely, (a)—is itself a statement. He writes:

> The intention for negation is that in negating a true statement, we will obtain a statement that is not true, and in negating a statement that is not true, we will obtain a true statement….[This] intention…is realized on most occasions when we negate a statement. It cannot be realized when the negation is applied in such a way as to yield a paradoxical attempted statement. There are…attempted Liar statements, but no actual Liar statements….[These are] cases of [speech] acts that aspire to be acts of a certain kind, but fail to make the grade. (p.53)

So, Kearns posits a one-off aberration in the behavior of negation as applied to statements: the negation of a false statement is usually itself a statement, but sometimes this fails. (Similarly, Kearns is committed to positing a one-off aberration in the behavior of 'not' (and its formal counterpart '~'): for certain, Liar-ish putative statements S (such as (a) above), in assertively uttering the result of concatenating '~' with a sentence that attributes truth to S, one fails to make any statement whatsoever. By contrast, for most other putative statements S', assertively uttering such a sentence will amount to (performing the speech act of) negating <S', truth>. In these respects, Kearns' view is like aberrationism, except that it targets 'not' and the like.)

Before I engage with Kearns in earnest, a word about negations of statements. The word 'statement' is often used to designate sentences or propositions, and the idea of the negation of a sentence or a proposition is clear and familiar. But for Kearns statements are speech acts, and it

[93] See p.55.

is less clear what the negation of a speech act is than what the negation of a sentence or proposition is. Of course, if Kearns' talk of negating statements is unclear, then so is his diagnosis of the Liar paradox. And even further, for Kearns, (a) is the result of negating a statement, namely (b); so if it is unclear what the result of negating a statement is, then even Kearns' formulation of the paradox is unclear.

Here, then, is my best attempt to clarify the idea of negation as applied to statements. Firstly, I'll assume henceforth that only statements that consist of assertions of propositions can be negated. Now, suppose we are given a statement S which consists in the assertion of some proposition *p*. Then, I propose, to negate S is to make a statement—following Kearns, let's give it the labels '~S' and 'the negation of S'. ~S, I suggest, is the speech act that consists of asserting the proposition ~*p*. Importantly, this way of characterizing statement negation makes room for Kearns' distinction between denying a statement and accepting its negation. (Though I do not see how one can define the act of accepting the negation of a statement except by identifying it with the act of negating the statement. But none of Kearns' claims relies on there being a distinction between these.) And my characterization of negation as applied to statements makes sense of Kearns' claim that (a) is the result of negating (b): (b) is the speech act consisting of asserting the proposition that (a) is true, and (a) is the speech act consisting of asserting the negation of that proposition, namely, the proposition that (a) is not true.

While Kearns might be able to accept my characterization of negation as applied to statements, it is worth noting that he would probably not endorse it as a definition. That is because for Kearns, speech acts, not propositions, are the most fundamental bearers of semantic properties, and therefore are the most fundamental things that can be negated, conjoined, disjoined, etc. By contrast, to take my characterization as a definition is to define negation as

applied to statements in terms of propositional negation. But whether it counts as a definition or merely a characterization, for the sake of clarity I'll henceforth understand Kearns' talk of negation as applied to statements in the way I have described.

With that description of negation for statements in hand, a Liar statement would presumably be a statement S which consists of asserting the proposition that S is not true.[94] (So, such a statement S would not be identical with its own negation. The negation of S would be the speech act consisting of the assertion of the proposition that *the proposition that S is not true* is not true.) That definition, at least, includes (a). Kearns' claim, again, is that there are no Liar statements, so on my characterization it would amount to the claim that there is no statement S consisting of the assertion of the proposition that S is not true.

Now to articulate my objections to Kearns' solution. The main problem is that the solution does not extend to Curry's paradox as adapted to statements.[95] Consider the following "Curry statement":

(c) If (c) is true then grass is purple

(Again, take '(c)' to denote not a sentence but a statement or putative statement.) Provided, as Kearns' discussion presupposes, that we can reason with statements as we do with sentences, reflection on (c) leads quickly to contradictions.[96] And whereas Kearns' solution was to posit a

[94] Kearns uses 'Liar statement' in several places, particularly to describe (a) above. But he never explicitly defines 'Liar statement'.

[95] See (Prior 1955) for the first articulation of the paradox in the form it has standardly assumed.

[96] Here is a simple proof:

1.	|(c) is true	(assumption)
2.	|If (c) is true then grass is purple	((T-out), (1))
3.	|Grass is purple	(conditional elimination, (1), (2))
4.	If (c) is true then grass is purple	(conditional introduction, (1)-(3))
5.	(c) is true	((T-in), (4))
6.	Grass is purple	(conditional elimination, (4), (5))

One might be concerned about the fact that this derivation invokes both conditional introduction and conditional elimination, and only the material conditional obeys both of these rules. Since the material conditional can be defined in terms of disjunction and negation, one might conclude that Curry's paradox invokes negation in disguise. However, there are versions of the paradoxical reasoning that do not appeal to both conditional introduction and

one-off aberration in the behavior of negation as applied to statements, the case of (c) appears to show that the conditional-as-applied-to-statements must also be subject to such aberrations, if paradox is to be avoided along anything like the lines Kearns proposes. Kearns could solve this problem by holding that both negation and the conditional exhibit aberrations. But rather than posit several different kinds of aberrations, a more efficient solution would be the analogue of aberrationism as applied to statements: posit some aberrations in the truth-attributing operator (that is, the operator that when attached to a statement S results in a statement consisting of the assertion that S is true).

A variation on this problem for Kearns arises in connection with sentences. Kearns would not deny that it makes sense to speak of sentences as being true or false.[97] But in that case, we can formulate Liar sentences such as sentence A and then ask after their alethic statuses. Initially, one simply worries that Kearns has no way to answer such questions, since his view addresses itself only to statements. However, the matter turns out to be more complex, since Kearns holds that sentences acquire their semantic features from the statements that they are conventionally used to make:

> It is [speech] acts which are the primary bearers of such semantic features as meaning and truth. Written and spoken expressions have syntactic features and can themselves be regarded as syntactic objects. Most expressions are conventionally used to perform acts with particular meanings; the meanings commonly assigned to expressions are the meanings of acts they are conventionally used to perform. However, these conventions are not the source of the meanings of meaningful acts, for the language user's intentions determine the meanings of his language acts. (p.32)

If Kearns is right that sentences acquire their semantic properties from the speech acts they are conventionally used to perform, then to see whether there are any Liar sentences (in my sense of

elimination. As (Beall 2013) illustrates, it suffices to have a conditional → for which the sentence ⌜(φ & (φ → ψ)) → ψ⌝ is true, for all sentences φ and ψ. Alternatively, it also suffices to have a conditional for which ⌜(φ → (φ → ψ)) → (φ → ψ)⌝ is true for all sentences φ and ψ.

[97] See his p.32.

the term[98]), one needs to examine the speech acts that such sentences and their constituents are conventionally used to perform. Kearns needs to show one of two things: either i) that there are no Liar sentences, or ii) that there are Liar sentences but that it follows from the way they or their constituents are conventionally used to perform speech acts that these sentences do not say what they appear to say.

(i) is quite implausible. Let us grant Kearns his view about how sentences acquire their semantic properties, and assume likewise for sub-sentential expressions—that is, we will assume that sub-sentential expressions acquire their semantic values from the way they are used in the performance of speech acts. Then it is clear that 'is not' is used to perform speech acts of negation and that 'true' is used to perform speech acts that consist of, among other things, attributing truth. Furthermore, the speech act performed by introducing the expression 'A' via a setoff and parentheses, as in

(A) A is not true

guarantees that in the ensuing discussion, 'A' will be used to perform speech acts which involve reference to the sentence displayed. Given the above facts about which speech acts the constituents of A are used to perform, it is hard to deny that these expressions *simpliciter* refer as we pre-theoretically expect. For that reason, then, it is hard to deny that there are Liar sentences (on my definition of 'Liar sentence'): sentences S that result from concatenating a negation symbol to a sentence whose grammatical subject is an expression that refers to S and whose grammatical predicate is an expression that refers to truth.

[98] Kearns uses 'Liar sentence' in a different way. For him, if a (putative) statement says of itself that it is not true, then any sentence(s) used to express that statement is a Liar sentence. (See p.51 and p.53 for evidence that this is how Kearns uses the term 'Liar sentence'.)

Given that there are Liar sentences, Kearns ought to embrace (ii). And given what Kearns wants to say about statements, his most natural move is to claim that the occurrence of 'not' in any Liar sentence undergoes an aberration, failing to refer to the standard truth function. However, similarly to the statement (c) above, there are Liar-like sentences that do not involve negation. The most prominent example is Curry sentences, such as 'If this sentence is true then 0 = 1'. As stated, Kearns' most natural approach to Liar sentences does not generalize to these other Liar-like sentences.

10.2. Self-Reference

Historically, blaming and then banning self-reference has been a common approach to the Liar paradox.[99] Why, one might wonder, is the expression in Liar sentences that refers to that sentence itself not a culprit? Does not self-reference show up in all seemingly Liar-ish paradoxes?

In fact, however, this second question has a straightforward, negative answer. (Yablo 1993) develops a seemingly Liar-ish paradox that does not rely on self-reference. The paradox consists of an infinite list of sentences:

(S_1) For all $n \geq 2$, S_n is not true.
(S_2) For all $n \geq 3$, S_n is not true.
…
(S_m) For all $n \geq m + 1$, S_n is not true.
…

A few moments' reflection on these sentences generates a paradox that is palpably Liar-ish. But it is clear that none of the sentences contains any expression referring to that sentence.

[99] For example, (Tarski 1935) can be read this way.

Still, however, while Yablo's paradox does not strictly speaking involve self-reference, it does involve a kind of circularity,[100] and one might take circularity of whatever form to be responsible for the paradox, self-reference being just a special case. Accordingly, then, one might take the culprit in each Liar-like sentence to be whatever expression is responsible for the circularity. In (Cook 2006), Roy Cook does an admirable job making this idea more precise. There, Cook defines four different kinds of *fixed points*. (For my purposes it doesn't matter exactly what fixed points are, or what Cook's particular kinds of fixed points are.) He shows that formalizing any of several familiar seemingly Liar-ish paradoxes, including Yablo's, will involve invoking the Diagonal Lemma (or some generalization thereof) to prove the existence of a fixed point of one of these four kinds. Thus, one might conclude, for any Liar-like sentence, whatever expression (putatively) makes the sentence count as a fixed point of one of these kinds is the culprit in that sentence. (In Liar sentences as I have defined them, the expression in question would be the one that putatively refers to the sentence itself.)

However, Cook also shows that one can get a palpably Liar-ish paradox without either the Diagonal Lemma or fixed points, by using infinitary conjunction:

(S_1') S_2' is not true & S_3' is not true &…
(S_2') S_3' is not true & S_4' is not true &…
…
(S_m') S_{m+1}' is not true & S_{m+2}' is not true &…
…

Thus, ubiquitous as fixed points are in formalizations of seemingly Liar-ish paradoxes, they are not strictly necessary for constructing a paradox that is palpably Liar-like.

[100] On some readings, each sentence begins with a universal quantifier that ranges over all natural numbers, followed by a conditional. E.g., 'For all n, if $n \geq 2$ then S_n is not true'. On these readings, there is circularity, in the sense that each sentence contains a quantifier that ranges over its own index.

One might respond here by doubting that Cook's paradox really is Liar-like, despite the fact that it strikes many philosophers as being so. Any genuinely Liar-like paradox, one might insist, must involve referential circularity of one kind or another. (In fact, a similar reaction to Yablo's paradox was available: one might have insisted that every paradox deserving of the title 'Liar-like' must involve self-reference.) But as it stands this response is dissatisfying. It amounts to simply legislating circularity into one's definition of 'Liar-ish', in defiance of the obvious similarities between Cook's paradox and Yablo's. The most striking such similarity is the fact that for each natural number i, (S_i) and (S_i') say very similar things; they have, for example, the same truth conditions, insofar as one can make sense of the notion of truth-conditions as applied to paradox-generating sentences. If Yablo's paradox counts as Liar-like, I contend, then surely Cook's Paradox should as well. (Note also, for what it is worth, that on my preferred way of defining *being Liar-like*, Cook's paradox counts: the reasoning involved makes central use of (T-intro) and (T-elim).)

Finally, whether or not Cook's paradox is Liar-like, it is a virtue of approaches to the Liar which target alethic expressions that they can be applied to it. By contrast, approaches that target fixed points cannot apply to Cook's paradox. Similarly, solutions that target alethic expressions can be applied to Yablo's paradox, whereas solutions that target self-reference strictly so-called cannot. So, approaches to the Liar paradox that target alethic expressions have the added benefit of applying to other paradoxes, whether or not these officially count as Liar-like.

I will end my discussion of alternative approaches on this optimistic note. In the next section, I will provide some support for (Aberrations) and (Determined) that is independent of considerations having to do with the Liar paradox.

11. Independent Motivations for Reference by Occurrences and (Determined)

11.1. Reference by Occurrences

In this subsection, I will argue that quite generally it makes sense to speak of an occurrence of an expression as having a referent. Then, in Section 11.2, I will argue that for any sentence, what Russellian proposition(s) it expresses, expresses the negation of, etc. is determined by the reference of the occurrences of the expressions it contains, rather than the expressions *simpliciter*.

To begin with, it helps to note that the notion of reference by occurrences has a long and distinguished history in analytic philosophy. (Frege 1892) holds that when a word occurs in the that-phrase of a propositional attitude description (such as 'Sylvia believes that it will rain tomorrow'), the word refers to what is usually its sense (its *customary sense*), rather than what is usually its referent (its *customary referent*). This amounts to the claim that the occurrence of the word in the that-phrase differs in reference from its other occurrences, or, if one prefers, from the word *simpliciter*. Moreover, it is worth noting, in such cases it is reference by the occurrence in the that-phrase that contributes to the truth-conditions of the attitude-attributing sentence, not reference by the word *simpliciter* or by its other occurrences. Thus, on Frege's account, the reason that the sentences 'Lois believes that Superman is Clark Kent' and 'Lois believes that Superman is Superman' differ in truth value is that the occurrences of 'Superman' refer to the customary sense of 'Superman' and the occurrence of 'Clark Kent' refers to the customary sense of 'Clark Kent'.

Even if Frege endorsed the notion of reference by occurrences, however, it will help to have some less controversial examples. Perhaps the most compelling case is that of pronouns. A pronoun *simpliciter* never refers, but in general, some of its occurrences do. E.g., the word 'him'

does not refer to any particular individual, but in the sentence 'Michelle Obama plays with Bo by throwing sticks for him to fetch', the occurrence of 'him' refers to Bo. (That is, relative to contexts in which 'Bo' is understood to refer to the Obama family's dog.) Pro-adjectives are similar. The word 'so' (when used as a proadjective, as well as more generally) does not refer *simpliciter*. However, in the sentence 'The voters wanted Barack Obama to be judicious, and he is so', the occurrence of the word 'so' refers to judiciousness. Similar claims can be made about other *pro-forms*—pro-verbs, pro-adverbs, etc.

Another convincing example of reference by occurrences arises when an ambiguity is resolved by a word's component context. As (Pustejovsky 1996) explains,

> there are some cases of…ambiguity that do not require context [of utterance] and pragmatic information for disambiguation, so much as the disambiguation that comes by virtue of the predication relation in the sentence. For example, in (15) below, the appropriate sense for the noun club is arrived at by virtue of sortal knowledge of the NP appearing in the inverted subject position…
>
> (15) a. Nadia's favorite club is the five-iron.
> b. Nadia's favorite club is the Carlton. (p.30)

If indeed the underlined occurrences in (a) and (b) are occurrences of a single word, 'club', then we have here a case in which the reference of the occurrence of a word depends on the component context. However, some philosophers and linguists are reluctant to regard the occurrences underlined above as occurrences of the same word.[101] So, it is worth noting that Pustejovsky's remarks also hold in the more straightforward case of *polysemy*, which is traditionally understood as the special case of ambiguity in which the distinct meanings involved

[101] If we identify words with lexical entries, then this denial sits naturally with the conception of ambiguity as involving two lexical entries which are associated with a single grapheme, phoneme, or grapheme-phoneme pair, as opposed to a single lexical entry with multiple meanings listed within it. A consideration that supports taking the occurrences underlined above to be occurrences of different words is that one might understand uses of 'club' as in (15)a but not as in (15)b, and such a failure need not be due to stupidity or inattentiveness—that is, it need not be a failure of performance. This suggests that the particular psychological competences involved in understanding these sentences are different; and it is reasonable to hold that words which are understood via the exercise of different psychological competences are different words. I am grateful to Harold Hodes for pointing this out.

are closely related.[102] In such cases, it is less controversial that it is the same word which occurs. For example, from his p.53:

(48) a. Mary regretted publishing the article in *Illustrated Semantics*.[103]
b. Mary regretted the article in *Illustrated Semantics*.

It is uncontroversial that the underlined occurrences above are occurrences of the same word.[104] But in (48)a the occurrence of 'regretted' refers to a relation between persons and action types, whereas in (48)b it refers to a relation between persons and objects. Because the disambiguation is achieved by the component context, it is, in effect, the occurrences that refer; and it is their reference that contributes to the truth-conditions of the respective sentences.

Having emphasized that polysemy furnishes us with plausible examples of reference by occurrences, I hasten to add that aberrationism is not committed to the claim that 'true' and other alethic expressions are polysemous. The most striking difference between one-off aberrations and polysemy is that polysemous words have multiple (though closely related) senses as well as multiple referents. By contrast, for example, although key occurrences of 'true' in Liar-like sentences fail to refer to truth, the word 'true' has only one sense.[105] Another difference is that polysemy is not "one-off"—that is, while the reference of a polysemous expression can depend on its component context, for each one of its distinct referents there is a wide range of component contexts that can trigger reference to that referent. By contrast, alethic expressions

[102] If ambiguity is characterized by distinct lexical entries associated with the same grapheme, phoneme, or grapheme-phoneme pair, then polysemy is characterized by a single lexical entry with multiple meanings listed within it. If ambiguity is characterized by a single lexical entry with multiple distinct meanings listed within it, then polysemy is the special case of ambiguity in which these meanings are closely related. According to (Sennet 2016), polysemy has traditionally been understood in the latter way.
[103] I have changed Pustejovsky's example slightly. His (48)a involves a propositional attitude construction, which complicates matters unnecessarily.
[104] Returning, for example, to the question of the competences involved in understanding these occurrences, it is plausible that understanding (48)a and (48)b involves the same competence, and that understanding one but not the other is indicative of a failure of performance, not of competence. So, if we individuate words via the competences involved in understanding them, then we will count the underlined occurrences in (48) as occurrences of the same word.
[105] Or, at least, aberrationists need not claim that it has multiple senses.

undergo aberrations only in a very narrow range of component contexts: when they occur as key occurrences in Liar-like sentences.

Before moving on, I will discuss two more controversial examples of (what I take to be) reference by occurrences. (Dummett 1973) observes that "[e]ven when a sequence of words has a sense, taken as a whole, and the words composing it occur in succession in a sentence, they may not, in that sentence, compose that or any other phrase having sense as a whole" (p.34). He gives two examples:

(Tore) The man wearing my coat tore up the letter

(Killed) Was what Henry killed a man?

Again, Dummett's claim is that the occurrences of the underlined phrases fail to have senses, despite the fact that the phrases of which they are occurrences, considered as expressions *simpliciter,* do have senses. But as with sense, so with reference—or, if one prefers, semantic value. The semantic value of the phrase 'my coat tore' *simpliciter* is a proposition (in many contexts of utterance). But its occurrence in the sentence (Tore) does not have a semantic value. Similarly, the phrase 'killed a man' *simpliciter* refers to an action-type. But its occurrence in (Killed) does not refer. These examples, then, are cases in which the semantic value of an occurrence of an expression (if there is one in these cases) can come apart from that of the expression *simpliciter*.[106]

Finally, consider the case of names. In (Jeshion 2015), Robin Jeshion argues that as they occur in most sentences, names designate individuals. However, she also holds that as they occur in certain kinds of sentences, names function as count nouns, designating properties.[107] That is to

[106] Thanks to Harold Hodes for these examples.
[107] Chomsky makes a similar claim in (Chomsky 1965) p.100.

say, for many names, some occurrences of the name refer to individuals, while others refer to properties. Here are some of Jeshion's examples:

[1] Alfred studies in Princeton.

[17] Some Alfreds are crazy; some are sane.

On Jeshion's view, the occurrence of 'Alfred' in [1] refers to an individual, whereas the occurrence in [17] refers to a property, *being called 'Alfred'*. The point is that if Jeshion is right, then in general the reference of a name is sensitive to its component context, and it makes sense to speak of an occurrence of a name as referring.

In light of all these examples, I now claim that even for an expression that is usually thought to refer *simpliciter*, it can also make sense to speak of some occurrences of this expression as referring. For instance, in the sentence 'Hillary Clinton spoke in the 2015 Democratic debate', the occurrence of the name 'Hillary Clinton' refers to Hillary Clinton. Admittedly, in the case of an expression *simpliciter* that refers to something, speaking of reference by the occurrences of the expression is usually unnecessary, because these usually refer to the referent of the expression *simpliciter*. Still, necessary or not, speaking of reference by the occurrences is harmless and makes perfectly good sense. To bring the point home, it is worth noting that people have made similar arguments concerning the relativization of reference to contexts of utterance. It is now utterly standard in semantic theory to speak of expressions as referring relative to contexts of utterance. On this way of speaking, it is trivial to treat a context-insensitive expression as referring to the same thing relative to all (or nearly all) contexts of utterance, rather than as referring *simpliciter*.

11.2. The Independent Case for (Determined)

Even if one were to grant that occurrences can refer and that they can differ in reference from the expressions of which they are occurrences (that is, that (Aberrations) is true), one might doubt that this helps with avoiding paradox. For example, one might think that whether or not A expresses the negation of the Russellian proposition <A, truth> is determined by what 'true' and 'A' refer to, and what any of their occurrences refer to is simply irrelevant. And one might make similar claims for any other sentence: quite generally, what (if any) Russellian propositions a sentence expresses, expresses the negation of, the conjunction of, etc., one might argue, depends on the reference of its expressions *simpliciter*, not of their occurrences. Thus, one might hold that (Determined) is false. (Recall: (Determined) is the claim that when an occurrence of an expression in a Liar-like sentence differs in reference from the expression of which it is an occurrence, what Russellian propositions (if any) the sentence expresses, expresses the negation of, the conjunction of, etc., is determined by what the occurrence refers to, rather than by what the expression simpliciter refers to.)

One might dodge this worry by maintaining that strictly speaking expressions as such do not refer. On that view, an expression *simpliciter* can refer to something only in the sense that most of its occurrences refer to that thing. For any sentence which contains an expression that refers in this sense, it is implausible that the referential behavior of the occurrences of the expression outside this sentence could have more influence on what Russellian propositions the sentence expresses, expresses the negation of, the conjunction of, etc. than the referential behavior of the occurrence(s) in the sentence. Thus, if expressions *simpliciter* do not strictly speaking refer, then (Determined) seems likely to come out true.

However, the claim that expressions as such do not refer is controversial, and aberrationism does not require that it be true. Instead, I propose to defend (Determined) by repeating my argumentative strategy from the last subsection, that is, by generalizing from examples in which it is reference by occurrences that determines what sentences say. More specifically, I now make the following claim:

(Generalized Determined) What Russellian propositions (if any) a sentence expresses, expresses the negation of, the conjunction of, etc., is determined by what the occurrences of its constituent expressions refer to, rather than by what the expressions themselves refer to.

To see why (Generalized Determined) is true, recall the case of pro-forms: again, a pro-form *simpliciter* never refers, but some of its occurrences do. E.g., the word 'him' does not refer to anything, but (relative to a context of use establishes that the name 'Bo' refers to the Obamas' dog, Bo,) the occurrence of 'him' in the sentence 'Michelle Obama plays with Bo by throwing sticks for him to fetch' refers to Bo. Now, relative to such a context, this example sentence attributes a relation to the pair <Michelle Obama, Bo>. And which pair this sentence attributes this relation to must be determined (in part) by facts about reference, in particular facts about the reference of 'him' or its occurrence in the sentence. But since the pronoun 'him' does not refer *simpliciter*, there is only one candidate referential relation here: the one between the occurrence of 'him' in the sentence and Bo. Thus the fact that this occurrence refers to Bo has to be what (in part) determines to which pair the sentence attributes the relation.

Pro-adjectives provide a similar example. E.g., the word 'so' does not refer to anything, but in the sentence 'The voters wanted Barack Obama to be judicious, and he is so', the occurrence of 'so' refers to judiciousness. Now, this sentence attributes a property to Barack Obama. And which property the sentence attributes must be determined (in part) by facts about

reference, in particular facts about the reference of 'so'. But since the pro-adjective 'so' does not refer *simpliciter*, there is only one candidate referential relation here: the one between the occurrence of 'so' in the sentence and Barack Obama. Thus, the fact that this occurrence refers to Barack Obama has to be what (in part) determines to which pair the sentence attributes the relation.

The same points hold for the other examples we have seen. Even in contexts in which the phrase 'my coat tore' expresses a proposition, it is not by expressing this proposition that this phrase contributes to the determination of what the sentence 'The man wearing my coat tore up the letter' says. Rather, the occurrence of the phrase in that sentence arguably fails to refer, and surely enough, the sentence expresses no proposition involving the tearing of a coat. Similarly with the sentence (Killed) from above. In all these cases, then, it is the semantic behavior of the occurrence, not the expression *simpliciter*, that matters for what proposition the sentence expresses. Generalizing on these examples, we arrive at (Generalized Determined).

Admittedly, in the case of sentences that contain referring expressions such as names and predicates, it will usually be harmless to take reference by the expression *simpliciter* to determine what Russellian propositions the sentence expresses, expresses the negation of, etc. That is because an occurrence of an expression that refers to something nearly always refers to that same thing. But the point concerns the comparative naturalness of rival semantic theories, not the ways that these theories might be put into practice. The point is that it would be oddly disjunctive to regard reference by occurrences as being the thing that contributes to what Russellian propositions a sentence expresses, expresses the negation of, etc. in the case of the kinds of examples we have seen, but to regard reference by expressions *simpliciter* as making this contribution in other, more familiar cases. Rather, the more natural view is that the same

thing—viz., reference by occurrences—makes this contribution in both settings, but that this contribution is masked by the fact that most referring expressions agree in reference with nearly all of their occurrences.

11.2.1. Making Mysteries

Another concern about (Generalized Determined) is that together with (Aberrations) it makes a mystery out of the fact that for any expression that refers to something, nearly all of the occurrences of that expression refer to that same thing. Pre-theoretically, one might have explained this by endorsing (Explanation):

(Explanation)	An occurrence of a referring expression in some sense gets its referent (if any) from the expression of which it is an occurrence.

Now, "gets its referent from" is obscure. But there is a particularly straightforward way of making it precise:

(Expressions Fundamental)	An occurrence of a referring expression can refer to an object x only in the sense of being an occurrence of an expression that refers to x.

Plainly, however, (Expressions Fundamental) contradicts (Aberrations), and so is unavailable to aberrationists.

Still, it is worth observing that there may be ways of clarifying (Explanation) that are compatible with aberrationism. Admittedly, this is not immediately obvious; (Generalized Determined) makes it hard to see how anything like (Explanation) could be true. If an occurrence of a referring expression in some sense gets its referent from the expression, then how come it is reference by the occurrences of the expression in a sentence that contributes to what Russellian

propositions the sentence expresses, expresses the negation of, etc., rather than reference by the expression itself?

My answer lies in the Lewisian claims about reference that I made in Section 3.3.2. As I explained there, for some kinds of expressions, in particular, expressions that are introduced via a theory, an expression of this kind refers to that thing (if there is one) that comes closest to respecting the principles that govern the expression. (Again, as I defined respect in Section 3.3.2, an object respects a principle for an expression if, were the expression to refer to that object, the principle comes out true, or truth-preserving. See Appendix A for an alternative definition.) It follows from these claims that any occurrence of such an expression that has a referent has that same referent, unless that would lead to a contradiction. In that event, the occurrence refers to that thing (if there is one) which is next in line for coming closest to respecting the principles for the expression. That, I propose, is the sense in which (Explanation) is true; an occurrence of a theoretical term acquires an object *o* (if any) as its referent in virtue of the fact that the term is governed by certain principles P, and the fact that *o* comes as close as possible to respecting P, or is next in line. So, one can correctly say that the occurrence acquires its referent by being an occurrence of an expression that has P as a principle.

Again, the point here is that I can tell a principled story about why there is generally agreement in reference between a referring expression and its occurrences. Indeed, as one might have desired, that story can involve appealing to (Explanation), provided that (Explanation) is clarified in the way that I propose.

12. Concluding Remarks

In this essay, I have argued that no Liar-like sentence says what it would have to say in order for the reasoning in the associated paradox to be legitimate. In particular, sentence A from the beginning fails to express the negation of the Russellian proposition <A, truth>, and similarly for all other Liar sentences. Assuming that the word 'true' refers to truth and the name 'A' as used in this essay refers to A, I showed that for A to fail to say of itself that it is not true, the following must hold:

(Aberrations)	For any Liar-like sentence, the key occurrences of its alethic expressions differ in reference from the expressions of which they are occurrences.

and

(Determined)	When an occurrence of an expression in a Liar-like sentence differs in reference from the expression of which it is an occurrence, what Russellian propositions (if any) the sentence expresses, expresses the negation of, the conjunction of, etc. is determined by what the occurrence refers to, rather than by what the expression *simpliciter* refers to.

One contribution of this paper is simply to articulate and endorse (Aberrations) and (Determined) explicitly. While I have argued that (Smith 2006) must rely on these claims, open endorsement of them, and concerted hashing out of their consequences, is, as far as I know, unique to me. In addition, in Section 11 I argued that these claims have some plausibility quite independently of Liar-like paradoxes.

Moreover, the availability of moderate aberrationism as an alternative to radical aberrationism (Section 3) is important. Radical approaches would have us give up the plausible idea that what our words refer to is determined in principled ways by our behavior, mental states and physical environment. The availability of moderate views shows that one can diagnose and solve the Liar-like paradoxes without giving up this idea. One especially appealing way to develop this view, I argued, is to hold that the key occurrences of the alethic expressions in any

Liar-like sentence acquire their referents via the same Lewisian pattern that determines reference for theoretical terms generally. In addition to allowing for the retention of Semantic Supervenience and Semantic Regularity, the availability of this view shows that moderate aberrationists have the tools to allow that Liar-like sentences come quite close to saying what they appear to say, and thus to do a significant measure of justice to our pre-theoretical impressions about these sentences.

While moderate versions of aberrationism are superior to their radical cousins, both are superior to the many alternatives in the literature. The no-proposition view can be cast as a distinct approach, but it soon emerges that the best way to develop it is by embracing aberrationism. For another thing, aberrationist approaches are not plagued by the incompleteness that afflicts views along the lines of (Kripke 1975). They are also preferable to contextualist views, such as that of Michael Glanzberg. I questioned Glanzberg's claim that the logical forms of Liar sentences involve quantification over propositions, and found a serious problem with his account of context-shifts in the Liar reasoning. However, I saw insight in his observation that 'true' does not pre-theoretically seem to be context-sensitive or the pass ordinary tests for context-sensitivity. I argued that these observations cause no problems for aberrationists, since aberrationists hold that the behavior of 'true' in Liar-like sentences is a one-off aberration from its context-insensitive behavior everywhere else. On the other hand, views on which 'true' contains an indexical element, such as Burge's, are vulnerable to Glanzberg's objection.

After discussing Burge's view, I turned to that of Keith Simmons, which runs into trouble when one tries to quantify over contexts of evaluation. Simmons has to say contradictory things about whether the uses of true in certain seemingly paradoxical sentences are context-independent. This contrasted with aberrationism, which gives the unproblematic diagnosis that

the occurrences of ‘true’ in the relevant sentences fail to refer to truth. More generally, for aberrationism, unlike for Simmons’ view, no problem is raised by the possibility of quantifying over all contexts—be they contexts of evaluation or component contexts.

Next, I examined Haim Gaifman’s view, which focuses on sentence tokens that comment on the semantic status of other sentence tokens. I concluded that Gaifman’s view may be good as far as it goes, but it does not address sentences (types or tokens) that comment on the status of sentence-types. Thus, as stated it fails to solve the problem that centrally concerns me in the essay; and it is not clear how Gaifman’s remarks about the examples that he presents can be adapted so as to solve this problem.

Next I criticized the prosentential views. My objection was that of (Wilson 1990): the principal claim made by these views—that sentences involving the word ‘true’ are prosentences—violates the linguistic data about English; thus, it is inferior to alternative approaches to the Liar that do better on this score, such as my own. For people who endorse the pre-theoretically plausible claim that truth exists, the inferiority of prosententialism as an approach to the Liar paradox should be welcome news.

I also argued that aberrationism is superior to wholesale indeterminism, although the two views are closely related. Aberrationism’s principal advantage is that it restricts its allegations of indeterminacy to the occurrences that are implicated in Liar-like paradoxes, rather than holding that all occurrences of ‘true’ are indeterminate. Another advantage is that aberrationism allows us to retain the pre-theoretically appealing idea that there is a single property that is determinately identical with truth.

The last family of views I criticized were those that posit some sort of one-off aberration, but which target something other than alethic expressions. I discussed (Kearns 2007), which

formulates a Liar paradox in terms of "statements"—speech acts in which propositions are asserted, denied, conjoined, etc., and which targets negation as applied to statements. I objected that Kearns' view is ill-equipped to handle Curry's paradox as formulated for statements, and, for similar reasons, Curry's paradox as formulated for sentences. In that section I also discussed views that posit one-off aberrations in the expressions that make for self-reference in Liar-like sentences of the most familiar kinds. Against these views, I argued that one can construct a palpably Liar-ish paradox, due to (Cook 2006), that does not involve self-reference. By a familiar methodological principle (applied also to Kearns' view in connection with Curry's paradox), the expressions (if any) that make for self-reference in a Liar-like sentence are not the appropriate ones to target for aberrant behavior.

Overall, the greatest strength of aberrationism lies in the balance that it achieves between two conflicting requirements: on the one hand, explaining why the Liar reasoning is so compelling, and, on the other hand, limiting the semantic abnormalities posited to what is necessary for diagnosing these sentences and avoiding paradox. I satisfy the first requirement by positing one-off aberrations. Except when they occur in Liar sentences, the expressions implicated in the Liar paradox behave exactly as they appear to behave. It is thus only natural for us to expect them to behave in these ways in Liar sentences as well. I resist paradox by denying that these expressions behave in these ways when they occur in Liar sentences. Still, aberrationists are able to satisfy the second, limiting requirement by holding that the aberrations are rare: only when they occur in Liar-like sentences do alethic expressions fail to refer as expected; this is a one-off aberration from their behavior everywhere else. Thus, aberrationists avoid making empirically unsubstantiated claims about how these expressions behave in unproblematic sentences.

Appendix A: Respect and Explanatory Circles

In Section 3.3.2, I described reference determination in terms of respect, and then described respect in terms of subjunctive conditions that involve reference. One might therefore worry that, insofar as they were supposed to explain anything about how reference is determined, these descriptions amount to an explanatory circle. However, one should keep in mind that my explanatory aim in invoking David Lewis' ideas about reference is rather modest. I am not purporting to define reference in entirely different terms, or demonstrate in full detail how the facts about what refers to what are determined from the facts about how people use words, though my remarks would naturally fit in as part of such a story. (Note how what I describe in (i)-(iv) contrasts with so-called "deflationary" views according to which all that can be said about the nature of reference is that 'dog' refers to dogs, 'rain' refers to rain, and so on.[108])

Still, one might hope to accomplish more than I do by way of explaining reference in independent terms. Indeed, perhaps Lewis intended to do more in (Lewis 1970). In particular, rather than defining it in terms of reference, one might have taken Lewis to be identifying respect with satisfaction-in-the-model-theoretic-sense—henceforth $\models$—and then characterizing reference in terms of respect-so-understood. Lewis' characterization could not count as a reduction or even partial reduction of reference, since it describes reference for terms that are introduced by a theory in terms of reference for the other terms that feature in the theory.[109] But still, one might find it more illuminating to characterize one expression's reference in terms of another's, as Lewis does, than to characterize it in terms of that expression's own reference in

[108] For an example of such a view, see (Horwich 1997). (Field 1994) presents an analogous view concerning truth. Compare also (Brandom 1984).

[109] He writes: "Let us assume that [these other terms] have conventionally established interpretations, already well-known to us" (p.429). Here 'interpretation' arguably refers to the term's referent.

other possible worlds, as I do in Section 3.3.2. This is what a definition of the respect relation from (i)-(iv) in model-theoretic terms—that is, in terms of $\models$—could hope to accomplish. I will now show that someone who insists on such a definition can adopt an aberrationist approach to the Liar paradox. However, it will turn out that she must take more assumptions on board than she would have to if she understood respect as I do, in terms of subjunctive conditions on reference.

Note first that $\models$ is a relation that holds between pairs $<M, v>$ and formulas φ, where *M* is a model and *v* a variable assignment. However, in (i)-(iv) I speak of respect as a relation between a property and a principle, where the latter could be an inference rule. Here, then, is how a definition of respect in terms of $\models$ would go. Let *e* be an expression and P an inference rule that is a principle for *e*. (I'll ignore cases in which the principle is a sentence, since my concern is with 'satisfies' and (Sat-intro) and (Sat-elim) are inference rules.) Let N be an instance of P. If P is an introduction rule, let Conclusion(X) be the formula that results from substituting all occurrences of *e* that are introduced via N with occurrences of a free variable X of the appropriate type (first order if *e* is a singular term, second order if *e* is a predicate). And if P is an elimination rule, let Conclusion(X) be the conclusion of N, unchanged. Similarly, if P is an elimination rule, let $\text{Premise}_1(X),\ldots,\text{Premise}_n(X)$ be the formulas that result from substituting all occurrences of *e* that were eliminated in N with occurrences of a free variable X of the appropriate type (first order if *e* is a singular term, second order if *e* is a predicate). And if P is an introduction rule, let $\text{Premise}_1(X),\ldots,\text{Premise}_n(X)$ be N's premises, unchanged. Let a *standard model* be a model whose interpretation function maps *e** to its referent[110] for each expression *e** in N other than *e*. Then *y respects*$_{MT}$ N if and only if: For every standard model *M* and every

[110] Again, recall Lewis, who writes "conventionally established interpretations" (p.429). Keep in mind that we are characterizing reference of one expression in terms of the reference of others that are linked to it by a theory.

variable assignment *v* that maps X to *y*, if <*M*, *v*> ⊨ Premise_1(X), and …, and <*M*, *v*> ⊨ Premise_n(X), then <*M*, *v*> ⊨ Conclusion(X). The idea on offer here is that the relation of respect which ought to feature in (i)-(iv) is respect_{MT}, defined in model-theoretic terms.

So, let us see how (i)-(iv) fare once they are understood in terms of respect_{MT}. First, consider my claim that the word 'true' *simpliciter* refers to truth. Plugging respect_{MT} into (i)-(iv), we get the claim that 'true' refers to that thing, if there is any, which comes closest to respecting_{MT} (T-intro) and (T-elim). So, we must figure out whether truth respects_{MT} (T-intro) and (T-elim). The answer depends in part on whether truth respects_{MT} the following instances (and likewise, others that involve Liar-like sentences, but for brevity I'll leave those to another occasion):[111]

A is not true
_____ (T-intro)
'A is not true' is true

And:

'A is not true' is true
_____ (T-elim)
A is not true

Let's start with the instance of (T-intro) above. Suppose that the occurrence of 'true' in the conclusion is replaced by a free second order variable, X. That gives us:

A is not true
_____ (T-intro)
'A is not true' is X

Now, suppose we are given a standard model *M* and a variable assignment *v* that maps X to truth. We want to know whether or not the following conditional is true:

(Sat_I) If <*M*, *v*> ⊨ 'A is not true' then <*M*, *v*> ⊨ ⌜'A is not true' is X⌝

[111] Here I simply assume that truth is well-behaved on instances of (T-intro) that do not involve paradoxical sentences.

How should we evaluate this conditional? Of course, if we assume the antecedent, we can disprove the consequent. But on the other hand, for the conditional to be false the antecedent must be true and the consequent false. However, the status of both the antecedent and the consequent just boils down to the alethic status of A.

More precisely: for (Sat_I) to be false, it must be that <*M*, *v*> ⊨ ‘A is not true’ but not <*M*, *v*> ⊨ ⌜‘A is not true’ is X⌝. Fix nonempty set U and interpretation function V such that *M* = <U, V>. Is the antecedent of (Sat_I) true? <*M*, *v*> ⊨ ‘A is not true’ if and only if V(‘Sentence A’) is not an element of V(‘true’) (assuming that the latter is a set!). By the hypothesis that *M* is a standard model, V(‘A’) = A and V(‘true’) = truth. So, (fudging the distinction between truth and the set of true sentences, and assuming the latter exists,[112]) <*M*, *v*> ⊨ ‘A is not true’ if and only if A is not true. The same situation arises for the question of whether <*M*, *v*> ⊨ ⌜‘A is not true’ is X⌝. <*M*, *v*> ⊨ ⌜‘A is not true’ is X⌝ if and only if V(‘A is not true’) is in *v*(X) (assuming the latter to be a set!). But by the assumption that *M* is a standard model, V(‘A is not true’) is A, and by our choice of *v*, *v*(X) is truth. Thus <*M*, *v*> ⊨ ⌜‘A is not true’ is X⌝ if and only if A is true. Looking ahead, it is easy to see that the same situation arises for the question of whether ‘true’ respects (T-elim). The conditional that is relevant to that question is simply the converse of (Sat_I).

So, whether or not truth $respects_{MT}$ (T-intro) and (T-elim) depends on the status of A. That means that my justification for the claim that A is indeterminate cannot rely on the claim that truth $respects_{MT}$ (T-intro) and (T-elim). But recall, my view was that A is indeterminate *because* its occurrence of ‘true’ is indeterminate in reference as between ascending truth and descending truth. And in turn I justified that claim by arguing that, after truth, ascending truth

[112] The model-theoretic definition of satisfaction is applicable only when the referents of the predicates in the language being considered are sets. If there is no set of all and only the true sentences, then it does not make sense to speak of truth’s $respecting_{MT}$ the principles for any predicate. This is a significant challenge for those who wish to identify respect with $respect_{MT}$ in settings that involve the word ‘true’.

and descending truth are tied for second place with respect to respecting (T-intro) and (T-elim). The most natural and convincing strategy for justifying that latter claim would involve showing that while truth respects these rules, neither ascending truth nor descending truth does. However, that strategy has now been ruled out, since it begs the question as to the status of A. I am not sure what other strategy might work.

Given this situation, it seems that a fan of my view who insists on understanding my talk of respect as talk of respect_{MT} has no choice but to adopt the claim that A is indeterminate as an additional assumption or as justified on independent grounds, rather than proving this claim from the thesis that truth comes closer than ascending truth and descending truth to respecting (T-intro) and (T-elim). While such a position is less satisfying than the one I set out to defend, I will now show that it is coherent nonetheless. That is, I will argue that if we plug respect_{MT} into (i)-(iv) and assume that A is indeterminate, then we can verify that truth comes closer than ascending truth and descending truth come to respecting (T-intro) and (T-elim). It then follows that the occurrence of 'true' in A is indeterminate in reference as between ascending truth and descending truth (making A indeterminate, consistent with our assumption). Thus, we will be able to conclude, a version of my view is available to theorists who insist on understanding my talk of respect in terms of respect_{MT}. That version is the package consisting of (Aberrations), (Determined), the claim that A is indeterminate, and the claims (i)-(iv) with 'respects' therein understood as referring to respect_{MT}. I will also show that someone who adopts this view would do best to hold that the conditional used to define respect_{MT} (e.g., the one that features in (Sat_I)) obeys the Łukasiewicz truth table.

Suppose that A is indeterminate. Then it is indeterminate whether A is true and indeterminate whether A is not true. Thus both the antecedent and the consequent of (Sat_I) are

indeterminate. We now want to know whether or not truth respects$_{MT}$ (T-intro) and (T-elim). That depends on the truth of the conditional (Sat_I) and its converse, *(Sat_E)*.[113] For both these conditionals, both the antecedent and the consequent are indeterminate. Whether that makes these conditionals true depends on what the correct logic of indeterminacy is. According to the Łukasiewicz truth table for three-valued propositional logic, a conditional both of whose constituent sentences are indeterminate is true. In that case, both (Sat_I) and (Sat_E) are true, and thus truth respects (T-intro) and (T-elim). Now, in Section 3.3.3 I showed that ascending truth fails to respect (T-elim) and descending truth fails to respect (T-intro), using the notion of respect described in terms of reference. But these arguments can easily be recast in terms of respect$_{MT}$.[114] Therefore truth comes closer than either ascending truth or descending truth to respecting both (T-intro) and (T-elim).

However, Łukasiewicz's logic of indeterminacy is not the only game in town. In Kleene's Weak truth table and his Strong truth table, a conditional both of whose constituent

---

[113] That is:

(Sat_E) If <*M*, *v*> ⊨ ⌜'A is not true' is X' then <*M*, *v*> ⊨ 'A is not true'

[114] Let us figure out whether d-truth satisfies (T-intro). Start with the sentence

(S_d) S_d is not d-true

Now consider the following inference:

S_d is not d-true

'S_d is not d-true' is true

I will show that d-truth fails to respect this instance of (T-intro). To show this, suppose we are given a second order variable X, and we replace the occurrence of 'true' introduced in this instance with X as follows:

S_d is not d-true

'S_d is not d-true' is X

Now suppose we are given a standard model *M* and a variable assignment *v*' that maps X to d-truth. Fix nonempty set U and interpretation function V such that *M* = <U, V>. Then since *M* is standard, V('S_d') = S_d and V('d-true') = d-truth. Recall from the definition of d-truth that every sentence S that expresses the negation of the Russellian proposition <S, d-truth> to itself is not d-true. Thus V('S_d') is not an element of V('d-true'), and so <*M*, *v*'> ⊨ 'S_d is not d-true'. However, because 'S_d is not d-true' expresses the negation of the Russellian proposition <S_d, d-truth>, it is not d-true. Thus it is not an element of *v*'(X), which is d-truth. Thus V('S_d is not d-true') is not an element of V(X) (which is *v*'(X)), so it is not the case that <*M*, *v*'> ⊨ ⌜S_d is not d-true' is X'. Thus d-truth fails to respect the displayed instance of (T-intro), and so fails to respect (T-intro). One can make a parallel argument concerning ascending truth and (T-elim).

sentences are indeterminate is itself indeterminate. That would make (Sat_I) and (Sat_E) are both indeterminate. Still, even in that case it is not obviously untenable that truth comes closer than ascending truth and descending truth to respecting both (T-intro) and (T-elim). The idea would be that, so to speak, two indeterminates comes closer than a definite respect and a definite non-respect. That is, by being indeterminate with respect to both rules, truth comes closer to respecting them both than either ascending truth or descending truth comes, given that each of these (latter) properties respects one rule but not the other. While I am unsure how to defend this view, at the same time it is not obviously untenable.

At any rate, it is worth noting that Łukasiewicz's truth tables have some appealing features. In particular, for Łukasiewicz the sentence ˹If P then P˺ comes out true for all sentences P that are indeterminate in truth value. Contrast Kleene's tables, for which ˹If P then P˺ comes out indeterminate if P itself is.[115]

[115] See (Sider 2010), p.77 footnote 29.

REFERENCES

1. Armour-Garb, Bradley, and Woodbridge, James. (2012) “The Story About Propositions.” *Nous*, Vol.46, No. 4, pp.635-674.
2. Balaguer, Mark. (1998) “Attitudes Without Propositions.” *Philosophy and Phenomenological Research*, Vol. 58, No. 4, pp.805-826.
3. Beall, J.C. “Curry’s Paradox”, *The Stanford Encyclopedia of Philosophy* (Spring 2013 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/spr2013/entries/curry-paradox/>.
4. Beebe, James R. (2015) “Prosentential Theory of Truth,” *Internet Encyclopedia of Philosophy*, James Fieser and Bradley Dowden (eds), URL = <http://www.iep.utm.edu/truthpro/>.
5. Braddon-Mitchell, David, and Nola, Robert. (1997) “Ramsification and Glymour’s Counterexample” *Analysis*, Vol. 57, No. 3, pp. 167-169.
6. Brandom, Robert. (1984) “Reference Explained Away.” *The Journal of Philosophy*, Vol. 81, No. 9, pp. 469-492.
7. Brandom, Robert. (1994) *Making it Explicit*. Harvard University Press, Cambridge, MA.
8. Burge, Tyler. (1979) “Semantical Paradox.” *The Journal of Philosophy*, Vol. 76, No. 4, pp. 169-198.
1. Burgess, Alexis. (2014) “Keeping ‘True’: A Case Study in Conceptual Ethics.” *Inquiry*, Vol. 57, Nos. 5-6, pp.580-606.
2. Chomsky, Noam. (1965). *Aspects of the Theory of Syntax*. Cambridge: MIT Press.
3. Dummett, Michael. (1973) *Frege Philosophy of Language*. Gerald Duckworth & Company Limited, 43 Gloucester Crescent, London, UK.

4. Eklund, Matti. (2002) “Inconsistent Languages.” *Philosophy and Phenomenological Research*, Vol. 64, No. 2, pp. 251-275.
5. Eklund, Matti. (2005) “What Vagueness Consists In.” *Philosophical Studies*, Vol. 125, No. 1, pp. 27-60.
6. Eklund, Matti. (2007) “Meaning Constitutivity.” *Inquiry*, Vol. 50, No. 6, pp.559–574.
7. Field, Hartry. (2008). *Saving Truth from Paradox*. Oxford: Oxford University Press.
8. Field, Hartry. (1994) “Deflationist Views of Meaning and Content,” *Mind*, New Series, Vol. 103, No. 411.
9. Frege, Gottlob. (1892) “On Sense and Reference.” In Geach, P. and Black, M. (eds.) (1980) *Translations from the Philosophical Writings of Gottlob Frege*, 3rd edition. Oxford: Basil Blackwell.
10. Gaifman, Haim. (1992) “Pointers to Truth.” *The Journal of Philosophy*, Vol. 89, No. 5, pp.223-261.
11. Glanzberg, Michael. (2001) “The Liar in Context”, *Philosophical Studies*, Vol. 103, pp.217–251.
12. Glanzberg, Michael. (2004a) “A Contextual-Hierarchical Approach to Truth and the Liar Paradox.” *Journal of Philosophical Logic*, Vol. 33, No. 1, pp.27-88.
13. Glanzberg, Michael. (2004b) “Truth, Reflection, and Hierarchies.” *Synthese*, Vol. 142, pp.289-315.
14. Goldstein, Laurence. (2009) “A Consistent Way with Paradox.” *Philosophical Studies*, Vol. 144, pp.377–389.
15. Graff Fara, Delia. (2015) “Names Are Predicates.” *Philosophical Review*, Vol. 124, No. 1, pp.59-117.

16. Grim, Patrick. (1995) “Book Review: Universality and the Liar: An Essay on Truth and the Diagonal Argument.” *The Philosophical Review*, Vol. 104, No. 3, pp. 467-469.
17. Horwich, Paul. (1997) “The Composition of Meanings.” *Philosophical Review*, Vol. 106, pp.503-32.
18. Jeshion, Robin. (2015) “Referentialism and Predicativism about Proper Names.” *Erkenntnis* Special Volume on Proper Names, Dolf Rami (ed.), No. 80, pp.363-404.
19. Karttunen, Lauri. (1976) “Discourse Referents.” *Syntax and Semantics*, Vol. 7, Academic Press, Inc., New York, San Francisco, London.
20. Kearns, John T. “An Illocutionary Logical Explanation of the Liar Paradox.” *History and Philosophy of Logic*, Vol. 28, No. 1, pp.31-66.
21. Kearns, Stephen, and Magidor, Ofra. (2012) “Semantic Sovereignty.” *Philosophy and Phenomenological Research*. Vol. 85(2), pp.322-350.
22. Kim, Jaegwon. (1987) “‘Strong’ and ‘Global’ Supervenience Revisited.” Reprinted in Kim, Jaegwon (1993) *Supervenience and Mind: Selected Philosophical Essays*, Cambridge: Cambridge University Press, pp.79–91.
23. Kirkham, Richard Ladd. (1992). *Theories of Truth: A Critical Introduction.* Cambridge, MA: MIT Press.
24. Kripke, Saul. (1975) “Outline of A Theory of Truth.” *The Journal of Philosophy*, Vol. 72, No. 19, Seventy-Second Annual Meeting American Philosophical Association, Eastern Division, pp.690-716.
25. Kripke, Saul. (1980) *Naming and Necessity*, Cambridge, MA: Harvard University Press.
26. Larson, Richard K. (1988) “Implicit Arguments in Situation Semantics.” *Linguistics and Philosophy*, Vol. 11, No. 2, pp. 169-201.

27. Ludlow, Peter. (1996) “The Adicity of ‘Believes’ and the Hidden Indexical Theory.” *Analysis*. Vol. 56, No. 2, pp.97-101.

28. Lewis, David. (1970) “How to Define Theoretical Terms.” *The Journal of Philosophy*, Vol. 67, No. 13, pp. 427-446.

29. Lewis, David. (1972) “Psychophysical and Theoretical Identifications,” *Australasian Journal of Philosophy*, Vol. 50, No.3, 249-258.

30. Lewis, David. (1994) “Reduction of Mind.” In Samuel Guttenplan (ed.), *Companion to the Philosophy of Mind*, Blackwell, pp.412-431.

31. Lewis, David. (1997) “Naming the Colours.” *Australasian Journal of Philosophy*, Vol. 75, No.3, pp.325-342.

32. Marcus, Ruth Barcan. (1947) “The Identity of Individuals in a Strict Functional Calculus of Second Order”, *Journal of Symbolic Logic*, Vol. 12, No.1, pp.12–15.

33. McGee, Vann. (1992) “Maximal Consistent Sets of Instances of Tarski’s Schema (T).” *Journal of Philosophical Logic*, Vol. 21, No. 3, pp. 235-241.

34. McLaughlin, Brian and Bennett, Karen, “Supervenience,” *The Stanford Encyclopedia of Philosophy* (Spring 2014 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/spr2014/entries/supervenience/>.

35. Melia, Joseph, and Saatsi, Juha. (2006) “Ramseyfication and Theoretical Content.” *The British Journal for the Philosophy of Science*, Vol. 57, No. 3, pp. 561-585.

36. Parsons, Charles. (1974) “The Liar Paradox”, *Journal of Philosophical Logic*, 3: 381–412.

37. Peacocke, Christopher. (1992) *A Study of Concepts*. Cambridge, Mass: MIT Press.

38. Prior, Arthur N., 1955. “Curry’s Paradox and 3-Valued Logic”, *Australasian Journal of Philosophy*, Vol. 33, pp.177-82.

39. Priest, Graham. (1987) *In Contradiction*. Dordrecht, The Netherlands: Martinus Nijhoff.

40. Priest, Graham. (2006) *Doubt Truth to be a Liar*. Oxford, U.K., Oxford University Press.

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

42. Quine, Willard van Orman. (1970) *Philosophy of Logic,* Englewood Cliffs: Prentice Hall.

43. Scharp, Kevin. (2013) *Replacing Truth*. Oxford, U.K., Oxford University Press.

44. Searle, John. (1958) “Proper Names,” *Mind* 67: 166-73.

45. Sennet, Adam, “Ambiguity”, *The Stanford Encyclopedia of Philosophy* (Spring 2016 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/spr2016/entries/ambiguity/>.

46. Sider, Theodore. (2010) *Logic for Philosophy*. Oxford, U.K., Oxford University Press.

47. Simmons, Keith. *Universality and the Liar*, Cambridge, U.K., Cambridge University Press, 1993.

48. Smith, Nicholas. “Semantic Regularity and the Liar Paradox.” *The Monist*, Vol. 89, No.1, (2006), pp.178-202.

49. Tarski, Alfred. (1935) “Der Wahrheitsbegriff in den formalisierten Sprachen,” *Studia Philosophica*, Vol.1, pp.261–405.

50. Tarski, Alfred. (1944) “The Semantic Conception of Truth: and the Foundations of Semantics.” *Philosophy and Phenomenological Research*, Vol. 4, No. 3, pp.341-376.

51. Wilson, W. Kent. (1990). “Some Reflections on the Prosentential Theory of Truth.” In J. M. Dunn & A. Gupta (eds.), *Truth or Consequences*, pp.19-32. Dordrecht: Kluwer Academic Publishers.

CHAPTER 2

THE ABERRATIONIST APPROACH TO THE LIAR PARADOX AND ITS KIN: REVENGE AND OTHER CHALLENGES

1. ***Aberrationist Approaches to the Liar Paradox***
2. ***Revenge Problems: An Introduction***
3. ***Aberrationism and Revenge***
4. ***Other Liar-like Paradoxes***
5. ***Articulating Aberrationism in Full Generality***
6. ***Concluding Remarks***

1. Aberrationist Approaches to the Liar Paradox

1.1. Aberrationism in General

To begin with, let us define *Liar sentences*. For any sentence x, let us say that x is a *Liar sentence* if x is the negation of a sentence whose grammatical subject is an expression that refers to x and whose grammatical predicate is an expression that refers to truth.[1] In Chapter 1, I focus on Liar sentences much like L* below:

(L*) L* is not true.

Consideration of L* leads quickly to contradictions. Suppose L* is true. Then it is not true, a contradiction. Thus L* is not true. But since that is precisely what L* says, it is then true, a contradiction.

I'll now take a moment to put this reasoning more formally. To that end, it is worth taking a quick detour to introduce some rules of inference that will be used in the formal version, as well as throughout the rest of the essay. These are the rules of inference below, which are intimately associated with the term 'true':

(T-intro) $\dfrac{\varphi}{\text{`}\varphi\text{' is true}}$

(T-elim) $\dfrac{\text{`}\varphi\text{' is true}}{\varphi}$

Suppressing issues of ambiguity and context-sensitivity, here the 21st letter of the Greek alphabet is a substitutional variable ranging over all declarative sentences of English. These rules "govern" the word 'true' in at least the sense that by and large we are disposed to follow them, and, when we are feeling helpful, to correct uses of 'true' that deviate from them. (T-intro) and

[1] Throughout, I will use 'refers' so as to include not only reference by singular terms but also the reference-like relation in which predicates stand to properties, if this is something other than reference.

(T-elim) will figure prominently in what follows, so it is worth taking a moment to make them familiar.

With (T-intro) and (T-elim) in hand, here is a formal version of the reasoning associated with L*. It is called the *Strong Liar* paradox:

1. L* = 'L* is not true'	(definition of 'L*')
2. | L* is true	(Assume for *reductio*)
3. | 'L* is not true' is true	(Substitution, (1), (2))
4. | L* is not true	((T-elim), (3))
5. | Contradiction	(Contradiction introduction, (2), (4))
6. L* is not true	(Negation Introduction, (1)-(4))
7. 'L* is not true' is true	((T-intro), (6))
8. L* is true	(Substitution, (1), (7))
9. Contradiction	(Contradiction introduction, (6), (8))

If the reasoning above is truth-preserving, then some contradictions are true. And as we classically reason, if some contradictions are true then all other sentences are true. Yet somehow, not absolutely every sentence is true. So, something in the reasoning which led to this conclusion has to be wrong. The task of diagnosing the paradox as it arises for natural languages is therefore not one of concocting a new, paradox-free language, but rather one of giving the best explanation of how natural languages manage already to be coherent. In what follows, I will defend what I take to be the best explanation: a view I will call *aberrationism*.

To work as an explanation, aberrationism must at the very least entail that the paradoxical reasoning fails to go through. Let us now see how that is achieved. Notice that if L* fails to say of itself that it is not true, but at the same time the sentence written on lines (2) and (8) succeeds in saying of L* that it is true, then steps (2) and (4) do not contradict, and neither do steps (6) and (8). Aberrationists develop this claim by arguing that the *occurrence* of the word 'true' in L* fails to refer to truth. In brief, the view is that when a word like 'true' or 'satisfies' occurs in a Liar-like sentence, the occurrence of that word in that sentence fails to refer to truth. Therefore,

these sentences fail to say what they appear to say, and so they generate no contradictions. In a similar way, it helps to note, no contradictions arise when I utter ‘I am hungry’ and you utter ‘I am not hungry’; in different mouths, the sentences fail to express propositions that are genuinely contradictory. The distinctive claim of aberrationism is that this sort of thing can happen on a one-off basis in sentences that would otherwise be paradoxical, without any indexicality as we had with the word ‘I’. Again, aberrationists defend this claim as the best explanation of how natural languages manage to be coherent in the face of the Liar paradox.

Since the above diagnosis invokes occurrences, I should say a few words about these. It is easy to confuse occurrences with tokens, but the two are importantly distinct. Tokens, but not occurrences, are concrete. With respect to abstractness, occurrences are more similar to words and sentences than to their tokens. For instance, if I write the sentence ‘An occurrence of an expression is a lovely thing’ on a blackboard, I have created two tokens of the word ‘an’. But even before these tokens existed there were two occurrences of ‘an’ in that sentence. Once I erase the board, the tokens go out of (present) existence, but the occurrences do not.

For getting a grip on occurrences, it also helps to have a way of representing them formally. It will be harmless in what follows to think of an occurrence of an expression as an ordered pair of the expression and a *component context*. One can think of an occurrence’s component context as the result of using a hole-punch to remove the occurrence from the sentence, and replacing it with a blank (or, as I’ll write, ‘_’). Thus the occurrence of ‘true’ in L* can be represented as the ordered pair <‘true’, ‘L* is not _’>.

Before I can articulate aberrationism precisely, I need to introduce a few more concepts. For now, let us say that an expression *e* is an *alethic expression* if

(Base Clause)	*e* is inter-substitutable with an expression that is governed by (T-intro) and (T-elim),[2]
or	
(Inductive Clause)	the result of replacing each non-primitive component expression in *e* with its definiens contains an alethic expression.

Now notice that the same alethic expression can occur more than once in a Liar-like sentence, with only some of these occurrences being responsible for the paradox to which the sentence gives rise. Let's refer to these problem-causing occurrences as *key occurrences*. Key occurrences are hard to define rigorously, but it is very easy to get the main idea. Consider (A_w) below:

(A_w) A_w is not true, and 'War causes suffering' is true.

Clearly, only the first occurrence of 'true' plays an essential role in generating the paradox associated with A_w. So only that first occurrence counts as a key occurrence.

Given these definitions, *aberrationism* consists of the following two claims:

(Aberrations)	For any Liar-like sentence, the key occurrences of its alethic expressions differ in reference from the expressions of which they are occurrences.[3]
(Determined)	When an occurrence of an expression in a Liar-like sentence differs in reference from the expression of which it is an occurrence, what (if anything) the sentence says[4] is determined by what the occurrence refers to, rather than by what the expression *simpliciter* refers to.

[2] In Section 4.6, Grelling's Paradox will give us a reason to expand this definition. But until then, this simpler version will suffice.

[3] A similar view would simply say that the reference of the occurrences in question differs from that of the other occurrences of the expression, insisting that an expression *simpliciter* can refer to a thing *x* only in the sense that most of its occurrences refer to *x*. While this view is consistent with the spirit of aberrationism, it is worth noting that the view is controversial and aberrationism is not committed to it. See Chapter 1 Section 11.2 for further discussion.

[4] In this essay, I will use 'says' in such a way that in any world *w* in which redness is the most interesting property, the sentence 'Roses have the most interesting property' counts, in *w*, as saying that roses are red. This contrasts with a different use of 'says', on which this sentence counts only as saying that roses have the most interesting property, whatever that property may be.

Before moving on, a few clarificatory remarks about this definition. I use the term 'aberrations' because when an occurrence of an expression in a sentence fails to co-refer with that expression, it makes sense to say that the expression is there undergoing an aberration in its referential behavior. These aberrations are "one-off", because they only arise in cases in which there would otherwise be paradox. The idea for this overall approach was loosely inspired by (Smith 2006), who I take to be implicitly committed to (Aberrations) and (Determined).[5] However, explicit formulation and endorsement of these claims is due to me.

It is worth taking a moment to understand the role of (Determined) in the aberrationist picture. For all (Aberrations) says, what a Liar sentence says could be determined solely by the reference of its expressions *simpliciter*, not by the occurrences of these expressions in that sentence. In that case, even if the key occurrences of its alethic expressions underwent referential aberrations, that would not change what the sentence said; the sentence would give rise to contradictions regardless of these aberrations. In that case, the sentence would still give rise to contradictions, and positing aberrations would be of no use as an explanation of how ordinary speakers manage to avoid incoherence. For aberrationism to work as an explanation, then, we need the aberrations to affect what the sentence says. (Determined) guarantees this.

As another matter, the reader may have noticed that the definition of aberrationism uses the term 'Liar-like sentence'. This raises the difficult question of exactly which sentences count as Liar-like. In Section 5, I'll take up this question in detail. For now, it suffices to leave this matter at an informal level. Since I will arrive at a precise definition of 'Liar-like' by reflecting on the wide variety of sentences to which I want it to apply, it will be best to look first at these sentences and the paradoxes to which they give rise.

[5] See Chapter 1, Section 3.

One last clarificatory remark. Aberrationism does not claim that alethic expressions are sensitive to their contexts of use, or that they contain hidden indexical elements. Those varieties of referential shiftiness are neither one-off nor aberrations. For one thing, on those views, an occurrence of an alethic expression in a non-Liar-like sentence will undergo referential shifts when the sentence is uttered in the same context as a Liar sentence. Nothing like this happens on the aberrationist picture; it is only Liar-like sentences that witness aberrations. As a related point, indexical and context-sensitive expressions undergo shifts in reference in a wide variety of different situations, not just in Liar-like sentences. For instance, on some views,[6] the reference of these terms shifts in response to the number of nested semantic attributions being made in the context of utterance. By contrast, according to aberrationism, it is a mistake to try to explain what happens in paradoxical situations in terms of what happens anywhere else, or, relatedly, to claim that it happens anywhere else. Precisely one of the lessons of Liar-like paradoxes is that something genuinely distinctive and unique is going on in such cases.

My main goal in this essay is to show that aberrationism is invulnerable to revenge problems, and that it can be fruitfully applied to a wide variety of Liar-like paradoxes. I will also argue throughout that *moderate* versions of aberrationism are superior to other, *radical* versions, associated with (Smith 2006). I will use my own preferred version of moderate aberrationism to illustrate this point. With those goals in mind, I'll use the rest of this section to introduce the moderate vs. radical distinction and sketch my own preferred moderate view. Then, in Section 2, I will introduce the issue of revenge.

[6] See (Burge 1979).

1.2. Moderate vs. Radical One-Off Aberrations Approaches

If aberrationism is correct, then an important lesson of the Liar paradox is that our sentences cannot always say what we pre-theoretically take them to say. As (Smith 2006) does and as I do at length in Chapter 1, one argues for this by showing that the consequences of allowing a Liar sentence such as L* to say of itself that it is not true are worse than the consequences of denying that it does this. Still, necessary though it may be, this denial is painful and unintuitive. L* really does appear to say of itself that it is not true.

According to Smith, we simply need to bite the bullet here; there is little we can do to mollify the discomfort. He claims that

> Sometimes our words do not mean what we want them to mean: not due to hidden complexities of our semantic mechanisms, operating behind the scenes to produce unforeseen results—i.e., not for some principled, bottom-up reasons; but because our words *cannot* mean what we want them to mean, and so our semantic mechanisms simply break or malfunction, and some of our words get assigned meanings more or less randomly (p.195).

In the vicinity of these remarks are two views, each of which consists of the rejection of one of the theses below:

(Semantic Supervenience)	For every semantic property, there are some non-semantic properties on which it supervenes.[7]
(Semantic Regularity)	There are *perfectly* reliable, principled relationships between our behavior, mental states and physical environment on the one hand, and what we mean by our utterances on the other hand (Smith 2006 p.188, emphasis mine).

Let us call any version of aberrationism that rejects either Semantic Supervenience or Semantic Regularity a *radical one-off aberrations approach*, henceforth *radical aberrationism*. I argue in Chapter 1 that because they cast into doubt the robustness of the relation between semantic

[7] For a definition of supervenience, see (Bennett and McLaughlin 2014).

phenomena and non-semantic phenomena, radical approaches are not in a good position to explain our very vivid sense of what Liar-like sentences seem to say by appealing to facts about language use, or to identify what these sentences in fact say by looking at facts about language use. (Here and henceforth, when I speak of how words are *used*, I will include such phenomena as our linguistic behavior involving these words, the mental states we use them to express, and how our uses of them are related to our physical environment.)

By contrast, let us call any version of aberrationism that accepts both Semantic Supervenience and Semantic Regularity a *moderate approach*, henceforth *moderate aberrationism*. This is the approach that I defend here and in Chapter 1. In particular, in Chapter 1 I argue that the appearance that a sentence like L* says what it appears to say is in general very nearly correct, because such sentences come very close to saying what they appear to say. That is because, on my view, when an occurrence of a word cannot co-refer with that word, it need not follow that what the occurrence refers to has nothing to do with how we use the word. More specifically, even when an occurrence of 'true' cannot, on pain of contradiction, refer to truth, what it refers to is still influenced by our use of 'true' in accordance with (T-intro) and (T-elim).

In the rest of this section, I will describe what I find to be a plausible account of the nature of this influence, and show how this account can allow that L* comes close to saying what it appears to say. Readers should keep in mind that the purpose of this account is merely to serve as an illustration of the advantages that moderate approaches can in principle have over their radical cousins. Accordingly, while I will strive to present the account in sufficient detail to make it plausible, I will not provide its every element with a sustained defense. The same goes for Chapter 1, although the reader can find more details there.

1.2.1. An Example Moderate View

1.2.1.1. The General Idea

In Chapter 1, I describe a relation of *respect* that a property can bear to an inference rule. (See below for more on this relation.) Following a suggestion from (Lewis 1970), and noting the intimate relationship between the rules (T-intro) and (T-elim) and the word 'true', I claim that the word 'true' refers to that property, if there is one, which comes closest to respecting these rules. That thing, I argue there, is truth. Then I argue that something similar holds for occurrences of 'true' that cannot, on pain of contradiction, refer to truth. These occurrences refer to that thing, if there is one, which comes next-closest (after truth itself[8]) to respecting (T-intro) and (T-elim). If there are multiple such things, the occurrence is indeterminate in reference as between these. In Chapter 1, I suggest that this is precisely the situation. From (Scharp 2013) we learn of two quite truth-like properties, *ascending truth* and *descending truth*, which, I argue there, are tied for second place (after truth itself) when it comes to coming close to respecting (T-intro) and (T-elim). The reference of the occurrence of 'true' in L* is thus indeterminate as between these properties. L* is then indeterminate in content; that is, it is indeterminate whether L* says of itself that it is not ascending true or says of itself that it is not descending true. Because L* is indeterminate in content, it is also indeterminate in truth value.[9] Moreover, similar remarks hold for all other Liar-like sentences.

The version of moderate aberrationism just described is moderate in two important respects. Firstly, it allows us to retain the plausible thesis that the reference of any (English) occurrence of 'true' is significantly influenced by how we use the word 'true'. Secondly, it does

[8] See Chapter 1 Section 3.3.3.2 for an argument that truth respects these rules.
[9] See Chapter 1 Section 3.3.3.3 for further explanation of how the one variety of indeterminacy leads to the other in this case.

some justice to our strong pre-theoretical impression that L* says of itself that it is not true. As just noted above, L* is indeterminate as between saying of itself that it is not ascending true and saying of itself that it is not descending true. Because ascending truth and descending are quite similar to truth, L* thereby comes quite close to saying what it appears to say; and similarly for other Liar-like sentences. Moderate approaches in general have the potential to share these two virtues, depending on what exactly they say about the relationship between semantic and non-semantic phenomena.

1.2.1.2. Getting a Concrete Sense of Ascending Truth and Descending Truth

Since I will be using the moderate view just described as an illustrative example throughout, I will now give the reader a more concrete sense of the properties that it invokes, *ascending truth* and *descending truth*. In later sections, some of the discussion will turn on claims about ascending truth and descending truth that I justify in what follows. Readers who want to skip the details in the rest of this section can simply take those claims on faith.

In the rest of this section, the essential idea to have in the background is that because (T-intro) and (T-elim) play a highly significant role in governing our uses of 'true', they also play a highly significant role in determining its reference, and that of its occurrences. In particular, when an occurrence of 'true' fails to refer to truth, if it succeeds in referring at all then it refers to something that, like truth, has an intimate relationship with these rules of inference. Ascending truth and descending truth, I claim, have such a relationship: ascending truth *respects* (T-intro) and descending truth *respects* (T-elim).

By 'respects' I mean the following. A property P *respects* (T-intro) if for any declarative, context-insensitive sentence S, if the occurrence of 'true' in ⌜'S' is true⌝ were to refer to P then

the inference from S to ⌜'S' is true⌝ would be truth-preserving. (That is, if the premise were true then the conclusion would be true.) Likewise, a property P *respects* (T-elim) if for any declarative, context-insensitive sentence S, if the occurrence of 'true' in ⌜'S' is true⌝ were to refer to P then the inference from ⌜'S' is true⌝ to S would be truth-preserving. The point about ascending truth and descending truth is that ascending truth respects (T-intro) but not (T-elim) (although it comes close), and descending truth respects (T-elim) but not (T-intro) (although it comes close). Thus, each property respects one rule and comes close to respecting the other.

We have reason to think that ascending truth and descending truth respect (T-intro) and (T-elim), respectively, because of the way these properties are introduced. (Scharp 2013) presents two predicates, 'A(x)' and 'D(x)', which are defined by a list of 20 axiom schemata, including two that are conditionalized versions of (T-intro) and (T-elim):[10]

(A1) If S then A('S')

(D1) If D('S') then S

Here I use the 19th letter of the uppercase Latin alphabet substitutionally, to range over declarative, context-insensitive sentences of English. Scharp calls the set of all instances of his axiom-schemata *ADT*—"the theory of ascending and descending truth." Ascending truth and descending truth, then, are introduced as being those properties, if there are any, to which the predicates 'A(x)' and 'D(x)' respectively refer, when added to English and defined by our adopting the axioms in ADT.

When understood in this way, we have indeed some reason to think that the predicates 'A(x)' and 'D(x)' refer. That is because Scharp proves the consistency of ADT relative to the axioms of set theory. He does this in the standard way, by constructing a set-theoretic model, M_2,

[10] See Scharp's p.154 for a full list.

of ADT.[11] For similar reasons we are justified in taking ascending truth and descending truth to respect (T-intro) and (T-elim). That is, the referents-in-M_2 of 'A(x)' and 'D(x)' respect these rules, and that gives us inductive evidence that when these predicates are added to English and defined by ADT, their referents respect (T-intro) and (T-elim), respectively. So, if we take Scharp's model-construction seriously, we can take it to be a defining feature of ascending and descending truth that they respect (T-intro) and (T-elim), respectively.

A different defining feature of ascending and descending truth is that neither property respects the rule that the other respects. To begin with, no instance of either (A1#) or (D1#) below is in ADT:

(A1#) If A('S') then S

(D1#) If S then D('S')

Furthermore, it is not just that no instances of these schemata show up in the theory. It turns out that there are many instances of (A1#) and (D1#) which come out false in every model of ADT. Indeed, Scharp claims (p.186) that this is what happens for all and only pre-theoretically Liar-ish sentences (though he does not provide a proof).

One pair of sentences is particularly illuminating. Consider the sentences below, together with their associated instances of (A1#) and (D1#):

(α) α is not ascending true
(δ) δ is not descending true

$(A1\#)_{\alpha}$ If A('α is not ascending true') then α is not ascending true

[11] See his Section 6.6 (p.157-169) and an appendix (p.178) to Chapter 6. Assuming that set theory is consistent, in general a proof that an expression defined by a list of axioms has an extension in a set theoretic model serves as inductive evidence that if the expression were added to English, defined by those axioms, it would refer. In particular, then, the fact that 'ascending true' and 'descending true' have extensions when interpreted in Scharp's model serves as inductive evidence that they too would refer, if added to English and stipulated to be governed by the axioms in ADT. Moreover, Scharp is careful not to impose any expressive limitations on the language of M_2 that might be relevant to its being extensible to a full natural language such as English. Indeed, he objects to other theorists' failure to take such precautions (see p.156). Therefore, the point is, Scharp's discussion gives us some reason to think that 'A(x)' and 'D(x)', considered as expressions of an expanded English, refer.

$(D1\#)_{\delta}$ If δ is not descending true then D('δ is not descending true')

Since they can be used to derive contradictions, $(A1\#)_{\alpha}$ and $(D1\#)_{\delta}$ are false in every model of ADT.[12] Thus we have reason to believe that the sentences $(A1\#)_{\alpha}$ and $(D1\#)_{\delta}$ are false when 'A' and 'D' are considered as predicates added to English.

This example enables us to make two generalizations about ascending and descending truth, which may help the reader get a further grip on these properties. Firstly, any sentence S which says of itself only that it is not ascending true is (contrary to what it says) ascending true, and therefore false. And similarly, any sentence S which says of itself only that it is not descending true is not descending true, and is thus true.[13]

We just saw some instances of (A1#) and (D1#) that come out false in every model of ADT, and thus are arguably false when 'A(x)' and 'D(x)' are interpreted as predicates added to English. On the other hand, there are some other instances of (A1#) and (D1#) which come out true in every model of ADT. A quick look at which sentences these are will further help to give a concrete sense of ascending and descending truth. Scharp proves that all instances of (A1#) and (D1#) that involve sentences which are "grounded" come out true. A sentence is *grounded in*

[12] Here is a derivation of a contradiction that uses $(A1\#)_{\alpha}$:

1.	α = 'α is not ascending true'	(definition of 'α')
2.	α is ascending true	(proven elsewhere by Scharp)
3.	'α is not ascending true' is ascending true	(proven elsewhere by Scharp)
4.	If 'α is not ascending true' is ascending true then α is not ascending true	$(A1\#)_{\alpha}$
5.	α is not ascending true	(*modus ponens*, (3), (4))
6.	Contradiction	(Contradiction Introduction, (2), (5))

[13] Here is the reasoning that justifies these generalizations. Since $(A1\#)_{\alpha}$ is false, it has a true antecedent and a false conclusion. So 'A('α is not ascending true')' is true and 'α is not ascending true' is false. From both these claims we can conclude that α is ascending true. Since this contradicts what α says, that makes α false. Similarly for $(D1\#)_{\delta}$. Since $(D1\#)_{\delta}$ is false, it has a true antecedent and a false conclusion. So, 'δ is not descending true' is true and 'D('δ is not descending true')' is false. From both these claims we can conclude that δ is not descending true. Since this confirms what δ says, that makes δ true. Now, presumably (A1#) and (D1#) fail not only for α and δ, but for all sentences that deny of themselves that they are ascending true or that they are descending true, respectively. At least, Scharp is certainly committed to this claim, since he holds that (A1#) and (D1#) fail for all pre-theoretically Liar-ish sentences. And it is surely sufficient for a sentence's being pre-theoretically Liar-ish that it says of itself that it is not ascending (descending) true. Such sentences mimic Liar-sentences in an obvious way.

Scharp's sense if "its ascending truth value and descending truth value are completely determined by the ascending truth values and descending truth values of sentences that have no occurrences of 'ascending true' or 'descending true'" (p.170). So, although (A1#) and (D1#) are not generally true, they hold for any sentences that are grounded in Scharp's sense.

More generally, every sentence which is *safe* comes out true in every model of ADT. A sentence is *safe* if it is either descending true or not ascending true (see his p.186). Every sentence that is grounded is safe, but some ungrounded sentences are also safe. These include the sentence 'Every sentence is either ascending true or not ascending true' and 'No sentence is both descending true and not descending true' (see p.170). These examples of grounded and safe sentences help to give at least a rough sense of which instances of (A1#) and (D1#) are true in every model of ADT, which in turn helps to give a sense of ascending and descending truth.

So much, then, for getting a sense of ascending and descending truth. Now for how they fit into the moderate aberrationist view that I favor. It is my view that not only do ascending and descending truth exist and come close (after truth itself[14]) to respecting (T-intro) and (T-elim), but moreover they come equally close, and nothing else comes closer. Thus, they are "tied for second place" when it comes to respect of (T-intro) and (T-elim). It is for that reason that when an occurrence of 'true' cannot refer to truth, it divides its reference between ascending and descending truth. However, I will not defend the "tied for second place" claim here.[15] For our current purposes, it suffices to note that if indeed ascending truth and descending truth are tied for second place, then by the Lewisian claims that I described earlier, the reference of the occurrence of 'true' in L* is indeterminate as between ascending truth and descending truth. Since these properties are quite similar to truth, that then gives a robust sense in which L* comes

[14] At the beginning of Section 1.2.1.1 I claimed that truth respects these rules.
[15] See Chapter 1 Section 3.3.3.1 for further discussion.

close to saying of itself that it is not true. This enables my preferred version of moderate aberrationism to go a significant way toward honoring our pre-theoretical impressions about what Liar-like sentences say.

Now that I have introduced moderate aberrationism and my own preferred version of it, I am in a position to discuss three questions:

A. Does aberrationism in general, or my own preferred version of it in particular, face a revenge problem?
B. Can moderate aberrationists invoke indeterminacy without falling victim to the classic revenge problems for approaches that invoke indeterminacy?
C. Even if aberrationism solves the Strong Liar paradox (which was used to introduce it), does it fall prey to any other Liar-like paradoxes?

In Sections 2, 3, and 4, I address these questions, defending both aberrationism in general and my own favored version of it. Then in Section 5 I explain how aberrationist approaches to the Strong Liar paradox can generalize to all other Liar-like paradoxes.

2. Revenge Problems: An Introduction

2.1. The Classic Liar and Strong Liar

Solutions to the Liar paradox tend, notoriously, to be vulnerable to revenge problems. A solution faces a *revenge problem* if the terms or concepts that the solution introduces can be used to state a palpably Liar-ish paradox, one that, by the lights of the solution being proposed, cannot be solved using those same terms or concepts. In this section, I will give several examples of approaches to the Liar paradox and revenge problems they face, to give the reader a sense of what revenge problems are. Let's start with the *Classic Liar Paradox*. There, consideration of the following sentence quickly leads to contradictions:

(L) L is false

Suppose L is true. Then L is false, a contradiction. So, then L must not be true. But then L is false. However, that is precisely what L says! So, L is true after all, contradicting what we have shown, namely, that L is false. Paradox!

A classic reaction to this situation is to hold that L is neither true nor false. L, one might say, has no truth value. However, a sentence that has no truth value is in particular not true; it lacks the truth value *truth*. And now that the notion of a sentence's not being true has been introduced, we can obtain a new version of the paradox, the *Strong Liar Paradox*, using the following sentence:[16]

(L*) L* is not true.

Suppose L* is true. Then L* is not true, a contradiction! Thus L* is not true. But that is precisely what L* says! So, L* is true after all, a contradiction! Thus, classic approaches to the Liar paradox which involve denying that Liar sentences such as L have any truth values face a revenge problem in connection with L*.

2.2. Contextualist Views

Views on which all Liar sentences contain some context-sensitive elements also tend to face revenge problems. For example, (Simmons 1993) takes the word 'true' to refer to different properties in different contexts of use. Thus, in no context of use does the word 'true' refer to any property that would render any sentences paradoxical relative to that context. However, this solution falls prey to a revenge problem. Once one has introduced contexts of use, one can construct sentences of the following sorts:

[16] There is also a different way to make trouble. L says of itself that it is false. Suppose L is neither true nor false. Then in particular it is not false. Then what L says is not the case. But that would make L false after all, contradicting the original assumption that L is neither true nor false.

(X) X does not fall under 'true' in any context of use

(X_1) X_1 has none of the properties over which 'true' ranges, across all different contexts of use

To deal with sentences like X and X_1, all contextualist views have to ban unrestricted quantification of some kind or other—either over contexts of use,[17] or over the properties available as referents for 'true' across all different contexts.[18] The point is, contextualists introduce contexts of use and various properties to which 'true' refers relative to different contexts of use; and both of these notions can be used to restate a new version of the paradox, one that cannot be solved without importing some further assumptions.

2.3. Dialetheism

Dialetheist approaches to the paradox also face revenge problems. A *dialetheist approach* is one that relies on the claim that some sentences, in particular Liar sentences, are both true and not true. Now, one might initially worry that dialetheist approaches are bound to be *trivial*. An approach to the Liar paradoxes is *trivial* if anyone who endorses the approach is thereby bound to accept that absolutely every declarative sentence is true, or to committed to assent to absolutely every declarative sentence. Triviality concerns about dialetheism stem from the fact that the classical law *ex falso quodlibet* allows one to derive any declarative sentence whatsoever from a *contradiction*—a sentence of the form 'P and not P'—and dialetheists accept some

[17] (Grim 1995), pp.468-469, shows that (Simmons 1993) has to ban unrestricted quantification over contexts of use.
[18] According to both (Burge 1979) and (Glanzberg 2004a), there are infinitely many truth-like properties, which are arranged in an infinite hierarchy and indexed according to their levels in the hierarchy. Burge and Glanzberg ban unrestricted quantification over indices. See (Glanzberg 2004b) p.289, and (Burge 1979) p.196 fn. 28 for further discussion, and see (Scharp 2013) pp.117-119 for related criticism of Glanzberg.

contradictions. However, dialetheists avoid triviality by rejecting *ex falso quodlibet*. They insist that contradictions, properly understood, do not entail absolutely everything.

As is typical with revenge problems, one of the notions that features centrally in the articulation of dialetheism can be used to construct a new, problematic sentence. Once the notion of a sentence's entailing absolutely everything (or, likewise, the notion of its not doing so) is introduced, one can introduce an expression, 'NOT', by simply stipulating that for any declarative sentence S, from S and ⌜NOT-S⌝ everything follows.[19] That is, from this pair one may infer any declarative sentence one likes.[20] With 'NOT' in hand, one can easily construct a sentence consideration of which leads to triviality:

(L_{NOT}) L_{NOT} is NOT true

To avoid triviality in the face of L_{NOT}, dialetheists need to insist that it is impossible to define 'NOT' or any other expression in the way just described.[21] Some philosophers[22] question how reasonable this insistence is, and I sympathize heartily with this skeptical reaction; still, we need not dig deeply into this complex matter here. The point for now is just that like many other responses to the Liar, the dialetheist's initial move introduces a concept that can, *prima facie*, be used to state a new, recalcitrant version of the paradox.

So far, we have seen some alternative approaches to the Liar paradox and the revenge problems that afflict them. One of my main goals in this essay is to show that aberrationism, and in particular moderate aberrationism, is invulnerable to such problems.

[19] Here I set aside considerations of context-sensitivity.
[20] Indeed, this is how logicians, except for devotees of dialetheism, have modeled our everyday use of the word 'not'. *Ex falso quodlibet* can be seen as a way of articulating the idea that one who accepts a contradiction might as well accept anything else, which in turn is a way of cashing out the near-universal idea that accepting contradictions is bad.
[21] Many prominent dialetheists do in fact deny this. See, for example, (Priest 1990), pp.204-209.
[22] (Scharp 2013) Section 4.3.2 (p.104).

3. Aberrationism and Revenge

3.1. A General Lesson

Before I get going discussing the revenge problems that might arise for aberrationism, I want to articulate an important general lesson concerning these problems. Nearly every approach to the Liar paradox introduces some special property P, and then appeals to P in order to (putatively) solve the paradox. Revenge problems arise when one defines a predicate that refers to P, and then uses that predicate to construct a new, problematic sentence. Most approaches are then stuck between two unsatisfying lines of response. The first kind of response is to deny that there is a predicate that refers to P; or, more strictly speaking, that refers to whatever P-ish property one appeals to in constructing the new problematic sentence.[23] But for two reasons, that line of response is a dead end. Firstly, theorists who want to appeal to a property P need a way to attribute P to things; invariably, they will end up needing a predicate that, by their own lights, refers to P. But then, if in fact that predicate cannot refer to P, then this calls into question the intelligibility of their own attributions of P. Secondly, philosophers who want to ban predicates that refer to P have to explain why we cannot simply introduce one. (The reader may anticipate a similar objection to aberrationism, which purports to introduce a predicate all of whose occurrences refer to truth. I address this objection in Section 3.3.)

A different kind of response, then, is to distinguish between an *object language*—the language one is describing, for which the paradox arises—and the *meta-language*—the language that one is using to describe the object language. Then one insists that one is solving the paradox only for the object language, doing so by introducing a new expression into (only) the meta-

[23] For example, recall from Section 2 that contextualists introduce the notion of falling under 'true' relative to a context, but then have to deny that one can speak of a sentence's failing to fall under 'true' relative to any context.

language.[24] But such approaches face a dilemma. If it is possible to apply the new expression to sentences in the metalanguage, then revenge problems arise. If the expression can be applied only to sentences in the object language, then the approach in question is incapable of providing a fully general solution to the Liar paradox. It can at best solve the versions of the paradox that arise for languages other than the one in which the solution itself is expressed. Of course, the solution's advocate may insist that she has still provided a solution recipe: the same sort of thing can be obtained for the current metalanguage as for the object language, by ascending to a meta-meta-language, and so on all the way up. But this response nonetheless makes a significant concession: that there is no single language in which we can say everything we want to say. To this, actual natural languages such as English strongly appear to be counterexamples. They seem—somehow!—to allow for just the kind of expressiveness that the thinker in question claims is impossible. It is worth developing approaches to the Liar paradox that do not force us to relinquish this idea.

Aberrationism, we will soon see, does better on all the matters just described. For one thing, it can allow that for any property, we can have a predicate that refers to that property. One avoids revenge by denying that this predicate can in turn be used to construct a new, problematic sentence. Importantly, that is not because reference to the property is altogether impossible. Rather, it is only the key occurrences of the expression in question in the offending sentence fail to refer to the property; by contrast, all the occurrences that are used in stating aberrationism succeed in referring to the property. Secondly, because unproblematic occurrences of the predicate succeed in referring to the property, reference to the property does not require a retreat to a meta-language. And since it does not require a meta-language, aberrationism can apply to

[24] See (Kripke 1975) for this kind of move.

the very language(s) in which it is stated. In short: for the price of a minor constraint on which sentences can say which things, we can have a single language that allows us (one way or another) to say everything we want to say.[25] Freedom from revenge-sentences and from the need to retreat to a meta-language are major advantages of aberrationism.

As we have seen, a revenge problem occurs when the notions that one introduces to solve one form of the Liar paradox can be used to formulate a new version of the paradox, one that cannot be solved by appealing to those same notions. The potentially problematic notions common to all versions of aberrationism are that of a component context, reference by an occurrence, and a Liar-like sentence. To these notions, my own preferred view (along with various other versions of aberrationism) adds to these the notions of indeterminacy and of ascending truth and descending truth. In the rest of Section 3, then, I will discuss a variety of attempts to formulate a new Liar-like paradox using these notions—though I will postpone discussion of revenge problems involving the term 'Liar-like sentence' until I have explored some definitions of this term in Section 5. The basic strategy for avoiding revenge is the same in all the cases I will discuss here, and is precisely the strategy employed to solve the original, Strong Liar Paradox: posit a one-off aberration. The challenge will be to show that the same considerations that one marshals to support an aberrationist diagnosis of the Strong Liar can be marshalled in these cases as well.

[25] In particular, one might have worried that my use of the expression 'the property of being true' throughout the essay requires that all of its occurrences refer to truth, and that this can only happen if the expression belongs to a meta-language. However, this concern fundamentally misunderstands the nature of aberrationism. All aberrationists require is that all the occurrences of 'the property of being true' employed in the course of arguing for aberrationism refer to truth. Aberrationists can allow that various other occurrences fail in this regard. Indeed, they must allow this in order to apply aberrationism to sentences such as 'This sentence lacks the property of being true'; precisely what they want to say about that sentence is that the occurrence of 'the property of being true' in that sentence fails to refer to truth, and so the sentence fails to say of itself that it is not true. One can of course say of that sentence that it is not true; it is just that one needs a different sentence to do it. For instance, the following will do: 'The sentence 'This sentence lacks the property of being true' lacks the property of being true'. That sentence is perfectly in order, as it fails to say of itself that it is not true. Rather, it speaks to the status of a different sentence.

3.2. Occurrences and Component Contexts

Let us begin by trying to construct a revenge sentence using the notions of an occurrence and a component context. Just as some contextualist approaches cannot accommodate unrestricted quantification over contexts of utterance, one might worry that aberrationism cannot accommodate unrestricted quantification over component contexts. For contextualists, the sentence that brings out this problem is one that says of itself that it does not fall under the extension of 'true' in any context of utterance. So, one might hope to formulate an analogous sentence that instead quantifies over component contexts. Here is my best attempt to formulate such a sentence:

(X*) In no component context does the occurrence of 'true' in X* refer to anything which would make the resulting sentence true.

In parsing X*, keep in mind that what I am asserting to be relativized to component contexts (strictly speaking) is reference by expression types, not reference by occurrences. On my picture, occurrences refer *simpliciter*. So, strictly speaking, it makes no sense to speak of the reference of an occurrence <*e*, C> as being relative to a component context C*, even if C* = C. Rather, if <*e*, C> refers at all then it refers *simpliciter*, not relative to anything. Therefore, strictly speaking, (X*) makes no sense, since it treats reference by an occurrence (namely, the occurrence of 'true' in X*) as being relative to the component context.

Less strictly speaking, though, one might simply reinterpret talk of what an occurrence refers to relative to all component contexts as talk of what it refers to. That view would take X* to be synonymous with X** below:

(X**) The occurrence of ‘true’ in X** fails to refer to anything that would make X** true.

X** makes perfectly good sense. However, it is straightforward what aberrationism says about sentences like this: they fail to say what they appear to say, because their occurrences of ‘true’ fail to refer to truth. Thus X** fails to say anything from which it follows that X** is not true; so, there is no paradox here.

3.3. Stipulations

In Section 2.3, I mentioned that according to dialetheists, the English connective ‘not’ fails to obey the rule *ex falso quodlibet*. This claim enables dialetheists to insist that for some sentences S the conjunction ˹S and not-S˺ is true, without allowing that every declarative sentence whatsoever is true. As we saw there, the question then arises whether it is possible simply to add to English a connective, ‘NOT’, that does obey *ex falso quodlibet*. If this is possible, then it hardly matters whether the dialetheists’ claims about the original word ‘not’ are correct. For as we saw, in that case it is easy to construct a Liar-like paradox that dialetheism cannot solve, a version involving ‘NOT’ rather than ‘not’.

One might have similar concerns about aberrationism. Suppose I am right that the key occurrence(s) of the word ‘true’ in any Liar-like sentence fails to refer to truth, and analogously for other alethic expressions. Still, cannot one simply define a new word, ‘TRUE’, by explicitly stipulating that all of its occurrences are to refer to truth, and then define a new Liar sentence using this new word? E.g.,

(L_T) L_T is not TRUE

Indeed, if such a stipulation is possible, then that raises the question why our existing intentions and conventions for using 'true' do not already guarantee that all of its occurrences refer to truth. For surely, speakers of English do not set out to make the word 'true' sensitive to its component contexts; rather, we set out, if only implicitly, to have every occurrence of 'true' refer to truth. But if we can achieve this by stipulation, then it is hard to see why we didn't achieve it already with our original intentions and conventions for using the word.

In response, a first point I want to make about L_T is that however unappealing some might find this move, an available response is simply to bite the bullet and insist that no stipulation like the one that defines 'TRUE' could succeed. In fact, adopting this response will not perturb any of aberrationism's central claims. One of the main lessons that aberrationists want to draw from the Liar paradox is that users of a language do not have complete control over what their expressions refer to and what their sentences say. It is perfectly congenial to this general outlook to hold that explicit stipulations can no more guarantee such control than tacit intentions can. So, the challenge posed by L_T is not that no aberrationist diagnosis of L_T is available in the logical space, but rather that, in light of the stipulation that defines 'TRUE', such a diagnosis is harder to swallow here than in other cases.

But, having pointed this out, we can now observe that nearly every competing diagnosis faces a similar problem. The problem arises from the fact that all diagnoses of the Liar paradox as it arises for natural languages must navigate between two conflicting requirements. On the one hand, a firm, widely accepted desideratum is to salvage as much as possible of our pre-theoretical conception of these languages.[26] This includes, in particular, doing justice to the idea

[26] This desideratum is one of the things that distinguishes *diagnoses* from *solutions*. Because it involves providing an account of natural languages, a diagnosis must give due deference to what we already have reasons to believe about these languages. One can see this desideratum at work in the widespread recognition that (Tarski 1935)'s approach to the Liar paradox, which bans self-reference, was too "restrictive". As (Feferman 1984) puts the point,

that natural languages are *maximally expressive*—that anything which can be said at all can be said in a single natural language, perhaps supplemented with some further vocabulary, and that it can be said in whatever ways the compositional grammar and lexical semantics of the language would seem to allow. On the other hand, however, unless the paradox renders natural language incoherent, some of the expressions involved must not behave as they pre-theoretically appear to behave; avoiding a diagnosis of incoherence requires one to violate our pre-theoretical conception at least somewhere. In practice, nearly all violations that one might posit involve some expressive limitations, and thus conflict to some degree with the idea that natural languages are maximally expressive. (See below for views that violate our pre-theoretical conception in other ways.) For example, Tarski-style approaches ban self-reference[27]; contextualists (see Section 4), and, in effect, (Kripke 1975),[28] ban unrestricted quantification; and dialetheists insist that there is no way of using 'not' so that it expresses *Boolean negation*, the kind of negation that comports with the rule *ex falso quodlibet*.[29]

At this point, we can start to see the situation for such diagnoses (those that violate maximal expressiveness) resemble the situation we had with aberrationism and L_T. Against any proposed expressive restriction, a critic can always object by introducing an expression that is (purportedly) stipulated to violate that restriction. And in response, the advocate of the approach under consideration can always dig in her heels and simply insist that no such stipulation can succeed, that no natural language can, without further changes, accommodate an expression

"natural language abounds with directly or indirectly self-referential yet apparently harmless expressions—all of which are excluded from the Tarskian framework" (p.77). If the aim were simply to construct a new language in which to do most mathematics, with no attempt made to simulate the features of natural languages, then the charge of restrictiveness could hardly count as an objection to Tarski's approach.

[27] See (Tarski 1935) and (Tarski 1944). Tarski constructs a hierarchy of languages, each with its own truth-predicate which can only be applied to sentences at strictly lower levels in the hierarchy.

[28] Kripke defines a language that is constructed over a transfinite hierarchy of stages. On pain of contradiction, the language does not allow for universal quantification over these stages.

[29] See (Scharp 2013) p.104-106 for discussion of this point.

defined in such a way and remain coherent. Deciding which of these views makes for the best diagnosis, then, is a matter of swallowing the pill that is least bitter. That is, the stipulation to sacrifice is the one whose loss inflicts the least mutilation on our pre-theoretical beliefs about natural languages, and in particular the thesis of maximal expressiveness.

There is no space here for an exhaustive comparison of the many competing approaches,[30] but two important points in favor of aberrationism are worth noting. Firstly, because it posits aberrations only in Liar-like sentences, aberrationism leaves the rest of language untouched. Thus, aberrationists can allow that the non-paradoxical areas of language are just as linguistics and casual inspection say they are, including when it comes to expressiveness. By contrast, for example, to reject unrestricted quantification[31] is to deny that it can be achieved anywhere, not just that it can be achieved in Liar-like sentences. Similarly, dialetheists do not merely claim that occurrences of 'not' in Liar sentences fail to express Boolean negation; they deny that 'not', or any other expression, can ever express Boolean negation. This claim has strong consequences concerning the meanings of all non-paradoxical sentences that contain 'not'. Thus, the expressive limitations posited by contextualism and dialetheism go way beyond the sentences involved in the Liar paradox.

A second, closely related point is that unlike many competing diagnoses, aberrationism has no need to posit any properties to which it is impossible to refer, or any proposition that cannot be expressed.[32] As discussed in Section 3.1, such claims pose a *prima facie* challenge to the coherence of approaches that make them, since these views end up purporting to refer to the

[30] See Chapter 1 for a more comprehensive discussion.
[31] See (Burge 1979), (Simmons 1987), and (Glanzberg 2004b) for such views. (Kripke 1975) can also be understood in this way, as denying the possibility that one can, while speaking a given language, quantify unrestrictedly over the stages in the construction of that language.
[32] Compare (Tarski 1935), which holds that no predicate in a language can refer to the property of being a true sentence of that language. See also the charge of incompleteness against (Kripke 1975) in Section 5 of Chapter 1.

property, express the proposition. Aberrationism avoids this problem. For one thing, although occurrences of 'true' in Liar sentences cannot refer to truth, plenty of other occurrences of 'true' (in non-Liar sentences) can—so there is no need for aberrationists to claim that no expression can refer to truth. Similarly, what a Liar-like sentence cannot say about itself (namely, that it is not true,) can be said by another, non-Liar-like sentence. For example, the sentence

(L') 'L* is not true' is not true

succeeds in saying of L* that it is not true. To get the idea, it helps to note that L' is not a Liar sentence: L' attributes untruth not to itself but rather to the (distinct) sentence L*.[33] So, there is no need for aberrationists to claim that no sentence can say of L* that it is not true. Again, the overarching point here is that while most diagnoses posit some expressive limitations on natural languages, the limitations posited by aberrationism are remarkably unrestrictive. To speak metaphorically, no area of reality is in principle inaccessible to language; it is just that certain areas cannot be reached by certain routes.

So far, I have only discussed competing views that violate the thesis of maximal expressiveness. Still, one might wonder whether cases like L_T work to the advantage of views that do not posit any expressive limitations. The answer is that even if they do, it is far from obvious that this advantage is decisive, because views that posit no expressive restrictions are bound to violate other aspects of our pre-theoretical conception(s) of language. One could, for example, take the Liar paradox to show that natural languages, and indeed any languages with

[33] Still, one might reasonably be concerned about the fact that L' can be converted into a Liar sentence by substituting the name 'L*' for the co-referential quote name ''L* is not true''. Should not L' therefore count as Liar-like? See Section 5 below for a full discussion of this issue, including a definition of 'Liar-like sentence' which includes L* but excludes L'.

the features blamed for generating the paradox, are simply incoherent.[34,35] But this claim is implausible on its face. It is just obvious that natural languages manage to be coherent, even despite the paradox; for not absolutely every sentence is true. If that is right, then positing pretty much any expressive limitation will make for a better diagnosis than positing incoherence.

There are, however, less trivial diagnoses that refrain from positing expressive limitations. (Brandom 1994) defends a *prosententialist* view, on which the expression 'is true' is not a predicate whose job is to refer to a property, but rather functions as a *prosentence-forming-operator*: when 'It is true that' is applied to a sentence, the resulting sentence is a *prosentence*—a sentence which inherits its content from the original sentence, in much the way that a pronoun can inherit its content from an antecedent expression in the discourse.[36] (E.g., 'It is true that snow is white' has the same content as 'Snow is white'.) Prosententialism posits no expressive restrictions on English, insofar as it does not involve positing any property to which no predicate can refer, or banning any kind of unrestricted quantification, or positing any one-off aberrations, etc. However, (Wilson 1990) argues convincingly that views like Brandom's conflict with the linguistic data about English: in sentences in which 'it is true that' occurs, prosententialists take that expression, rather than 'is true', to be a grammatical constituent. As Wilson illustrates, this hypothesis has a number of consequences that contradict speakers' judgments of grammaticality.[37] Weighing prosententialism against aberrationism, then, is a matter of weighing

[34] (Tarski 1935) is most naturally read as taking self-reference to be incoherent, thus rendering incoherent any language that allows for it.

[35] One should be careful to distinguish this from the claim that natural languages do not genuinely possess the feature in question, and so are coherent after all. While this is not an accurate reading of Tarski, it is one way his ideas might be used.

[36] See Chapter 1 Section 8 for further elaboration of Brandom's view.

[37] For example, if prosententialism takes the syntax of sentences containing 'true' to reflect their logical form, then it entails that the sentence 'What is true is that Bleda is vicious' is ungrammatical and that 'What is true that is Bleda is vicious' is grammatical. But these claims are clearly false. By contrast, the hypothesis that 'is true' functions as a predicate entails the reverse, that 'What is true is that Bleda is vicious' is grammatical and that 'What is true that is Bleda is vicious' is ungrammatical. (These examples are due to (Wilson 1990) pp.23-24.) Prosententialists can

these violations against the aberrationist's denial that the stipulation which defines 'TRUE' could succeed. Without reconstructing that debate in full here, I will point out that whereas the theory of the relation of grammatical constituency is on comparatively solid footing, reference remains a topic of active research and competing theories. Accordingly, then, it is preferable to take the Liar paradox to teach us something new and surprising about reference than to take it to refute linguists' canonical conception of grammatical constituency.

Before moving on, let me make one final, important point about stipulations like the one that (purportedly) defined 'TRUE' above. When presented with L_T, aberrationists should concede straightaway that it is possible to define an expression, 'TRUE', in a way that guarantees that all of its occurrences refer to truth. It is just that by aberrationist lights, one cannot—as was alleged in the original example—do this while keeping fixed everything else about one's language. On the contrary; in the face of such a stipulation, the language would have to compensate in some other way, forestalling the construction of a sentence which says of itself that it is not true. For instance, the language could be such that no occurrence of 'not' in any Liar sentence expresses negation. *Qua* sentence of such a language, L_T is quite unproblematic, since *qua* sentence of that language, it straightforwardly fails to say of itself that it is not true. The distinctive claim of aberrationism, then, is not that no language whatsoever can contain an expression like 'TRUE', but rather that natural languages cannot do so without undergoing various deformations that would result in a different language. Absent such deformations, in a natural language what breaks down in the face of the paradox is the reference of the key

respond by pointing out that syntax and logical form can come apart. E.g., in the sentence 'I saw a woman', the phrase 'a woman' serves as a noun phrase, even though at the level of logical form it is a quantifier phrase. Still, while there are known instances of mismatch between surface syntax and logical form, there is a general presumption against positing such mismatches.

occurrences of ‘true’ in Liar-like sentences. As I explain at greater length elsewhere,[38] what justifies this strategy of targeting ‘true’ rather than, for example, ‘not’, for unusual behavior is that this strategy affords us the most uniform diagnosis of the many different Liar-like paradoxes as they arise for natural languages. Every such paradox involves ‘true’ or some similar expression,[39] whereas only some involve the connective ‘not’, and similarly for many other traditionally targeted expressions such as those that make for self-reference.

While there is of course more to discuss concerning revenge, the foregoing should be enough to give aberrationists reason for optimism. Accordingly, I will press on and consider threats of revenge that might come from aberrationist appeals to indeterminacy.

3.4. Indeterminacy, and an Introduction to Parity Problems

Let us now try to obtain a revenge sentence using the notion of indeterminacy. A first thing to note here is that my own preferred version of moderate aberrationism in fact invokes two kinds of indeterminacy: indeterminacy as to the content of a sentence (what it says), and indeterminacy as to the sentence’s truth value. On my view, Liar-like sentences have both: they are indeterminate in truth value because they are indeterminate in content. This feature of the view will figure importantly throughout.

We have reasons quite independent of the Liar paradox and its kin to hold that some sentences are indeterminate in content.[40] Anyone who grants this will have to embrace some

[38] See Chapter 1, Section 10.
[39] Grelling’s paradox involves the relation symbol ‘satisfies’.
[40] See (Heck 2003) and (McGee and McLaughlin 2004) for exposition of some views that take the Sorites paradox to involve indeterminacy in content. (Quine 1960) and (Kripke 1982) argue that such indeterminacy is in fact quite pervasive; they hold that no sub-sentential expression in any language is determinate in reference, and that the reference of each putatively-referring expression is indeterminate as between many radically different entities. Similarly, then, all declarative sentences are radically indeterminate as to their contents. Whether or not one thinks the situation is as bad as this, the considerations Quine and Kripke raise make it plausible that at least some small amount of indeterminacy in content holds for at least some expressions.

semantics of indeterminacy—that is, some account of how the indeterminacy of a word can influence the contents and truth values of sentences that contain it, and how indeterminacy as to a sentence's truth value can influence the truth values of logically more complex sentences of which it is a constituent. So, the challenge of constructing a workable semantics of indeterminacy is one faced by many different theorists, not just those who incorporate indeterminacy into their approaches to the Liar paradox.

That said, the question remains whether the appeal to indeterminacy in truth value in connection with the Liar paradox gives rise to a revenge problem. And many classic approaches that invoke such indeterminacy are indeed vulnerable to such problems. The standard way to make trouble is with a sentence like L**:

(L**) L** is either indeterminate or false

If L** is indeterminate, then it is either indeterminate or false. But that is precisely what L** says, making it true. However, a sentence cannot be both indeterminate and true,[41] so we have a contradiction. Thus L** is not indeterminate. So, it is either true or false. If it is true, then, given what it says, it is either indeterminate or false. We showed already that L** is not indeterminate, so it must be false. But then it is not true, contradicting our assumption that it is true! We got this contradiction by assuming that L** is true. Since we showed already that L** must be either true or false, we now know that L** must be false. But then it is either indeterminate or false. However, that is precisely what L** says; so, L** is true after all! This contradicts what we just

[41] (Barnes and Williams 2011) shows that this does not hold on all conceptions of indeterminacy. However, there is no point to invoking indeterminacy as an approach to the Liar paradox unless one can claim that some indeterminate sentences are neither true nor false. The appeal to indeterminacy as an approach to the Liar is motivated by sentences like L from above ('L is false'). As we saw, we have trouble if L is true and trouble if L is false. The point of claiming that L is indeterminate is to escape the choice between these two scenarios. Positing indeterminacy is therefore of no help unless it can allow for such escape.

showed (namely, that L** is false). And now we have derived a contradiction with no undischarged assumptions.

The appeal to indeterminacy in truth value, as we just saw, leads to trouble. That is because once we have a word 'indeterminate' that refers to the property *being indeterminate in truth value*, we can, it seems, use that word to formulate a revenge sentence.[42] However, aberrationism can avoid this problem. L** generates contradictions only if it succeeds in saying of itself that it is either indeterminate or false. But that is precisely the sort of thing that aberrationists would deny. Such a denial can be secured by positing a one-off aberration in the occurrence of either 'indeterminate' or 'false' in L**. I will now explore some proposals to that effect. (Here again, I refer readers to Section 3.3 for discussion of the concern that one can simply introduce a predicate in a way that guarantees that its occurrences refer as desired.)

There are several constraints on a consistent one-off aberrations diagnosis. For one thing, if the occurrence of 'indeterminate' in L** refers to the property *being indeterminate in truth value*, then on pain of inconsistency it can't be that L** is indeterminate in truth value.[43] In that case, the first disjunct of L** would say of L** that it is indeterminate, and because L** is indeterminate that disjunct would be true. But then that would make L** true, and so not indeterminate.[44] Likewise, if the occurrence of 'false' in L** refers to falsehood, then it can't be that L** is false. If the occurrence of 'false' refers to falsehood, then L**'s second disjunct says of L** that it is false. If L** is indeed false, then that disjunct is true, making L** true, and so not false, a contradiction. One can also check that on pain of contradiction, the occurrence of

[42] Of course, theorists who favor the appeal to indeterminacy as a solution to the Liar paradox have attempted to resist the revenge problem I have just described. See (Soames 1999).
[43] Again, here I assume that indeterminate sentence cannot also be true. Recall footnote 25.
[44] For similar reasons, it can't be that the occurrence of 'false' refers to *being indeterminate* and L** is indeterminate. But it is anyway difficult to see how the occurrence of 'false' could come to refer to *being indeterminate*. At least, it is hard to identify anything about our use of 'false' that could lead to this.

'false' in L** cannot refer to *not being ascending true* or to *not being descending true*—no matter whether L** is true, false, or indeterminate.

Another constraint is the following. Suppose that the occurrence of 'false' refers, not to falsehood, but to the property *not being true*. And suppose that L** is indeterminate. By the first supposition, the second disjunct of L** says of L** that it is not true. Since L** is indeterminate (and so not true), that second disjunct is true, making L** true after all, a contradiction.

Even with these constraints, however, there are a number of possibilities left open to aberrationists. Here are several:

a) the occurrence of 'indeterminate' fails to refer to indeterminacy, the occurrence of 'false' refers to falsehood, and L** is indeterminate.

b) the occurrence of 'false' fails to refer to any of the following properties: falsehood, *not being true*, *not being ascending true*, *not being descending true*. It refers to some other property that L** lacks. The occurrence of 'indeterminate' refers to indeterminacy. L** is false.

c) Both the occurrence of 'indeterminate' and the occurrence of 'false' in L** are indeterminate in reference. L** is indeterminate in truth value. But no contradiction follows since it is indeterminate what L** says.

In each of these cases, the question 'to what do the aberrant occurrences refer?' is an interesting and difficult one. Unfortunately, I will have to postpone full discussion of these questions to another occasion.[45] For now, I want to address an even more pressing question.

Let us assume that either the occurrence of 'indeterminate' or the occurrence of 'false' in L** witnesses an aberration. Which occurrence does this? Or do both of them do it? This

[45] Here is my preferred way to develop (c). Let's say a sentence is *ascending false* if it is safe and not ascending true. (Recall Scharp's definition of safety from my Section 1.2.1.2. See (Scharp 2013), p.170.) Likewise, let's say a sentence is *descending false* if it is safe and not descending true. Now, let's say a sentence is *ascending indeterminate* if it is indeterminate whether or not the sentence is ascending true. And likewise, say a sentence is *descending indeterminate* if it is indeterminate whether or not the sentence is descending true. Then the idea would be that the reference of the occurrence of 'indeterminate' in L** is indeterminate as between ascending indeterminacy and descending indeterminacy, and the reference of the occurrence of 'false' in L** is indeterminate as between ascending falsehood and descending falsehood. I thank Harold Hodes for this suggestion.

question is an instance of the more general phenomenon of *parity problems*, which will crop up frequently in what follows. It is worth taking a brief moment to introduce these problems, which I will now do. There are many cases in which we are faced with a pair of sentences, words, or occurrences that are quite similar in nearly all respects that seem relevant for diagnosing their semantic statuses—statuses like *true*, *false*, or *indeterminate*—but such that we face significant pressure to attribute different statuses to them. The problem of attributing semantic statuses is a *parity problem*. Solving a parity problem amounts to providing a non-paradoxical diagnosis, plus an explanation of how the expressions involved acquire the statuses they have under that diagnosis. As I proceed, it will emerge that fans of radical and moderate aberrationism will want to diagnose parity problems in characteristically different ways. Radical aberrationists will find it easier to assign different statuses, while moderate aberrationists will find it easier to assign the same status.

Return now to the case of L** (that is, 'L** is either indeterminate or false'). The question was, assuming that at least one occurrence in L** witnesses an aberration, and assuming that it could be either the occurrence of 'false' or that of 'indeterminate', which is it? Other things being equal, it is best to minimize the aberrations one posits to what is necessary for avoiding paradox. Thus, at least on that count, it is better to posit only one aberration. However, other things are not equal. It is hard to identify any feature of our use of either 'false' or 'indeterminate' that would make that word a more likely candidate for undergoing an aberration. And neither expression is such that its behavior seems more responsible for the revenge problem than that of the other. Finally, we have assumed that each expression is such that targeting it for a one-off aberration can make for a consistent diagnosis of L*; considerations of consistency do

not force us to choose one or the other. Thus, there are pressures to assign both occurrences the same status. The presence of these pressures amounts to a parity problem.

Different aberrationists are free to take different stands on this problem. First consider radical aberrationists. Because they deny Semantic Supervenience or Semantic Regularity, radical theorists are at greater liberty than their moderate counterparts to posit differences in the reference of words (or occurrences) in cases in which there are no seemingly relevant differences in our uses of these words. So, at the cost of treating semantic facts as *sui generis*, one can assign different statuses to the occurrence of 'false' and the occurrence of 'indeterminate' in L**. Again, that is, one occurrence witnesses an aberration and the other does not.

Which occurrence gets which status? It is hard to see how to defend either of the two possible answers here. Normally, semantic claims are defended by appealing to facts about how expressions are used. However, in the present case we have already observed that nothing in the use of either expression justifies targeting either expression rather than the other, and in any case a radical aberrationist will tend to be skeptical about attempts to derive semantic facts from facts about language use. For her, then, the most promising strategy here is to endorse a sort of epistemicism, according to which when it comes to sentences like L** there is only one aberration, but no one can know which expression undergoes it. Since I myself do not want to reject Semantic Supervenience, I will leave the defense of this position here. But what I have said should suffice to make clear that radical aberrationists have a way to respond to the parity problem posed by L**, if there is one. (Again, I assumed for the sake of argument that we can consistently target either expression for an aberration. There may be more to say about this, but I cannot pursue the matter here.)

A less epistemological way of respecting our sense that the occurrences of 'indeterminate' and 'false' in L** are on equal footing is, of course, to say that both occurrences undergo aberrations. The obvious problem for this diagnosis is that it fails to minimize the aberrations posited to exactly what is necessary for avoiding contradictions. However, one might insist that the need to minimize aberrations is not the only constraint that a diagnosis must satisfy in order to be adequate; it must also assign comparable statuses to expressions which are comparable in relevant respects, such as use. This is an appealing view for people who, like me, accept Semantic Supervenience and Semantic Regularity. Another reasonable consideration is the following: because 'indeterminate' and 'false' share a boundary, one expects that the occurrence of 'false' shifts reference if and only if the occurrence of 'indeterminate' does so.

Still, it is worth noting that positing two aberrations is not the only way to assign comparable statuses to L**'s occurrences of 'indeterminate' and 'false'. Another option is to claim that while one of these occurrences witnesses an aberration, it is indeterminate which one that is. An objection to this view is that our linguistic behavior could not determine that there is an aberration without determining which one it is. Here, an aberrationist might stress that when one compares cases of indeterminacy of reference, the concern looks unconvincing. For we typically do want to say, of expressions that are indeterminate in reference as between just two candidates, that these expressions refer; though we deny that the use of these expressions makes it determinate exactly to what they refer. If we can say that in those cases, then we can also say it in the present case.

Another objection is that it is unsatisfying not to be able to state definitively of either occurrence that it does, or that it does not, witness an aberration. Of course, as a quite general feature of the notion of indeterminacy, if one holds that it is indeterminate whether X or Y

obtains then one is committed against definitively stating whether it is X or Y that obtains. Still, an objector might insist, the awkwardness of this commitment creates pressure to minimize the positing of indeterminacy. Aberrationists have no special answer to this objection. It is simply another problem whose seriousness should be evaluated in light of the problems that afflict alternative approaches to the Liar paradox. Though it is worth noting that the concern is particularly biting for aberrationists who also want to claim that the expression that undergoes an aberration (whichever expression that is) is indeterminate in reference. Such a position in effect posits three different layers of indeterminacy: indeterminacy as to which expression undergoes an aberration, indeterminacy in the reference of that expression, whichever expression it is, and then indeterminacy as to the truth value of the sentence containing the expression.

3.5. Ascending truth and Descending truth

As we saw, my own preferred version of moderate aberrationism appeals to the properties *ascending truth* and *descending truth*.[46] That is how I flesh out my claim that sentences like L* come close to saying what they appear to say, and that what these sentences say is influenced by how we use them and their component words. However, proponents of this version of moderate aberrationism must consider whether ascending truth and descending truth can be used to generate any new, recalcitrant, Liar-like paradoxes. Before answering this question in the negative, let me emphasize that the question is quite particular to my own preferred version of moderate aberrationism. Even if my answers should prove unsatisfactory, that need not affect the fate of moderate aberrationism in general.

[46] My talk of ascending truth and descending truth raises the possibility of a different view, on which the reference of every occurrence of 'true' is indeterminate between these different properties, but for most purposes this does not matter since it is only on paradoxical contexts that the properties diverge from our pre-theoretical judgments of truth. I address this view in Chapter 1, Section 9.

When it comes to ascending truth and descending truth, an obvious first place to look for revenge is the sentences α and δ from Section 1.2.1.2:

(α) α is not ascending true

(δ) δ is not descending true

However, as we already saw, these sentences turn out to be quite unproblematic. As I explained shortly after introducing α and δ, every sentence which says of itself only that it is not ascending true is ascending true (and thus false). So α is simply false; no paradox here. Similarly with δ. Every sentence that says of itself only that it is not descending true is not descending true, and is thus true. So δ is not descending true, and so it is simply true; no paradox here.

One could experiment with more complex examples, but a moment of mindful reflection suggests this will be futile. As I explained in Section 1.2.1.2, (Scharp 2013) Chapter 6 and its appendix present a consistency proof for ADT. This proof takes the form of construction of a set-theoretic model, M_2. Thus, relative to their interpretation in M_2, ADT and all of its classical consequences are consistent (relative to set theory). We can take the existence of M_2 as evidence that ADT is consistent when 'ascending true' and 'descending true' are understood as predicates added to English, defined by adopting the elements of ADT as axioms. This line of reasoning provides a straightforward response to anyone claiming to have identified a problematic sentence involving 'ascending true' and 'descending true'.[47] For one thing, provided that Scharp's consistency proof is correct, one cannot prove a contradiction from just this sentence, ADT, and the classical inference rules. Secondly, then, whatever additional principles one used to arrive at a contradiction are simply false in every model of ADT, since by hypothesis they are

[47] It is worth noting that this argument also addresses the closely related concern that one can use ascending truth and descending truth to define truth, and arrive at a paradox that way. Suppose we have some formula φ such that 'for all x, φ(x) if and only if x is true' is provable from ADT (and classical logic). Then using φ one can prove a contradiction from ADT. But this contradicts Scharp's result that ADT has a model.

syntactically inconsistent with ADT. That gives us reason to believe that these additional principles will come out false if 'ascending true' and 'descending true' are added to English and defined by stipulating the elements of ADT as axioms.

So much, for now, for the issue of revenge. I will return to this matter in Section 5, where I will examine whether two putative definitions of 'Liar-like sentence' that I develop give rise to revenge. In the meantime, I will consider a variety of different, pre-theoretically Liar-ish paradoxes and briefly explain how an aberrationist diagnosis can be applied to them.

4. Other Liar-like Paradoxes

Every approach to the Liar paradox is introduced by applying it to some particular version of the paradox. However, it is a fatal flaw if a putative solution works only for the specific version that is used to introduce it; to be satisfactory, a solution must be applicable to many different versions. In this section, therefore, I will show how the essential moves of aberrationism can be successfully used to diagnose and solve a number of paradoxes that are distinct from the Strong Liar, the version of the paradox with which I began. Then, in Section 5, I will examine two ways to define the term 'Liar-like sentence' that would allow aberrationism as defined in Section 1 to apply to these paradoxes.

A quick word on the motivation for discussing the many paradoxes below. Of course, if an approach to the Liar paradox can also solve lots of other, distinct paradoxes, then that can only be a point in its favor. But on the other hand, if an approach is defective as a solution to the Liar then it cannot compensate for this by being applicable to lots of other paradoxes. For this reason, when assessing different approaches, it matters what exactly counts as a Liar-like paradox. Still, while there is room for disagreement about exactly what paradoxes count as Liar-

like, in practice it is uncontroversial to count all the paradoxes that I will discuss as Liar-like. (See Section 5.1.1 for a definition of 'Liar-like paradox' that builds on their common features.) As I proceed, I will show that in cases in which the solution afforded by aberrationism comes at a significant cost, the overall balance of considerations nevertheless favors aberrationism, and within that, moderate aberrationism.

4.1. Contingent Liars

Imagine that sentence S_C below, and nothing else, is written on a portable blackboard, *B*:[48]

(S_C) The sentence written on the blackboard in room 103 is not true.

Let C be a context in which *B* and no other blackboard is in room 103. Then S_C is a Liar sentence, relative to C. For relative to C, the definite description 'the sentence written on the blackboard in room 103' refers to S_C itself, and this makes S_C the negation of a sentence whose grammatical subject refers to S_C and whose grammatical predicate is an expression that refers to truth—a Liar sentence.

On the other hand, imagine a different portable blackboard, *B**, containing nothing but the sentence 'Snow is purple'. And let C* be a context in which *B** and no other blackboard is in room 103. Then relative to C*, S_C is simply true. Similarly, imagine that the sentence 'Snow is white' is written on a blackboard *B***, and let C** be a context in which *B*** alone is in room 103. Relative to C**, S_C is simply false; no paradox here. These examples demonstrate that the definite description 'the sentence written on the blackboard in room 103' is context-sensitive,

[48] This example is taken from (Simmons 1993), p.101.

and that the referent, in a given context, of this description determines whether S_C is true, whether it is false, and whether or not it is a Liar sentence.

These examples also demonstrate something that is widely recognized by philosophers: that reference, at least for definite descriptions, must be relativized to contexts of use. Rather than being a two-place relation between an expression and an object or property, reference is now widely regarded as being, strictly speaking, a three-place relation between an expression, a context of use, and an object or property. Thus, rather than simply referring to S_C, the description 'the sentence written on the blackboard in room 103' refers to S_C relative to C, to 'Snow is purple' relative to C*, and to 'Snow is white' relative to C**. For the same reasons, one who speaks of reference by occurrences will want to say that the occurrence of that description in S is context-sensitive: in C, it refers to S_C, in C* it refers to 'Snow is purple', and in C** it refers to 'Snow is white'.

As we saw at the beginning, whether or not a sentence is a Liar sentence depends on the reference of its grammatical subject and predicate expressions. If the reference of these expressions is sensitive to context of use, then *being a Liar sentence* must be not a property but a relation, one of whose relata is a context of use. And presumably, the same should hold for all manner of Liar-like sentences. (It is easy to verify that my definition of 'Liar-like' in Section 5 below can be trivially modified so as to do justice to this desideratum.) Aberrationism must then be modified accordingly, so that it posits aberrations in a sentence relative to all and only the contexts in which that sentence is Liar-like. Let *CS-aberrationism* (for "context-sensitive aberrationism") be the conjunction of the following three claims:

(Context Sensitive Occurrences)	Occurrences refer only relative to contexts of utterance

(Context Sensitive Aberrations) For any sentence S and any context of utterance C, if S is Liar-like relative to C then, relative to C, the key occurrences of the alethic expressions in S differ in reference (relative to C) from the expressions of which they are occurrences.

(Context Sensitive Determined) For any sentence S and any context of utterance C, if S is Liar-like relative to C then what S says relative to C is determined by the reference-in-C of the key occurrences of S's alethic expressions, rather than by the reference-in-C of the expressions of which they are occurrences.

Returning to the sentence S_C from above, CS-aberrationism says that relative to any context in which the sentence in room 103 is S_C itself, the occurrence of 'true' in S_C fails to refer to truth; but relative to all other contexts, this occurrence refers to truth. (For my own part, I would add that in any context C in which S's occurrence of 'true' or similar fails to refer to truth, the reference of this occurrence is indeterminate as between ascending truth and descending truth. This allows that in C, S_C comes close to saying what it appears to say (in C).) Thus, CS-aberrationism allows us to avoid paradox in precisely the same way as regular aberrationism. It is only for simplicity of presentation that I have suppressed issues of context-sensitivity throughout, defending a de-contextualized version of aberrationism rather than CS-aberrationism.[49]

A final, different concern is that relativizing reference by (some) occurrences to contexts gives rise to a new revenge problem. That would involve quantifying over contexts and mentioning reference by occurrences, as in the following variant on sentence X from Section 3:

(X***) In no context does the unquoted occurrence of 'true' in X*** refer to anything which would render X*** true.

[49] One might reasonably wonder how CS-aberrationism is different from classic contextualist solutions in the literature. For discussion of this issue, please see Chapter 1, Section 6.3.

However, aberrationism can handle X*** much as it handles other variants on X. The one difference is that we must now admit that in some contexts other than the present one, 'X***', and likewise its occurrence in X***, refers not to X*** but to something else. (Keep in mind: in the present context 'X***' refers to the sentence-type displayed above.) Relative to any context in which 'X***' does not refer to X***, there is no danger whatsoever of revenge. And relative to other contexts, such as the present one, aberrationism applies. That is, for any context C*** in which 'X***' refers to X***, the occurrence of 'true' in X*** fails to refer to truth relative to C***. Thus, in no context does X*** say anything from which it follows that it (X***) is not true in any context.

4.2. Curry's Paradox:

A famous Liar-like paradox is the Curry paradox, which arises from consideration of *Curry sentences*. For all sentences S, S is a Curry sentence if and only if S is a conditional whose antecedent is a sentence that attributes truth to S and whose consequent is something entirely irrelevant to the antecedent.[50] Cu below is a good example:

(Cu) If Cu is true then grass is purple

As is the case with any Curry sentence, if indeed Cu says what it appears to say, then from the empty set of assumptions one can prove Cu's conclusion:

1. Cu = 'If Cu is true then grass is purple'	(definition of 'Cu')
2. |Cu is true	(assumption)
3. |'If Cu is true then grass is purple' is true	(Substitution, (1), (2))
4. |If Cu is true then grass is purple	((T-out), (3))
5. |Grass is purple	(conditional elimination, (2), (4))
6. If Cu is true then grass is purple	(conditional introduction, (2)-(5))

[50] One might have thought that only sentences that lead to unacceptable conclusions can count as Liar-like. However, if a sentence allows for the (apparent) deduction of an irrelevant conclusion, then the pattern of reasoning to which the sentence gives rise can easily be modified to obtain an unacceptable conclusion. See my definitions of 'paradox' and 'Liar-like$_{DEF}$' in Section 5.1.2.

7. ‘If Cu is true then grass is purple’ is true ((T-in), (6))
8. Cu is true (Substitution, (1), (7))
9. Grass is purple (conditional elimination, (6), (8))

It is fairly clear what aberrationists should say about Curry sentences: the occurrence of ‘true’ in the antecedent of any Curry sentence fails to refer to truth. Cu thereby fails to say what it appears to say, and so entering lines 5 and 6 is illegitimate. The legitimacy of these steps requires that Cu say what it appears to say, which it does not. More generally, the admissibility of an inference depends not only on the syntax of the sentences involved, but also on what propositions (if any) they express. Compare: if we adopted a convention according to which the 4th sentence in any derivation meant that grass is blue, then step 4 in the above would likewise be inadmissible, despite being licensed by the syntax of the sentences involved.

4.3. The Bad Pair:

The paradox associated with the following pair of sentences is like the Strong Liar, except that it does not involve direct self-reference.

(A) B is not true.

(B) A is true.

Suppose A is true. Then B is not true. But B says that A is true; so, for B not to be true, A must not be true. Contradiction! Thus, A must not be true. But A says that B is not true; so, if A is not true then B is true. But B says that A is true; so, if B is true then A is true. A contradiction![51]

[51] Here is a more formal presentation, leading to the same conclusion (continued on next page):

1. A = ‘B is not true’ (definition of ‘A’)
2. B = ‘A is true’ (definition of ‘B’)
3. |A is true (assumption)
4. |‘B is not true’ is true (substitution, (1), (3))
5. |B is not true ((T-elim), (4))
6. |‘A is true’ is not true (substitution, (2), (5))
7. |‘A is true’ is true ((T-intro), (3))
8. |Contradiction (contradiction introduction, (6), (7))

Aberrationism gives us several ways to avoid paradox in cases like this. On one diagnosis, A fails to say what it appears to say, but B does not; that is, the occurrence of 'true' in A fails to refer to truth, but the occurrence of 'true' in B refers to truth. Another option is that B fails to say what it appears to say, but A does not: the occurrence of 'true' in B fails to refer to truth, but the occurrence of 'true' in A refers to truth. A third option is to claim that neither A nor B says what it appears to say; both occurrences witness aberrations. And a fourth option is to say that while there is just one aberration here, it is indeterminate which occurrence witnesses it. (I leave it to the reader to check that all these diagnoses avoid paradox.)

By now, these options are familiar, as are the considerations weighing for and against each. As we saw with the putative revenge paradoxes involving L** in Section 3.4, other things being equal it is best to posit only such aberrations as one must to avoid paradox. A appears to say of B that it (that is, B) is not true, and B appears to say of A that it (that is, A) is true; and surely, other things being equal, the more one can respect these appearances the better. However, as we also had with L**, here we have a parity problem. It is hard to conjure up any principled reason for targeting one occurrence of 'true' that does not also justify targeting the other. A and B play equal and comparable roles in the derivation of a paradox; we are no less sanguine about steps that make use of A than we are about steps that make use of B, and vice versa. Only a diagnosis that gives A and B the same status can respect our equal treatment of these sentences.

9.	A is not true	(negation introduction, (3)-(8))
10.	|B is true	(assumption)
11.	|'A is true' is true	(substitution, (2), (10))
12.	|A is true	((T-elim), (11))
13.	|Contradiction	(contradiction introduction, (9), (12))
14.	B is not true	(negation introduction, (10)-(13))
15.	'B is not true' is true	((T-intro), (14))
16.	A is true	(substitution, (1), (15))
17.	Contradiction	(contradiction introduction, (9), (16))

The situation for different aberrationists is the same as it was for the earlier parity problems we saw. Radical aberrationism is freer to ignore similarities in use, and insist that only one occurrence witnesses an aberration. These theorists can avoid having to explain which one it is by embracing an epistemicist view, according to which it is impossible to know which occurrence witnesses the aberration. On the other hand, moderate approaches are more congenial to diagnoses that assign the same status to the occurrences of 'true' in A and B, since they are committed to respecting the links between reference and language use. These approaches must choose between positing two aberrations or claiming that there is one aberration and holding that it is indeterminate which occurrence witnesses it.

4.4. The No-No Paradox:

The following "paradox" is a variant of one much discussed by Roy Sorensen, e.g., in (Sorensen 2001). It begins with the following sentences:

(A*) B* is not true.

(B*) A* is not true.

Note first that one can consistently allow that one of these sentences is true and the other false, or allow that one is true and the other lacks a truth value. Accordingly, an available position is that there is no genuine paradox here, since there are assignments of truth values that one can consistently make.

Paradox or no, the No-No sentences confront us with a very real parity problem. A* and B* are symmetrical with respect to all properties that seem relevant for attributing truth values, or lack thereof, and so any asymmetrical diagnosis is apt to seem quite arbitrary. Yet, as long as one allows that A* and B* say what they appear to say, one cannot consistently attribute the

same semantic status to both. If A* and B* are both true, then since A* is true, B* is not true, a contradiction. And one can reason similarly with A* and B* reversed. Likewise, if A* and B* are both false, then since B* is then not true, A* is true after all, a contradiction (since A*'s being false entails that it is not true). Again, one can reason similarly with A* and B* reversed. And finally, if A* has no truth value, then in particular it is not true. That would make B* true! Similarly with A* and B* reversed. So, if either sentence lacks a truth value then the other is true.

As we had with earlier parity problems, No-No sentences sit differently with different aberrationists. Let us start with the radical theorists—those that deny either Semantic Supervenience or Semantic Regularity. Such philosophers are at greater liberty to make an asymmetric attribution of truth values, and they can do this without positing any aberrations. That is, they can allow that both A* and B* say what they seem to say, and simply hold that one is true and the other false. To avoid having to explain which is which, they can make the by-now-familiar appeal to epistemicism, claiming that it cannot be known which of A* and B* is true and which untrue.[52] They can say all this, because it is part of their general outlook to be happy with divorcing the semantic properties of A* and B* from the properties that these sentences so plainly have in common, such as their syntactic and compositional properties. I find this general outlook to be deeply flawed, but a detailed critique is a project for another occasion.

By contrast with radical aberrationists, moderate theorists are less inclined to ignore the similarities between A* and B*, and more inclined to respect the pressure that most of us feel to attribute the same status to these sentences. Still, there may in the end be ways for such theorists

[52] Of course, it is also open to a radical theorist to claim that there is a single one-off aberration here. But that solution carries the cost of positing an aberration without providing the benefit of avoiding arbitrariness. So, I'll say no more about it. Radical theorists can also posit two aberrations, but they have little reason to, given that their general outlook justifies ignoring the similarities between A* and B*.

to resist this pressure and give an asymmetric diagnosis, under the force of other considerations. In such a case, it would even be open to them to take an epistemicist stance as a way of defending such a diagnosis. Going this route, they would say that there are some non-semantic facts that determine, according to some unknown patterns, that A* and B* have different truth values, but that it is impossible to know which facts and patterns these are.

However, the asymmetric route is rockier for a moderate theorist than for her radical cousins. It is hard to imagine what the difference-making non-semantic facts could be, given that all the relevant-seeming facts we can observe point to similarity between A* and B*. Thus, it is worth noting that several different ways of positing one-off aberrations allow one to assign A* and B* the same status. The options here are the ones familiar from the other paradoxes we have seen: either both of the occurrences of 'true' witness aberrations, or one does but it is indeterminate which one that is.

These diagnoses are not without complications. One might hold that the No-No sentences do not present us with a genuine paradox; rather, what we have is simply a *reductio* of the idea that sentences which are similar in the manner of A* and B* always have the same semantic status. On that sort of view, one might hold that, all things considered, making an asymmetric assignment of semantic values is less costly than positing a one-off aberration. On the other hand, if asymmetric assignments of semantic values are unacceptable—whether because Semantic Regularity is true or for some other reason—then the No-No sentences confront us with a genuine paradox. In that case, there is as much justification for positing a one-off aberration here as there is in the case of the other paradoxes we have seen.

4.5. Yablo's and Cook's Paradoxes:

(Yablo 1993) develops a Liar-like paradox that does not rely on self-reference. The paradox involves an infinite list of sentences:

(S_1) For all $n \geq 2$, S_n is not true.
(S_2) For all $n \geq 3$, S_n is not true.
…
(S_m) For all $n \geq m + 1$, S_n is not true.
…

A few moments' reflection on these sentences generates a paradox that is palpably Liar-like.[53] But it is clear that none of the sentences contains any expression referring to that sentence.

While Yablo's paradox does not strictly speaking involve self-reference, (Cook 2006) describes a more general kind of circularity—*fixed points*—that it does involve. (For my purposes here, it does not matter exactly what fixed points are. See Cook's paper for a definition.) Using a language that allows for infinite conjunctions, Cook is able to construct a palpably Liar-like paradox that does not even involve fixed points. Here are the relevant sentences:

(S_1') S_2' is not true & S_3' is not true &…
(S_2') S_3' is not true & S_4' is not true &…
…
(S_m') S_{m+1}' is not true & S_{m+2}' is not true &…
…

The paradox-generating reasoning here is quite similar to that associated with Yablo's paradox.

The question now is, what should aberrationism say about the sentences in these paradoxes? To start, there is reason to think that all the sentences must have the same semantic status—truth, untruth, falsehood, indeterminacy, witnessing an aberration, etc. This is because given any sentence in the list, what is essentially the same paradox can be raised for the sequence

[53] Here is how to get a contradiction. Given any number m, suppose S_m is true. Then for all $n \geq m + 1$, S_n is not true. This includes S_{m+1}. But then also, for all $n > m + 2$, S_n is not true. That makes S_{m+1} true after all, a contradiction. So, it must be that for all m, S_m is not true. But then in particular, for any $m \geq 2$, S_m is not true. Thus, S_1 is true, a contradiction! Now we are in trouble.

of sentences that succeed it. Accordingly, then, any argument that justifies attributing a given status to one of the sentences can be mustered to justify attributing this status to any other sentence in the list.[54]

By this reasoning, then, the only option for aberrationism is to deny that any of these sentences says what it appears to say. In particular, moderate approaches of the kind I favor will say that the occurrences of 'true' in each of these sentences are indeterminate as between referring to ascending truth and referring to descending truth, and so all the sentences are indeterminate. (Keep in mind: that does not make any of them true after all, since none succeeds in saying of all its successors that they are not true.)

4.6. Grelling's Paradox

All of the paradoxes we have seen so far involve the notion of truth. However, there is a natural generalization of this notion which also gives rise to a paradox that is palpably Liar-like. To get the idea, note first that just as 'true' is governed by (T-intro) and (T-elim), we can formulate a predicate, 'satisfies', which applies to ordered tuples of arbitrary arity, and which is defined by the following rules:[55]

---

[54] In particular, one might have thought that the first sentence is different from the rest, since it, and only it, has no predecessors. But as just explained, any other sentence can simply be treated as the first sentence in a new list, consisting of all subsequent sentences. For contrast, consider the following sentences, brought to my attention by Matti Eklund:

(S_1*) It is not the case that for all $1 \leq n < 1$, if S_n* exists then S_n* is true
(S_2*) It is not the case that for all $1 \leq n < 2$, if S_n* exists then S_n* is true
(S_3*) It is not the case that for all $1 \leq n < 3$, if S_n* exists then S_n* is true
…

S_1* has no predecessors, and so (trivially) all of its predecessors are true, making S_1* false. But then for each $n > 1$, S_n* is true, since S_n* has at least one false predecessor (namely, S_1*). The difference between this case and Yablo's and Cook's paradoxes is that here the sentences talk about their predecessors rather than their successors. This enables the S_1*'s lack of predecessors to affect its truth value, and via this, the truth values of all subsequent sentences. In Yablo's and Cook's paradoxes, by contrast, each sentence addresses only those that succeed it, and so the first sentence's lack of a predecessor is irrelevant to its truth value and to those of all subsequent sentences.

[55] It is worth noting that the notion of satisfaction is quite similar to the notion of respect, introduced in Section 1.2.1.2. The difference is that respect is defined in terms of a modal condition on reference and in terms of truth,

(Sat-intro) $\dfrac{\varphi(a_1,\ldots,a_n)}{\text{satisfies } (\langle a_1,\ldots,a_n\rangle, \text{`}\varphi(x_1,\ldots,x_n)\text{'})}$

(Sat-elim) $\dfrac{\text{satisfies } (\langle a_1,\ldots,a_n\rangle, \text{`}\varphi(x_1,\ldots,x_n)\text{'})}{\varphi(a_1,\ldots,a_n)}$

Suppressing issues of context-sensitivity, here the 21st letter of the Greek alphabet is a substitutional variable ranging over all formulas of English, for each Arabic numeral, the 24th letter of the lowercase Latin alphabet subscripted by that numeral is a substitutional variable ranging over all objectual variables in English, and for each Arabic numeral, the 1st letter of the lowercase Latin alphabet subscripted by that numeral is a substitutional variable ranging over all English singular terms.

A moment's reflection reveals (T-intro) and (T-elim) to be, modulo some trivial cosmetic differences, the special cases of (Sat-intro) and (Sat-elim) where n = 0. This follows straightforwardly from the fact that declarative sentences are formulas with no free variables; when n = 0, for example, 'satisfies ($\langle a_1,\ldots,a_n\rangle$, 'S($x_1,\ldots,x_n$)')' becomes 'satisfies ('S')'. (The trivial cosmetic differences are those between, for example, 'satisfies ('S')' and ''S' is satisfied', and between ''S' is satisfied' and ''S' is true'.[56])

With (Sat-intro) and (Sat-elim) on the table, we can articulate Grelling's Paradox, displayed below:

whereas the notion of satisfaction is introduced simply by stipulating that the rules (Sat-intro) and (Sat-elim) are henceforth to define the word 'satisfies'.

[56] Setting the two rules side by side makes the point obvious. E.g.:

(T-intro) $\dfrac{S}{\text{'S' is true}}$

(Sat-intro) $\dfrac{S}{\text{satisfies ('S')}}$

1.	|satisfies (‘~satisfies (x, x)’, ‘~satisfies (x, x)’)	(assume for *reductio*)
2.	|~satisfies (‘~satisfies (x, x)’, ‘~satisfies (x, x)’)	(Sat-elim), (1)
3.	|Contradiction	Contrad. Intro, (1), (2)
4.	~satisfies (‘~satisfies (x, x)’, ‘~satisfies (x, x)’)	*Red. ad abs.*, (1)-(3)
5.	satisfies (‘~satisfies (x, x)’, ‘~satisfies (x, x)’)	(Sat-intro), (4)
6.	Contradiction	Contrad. Intro, (4), (5)

In what follows, let ‘U’ (for “unsatisfied”) be the name of the sentence written in steps (2) and (4):

(U) ~satisfies (‘~satisfies (x, x)’, ‘~satisfies (x, x)’)

One can quickly see that aberrationism can provide a solution to Grelling’s paradox. All we need to do is expand the notion of an alethic expression from the definition of aberrationism in such a way so that instead of applying only to expressions that are intersubstitutable with ‘true’, it also includes ‘satisfies’. This allows us to count U as a Liar-like sentence, and then we can say that the unquoted occurrence of ‘satisfies’ in U fails to refer to satisfaction, since this is a key occurrence of an alethic expression in a Liar-like sentence.[57] It then follows that U fails to say of the predicate ‘~satisfies (x, x)’ that it fails to satisfy itself. Thus, the sentences written on lines (1) and (2) fail to express genuinely contradictory propositions. Similarly for lines (4) and (5). Thus, the derivation does not go through.

This concludes my presentation of different pre-theoretically Liar-ish paradoxes. In the next section, I will work toward a definition of ‘Liar-like sentence’ that will allows aberrationism to apply to all the paradoxes above.

[57] These claims are uncontroversial. Thus, making them come out true will be a desideratum on our definitions of ‘Liar-like sentence’ and ‘key occurrence’, once we have occasion to develop those.

5. Articulating Aberrationism in Full Generality:

5.1. The Classification Problem:

An important task for any diagnosis of the Liar paradox is to solve the *classification problem*: to identify in a precise way all and only the sentences to which that diagnosis applies. This task is even more important for diagnoses that are simultaneously to serve as solutions, as aberrationism purports to do. A diagnosis that does not apply to the sentence(s) that generate a given paradox cannot be used to solve that paradox. And, as discussed at the beginning of last section, any adequate solution to the Liar paradox must solve all Liar-like paradoxes, not just the ones that are used to introduce that solution. Thus, a diagnosis cannot count as an adequate solution unless it applies to every sentence that gives rise to some Liar-like paradox.

How can one can identify all and only the sentences that give rise to Liar-like paradoxes? A simple strategy is just to list them. For instance, an aberrationist might say that if a sentence is a Classic Liar sentence, a Strong Liar sentence, a Yablo sentence, etc., then the key occurrences of its alethic expression(s) undergo one-off aberrations. However, such disjunctive formulations are bound to be unsatisfying. For one thing, there are many Liar-like paradoxes, and so in practice there is always the danger of failing to include some overlooked category of sentences, namely, those that give rise to an overlooked version of the paradox. There is also another, deeper concern. Ideally, a diagnosis will have a criterion of application that brings out a shared, underlying feature of the sentences in question, one that explains why they all (appear to) induce paradox. Unless a diagnosis bases its applicability on such a feature, there is a worry that it fails to illuminate the underlying nature of Liar phenomena.

In lieu of the simple listing strategy, the task of specifying all and only the sentences that give rise to Liar-like paradoxes can be approached from two different perspectives. One

perspective takes the second concern raised above to heart: it seeks to identify a shared, underlying feature of the Liar-like sentences, one that explains why they all (appear to) induce paradox. Philosophers with this perspective will be naturally inclined to pursue the *abstraction strategy*: consider a wide variety of different Liar-like paradoxes, note the sentences that give rise to them, and then try to abstract out a common feature of these sentences, one that explains how they give rise to these paradoxes. If the strategy yields a correct answer, then it will identify a feature the possession of which is both necessary and sufficient for giving rise to a Liar-like paradox. I explore this strategy in Section 5.1.1 below.

It is worth noting that the abstraction strategy is compatible with aberrationism. Aberrationism claims that all and only the sentences that give rise to Liar-like paradoxes witness referential aberrations in the key occurrences of their alethic expressions. This claim is consistent with there being no interesting, shared feature of these sentences that explains why they all give rise to these paradoxes. But it is also consistent with there being such a feature.

The abstraction strategy faces a challenge, since in principle there is room for disagreement as to exactly which paradoxes count as Liar-like. That creates unclarity as to exactly which sentences give rise to Liar-like paradoxes, and, in turn, unclarity about which sentences a diagnosis must apply to in order to be able to count as a solution. In light of the possibility of such disagreement, I can see only one method for identifying the sentences that must be targeted: acquiesce to one's pre-theoretical impressions about which paradoxes are Liar-like, and then look for common, problem-causing features of the sentences that give rise to these paradoxes. That is the method I will employ in Section 5.1.1.

So much, for now, for the abstraction strategy and the perspective that motivates it. A different perspective rejects the demand for an underlying feature that explains why various

sentences give rise to paradoxes. Rather, on this view, being a Liar-like sentence is simply identical with being a sentence that gives rise to a Liar-like paradox. (For convenience, but only for convenience, this is the definition of 'Liar-like' that I adopted above.) So, to get clear on which sentences are Liar-like, the important task is to define 'Liar-like paradox' and 'gives rise to' with some measure of clarity and precision. Call this task the *definition strategy* for identifying all and only the Liar-like sentences. I will explore this strategy in Section 5.1.2.

Like users of the abstraction strategy, users of the definition strategy also need to acquiesce to their pre-theoretical impressions about which paradoxes are Liar-like. For the way that they refine their definitions of 'Liar-like paradox' is by checking against examples, and an example paradox can only be used to test a putative definition of 'Liar-like paradox' if it is known whether or not the example falls under that classification. That said, as with the abstraction strategy, the appeal to pre-theoretical impressions here is not a serious concern in practice, as philosophers' pre-theoretical impressions tend to agree on most of the cases discussed.

Although they are naturally associated with two different perspectives, the abstraction strategy and the definitions strategy can be pursued in tandem. For all that has been said, the definition of 'gives rise to a Liar-like paradox' yielded by the definitions strategy may specify a deep, explanatory feature of the sentences to which it applies. And even if not, even philosophers in search of such a deep, explanatory feature will in the meantime find it helpful to have an extensionally adequate characterization of the sentences that they need to target in order to solve the paradox.

Before moving on and exploring the two strategies, I want to emphasize that because the task of formulating general definitions of 'Liar-like paradox' and 'Liar-like sentence' is shared

by many different philosophers who study the Liar, it is acceptable that the project of developing fully satisfactory definitions lies ultimately beyond the scope of this paper. The work of the next two subsections should be seen in the spirit of providing working approximations.

5.1.1. The Abstraction Strategy

As noted above, the abstraction strategy begins by acquiescing to one's pre-theoretical impressions as to the Liar-like-ness of various paradoxes. Of those paradoxes discussed above, the only ones about which I can envision the slightest controversy are the Yablo, Cook, Bad Pair, Curry, and Grelling Paradoxes.[58] I myself am inclined to count all of these as being Liar-like. In the rest of this subsection, I will explain the consequences of this choice. I will begin by developing a proposed necessary condition for being Liar-like, motivated by reflection on some of these sentences, and then move toward refining it into a necessary and sufficient condition.

Looking at the Strong Liar, Yablo, Cook, and Bad Pair sentences, each of these contains an occurrence of 'true' within the scope of negation. Thus, one might be tempted to say that a necessary condition for a sentence's being Liar-like is that it contain such an occurrence. However, the Grelling sentence contains in the scope of negation not 'true' but 'satisfies'. Since we can view 'true' as a special case of 'satisfies' (the case where the relation symbol being satisfied is 0-ary), we can treat all the other sentences just mentioned as involving occurrences of 'satisfies', modulo some trivial cosmetic differences. So, for inclusiveness, we can reformulate the condition as saying that a sentence should contain an occurrence of 'satisfies' in the scope of the negation operator.

[58] The No-No "paradox" raises complicated issues, which I will discuss in Section 5.1.2.

Even this newly qualified condition will not do, however, since there are Liar-like sentences which contain neither 'true' nor 'satisfies'. E.g., relative to the assumption that truth is the most philosophically interesting property, the sentence L_1 below is a Liar sentence:[59]

(L_1) ~(L_1 has the most philosophically interesting property)

Notice, however, that since (T-intro) and (T-elim) are needed to arrive at a contradiction, and since the predicate 'has the most philosophically interesting property' is not directly defined in terms of these rules, we need to substitute 'true' for 'has the most philosophically interesting property' in order to generate a paradox.[60] Let us now say that an expression *e* is an *alethic expression* if

(Base Clause) *e* is inter-substitutable with an expression that is governed by (Sat-intro) and (Sat-elim), for some n,

or

(Inductive Clause) the result of replacing each non-primitive component expression in *e* with its definiens contains an alethic expression.

(Here intersubstitutability can be relativized to a context of utterance or to further philosophical assumptions, as in the case of L_1.) Given this notion, we can accommodate sentences such as L_1

[59] Here is why. Relative to that assumption, 'has the most philosophically interesting property' refers to truth. Thus, the sentence 'L_1 has the most philosophically interesting property' attributes truth to L_1. L_1 is thus the negation of a sentence whose grammatical subject refers to L_1 and whose grammatical predicate refers to truth. But that is the definition of a Liar sentence from the Introduction.

[60] Here is how the derivation would go:

1.	L_1 = '~(L_1 has the most philosophically interesting property)'	(definition of 'L_1')
2.	The most philosophically interesting property = truth	(assumption)
3.	|<u>L_1 has the most philosophically interesting property</u>	(assumption)
4.	|'~(L_1 has the most philosophically interesting property)' has...	(Substitution, (1), (3))
5.	|'~(L_1 has the most philosophically interesting property)' is true	(Substitution, (2), (4))
6.	|~(L_1 has the most philosophically interesting property)	((T-out), (5))
7.	|Contradiction	(contradiction intro., (3), (6))
8.	~(L_1 has the most philosophically interesting property)	(negation introduction, (3)-(7))
9.	'~(L_1 has the most philosophically interesting property)' is true	((T-in), (8))
10.	L_1 is true	(Substitution, (1), (9))
11.	L_1 has the most philosophically interesting property	(Substitution, (2), (10))
12.	Contradiction	(contradiction intro., (8), (11))

by specifying that to be Liar-like, a sentence must contain an occurrence of an alethic expression within the scope of negation.

Although it is good as far as it goes, this condition does not apply to Curry sentences. In these sentences, recall, 'true' occurs not in the scope of negation but rather in the antecedent of a conditional. With a little extra machinery, this complication can be accommodated. The notion of a relation symbol's occurring within the scope of negation has a well-known generalization that also includes the antecedent clauses of conditionals: they are both ways that relation symbols can *occur negatively* in a formula. To see this notion, let us first define the notion of the *deconditionalization* $d(\chi)$ of a formula χ. This is simply the result of rewriting all occurrences of every conditional in χ (material conditional or otherwise) in terms of negation and disjunction, as though they were all occurrences of the material conditional.[61] Thus, for example, if 'F' and 'G' are atomic one-place predicates and 'a' and 'b' are singular terms, then d('Fa → Gb') is '(~Fa) v Gb'. We can now define the notion of a negative occurrence of an expression in a formula. If *e* is an expression of any kind, χ is a formula, and E is an occurrence of *e* in χ, then E is a *negative occurrence* of *e* in χ if, when χ is transformed in $d(\chi)$, the occurrence of *e* that corresponds to E lies under the scope of an odd number of negation operators.[62] For example, the occurrence of 'F' in 'Fa → Gb' is negative, since once we transform 'Fa → Gb' into '(~Fa) v Gb', the occurrence of F lies under the scope of an odd number of negation operators.

[61] More precisely, for any formula χ,

$$d(\chi) = \begin{cases} \chi, & \text{if } \chi \text{ is an atomic predication,} \\ \ulcorner\sim d(\varphi)\urcorner, & \text{if } \chi \text{ is } \ulcorner\sim\varphi\urcorner \text{ for some formula } \varphi, \\ \ulcorner d(\varphi)\ \&\ d(\psi)\urcorner, & \text{if } \chi \text{ is } \ulcorner\varphi\ \&\ \psi\urcorner \text{ for some } \varphi \text{ and } \psi, \\ \ulcorner d(\varphi)\ \mathrm{v}\ d(\psi)\urcorner, & \text{if } \chi \text{ is } \ulcorner\varphi\ \mathrm{v}\ \psi\urcorner \text{ for some } \varphi \text{ and } \psi, \\ \ulcorner(\sim d(\varphi))\ \mathrm{v}\ d(\psi)\urcorner, & \text{if } \chi \text{ is } \ulcorner\varphi \rightarrow \psi\urcorner \text{ for some } \varphi \text{ and } \psi, \\ \ulcorner\exists z d(\varphi)\urcorner, & \text{if } \chi \text{ is } \ulcorner\exists z\varphi\urcorner \text{ for some } \varphi \text{ and some variable z,} \\ \ulcorner\forall z d(\varphi)\urcorner, & \text{if } \chi \text{ is } \ulcorner\forall z\varphi\urcorner \text{ for some } \varphi \text{ and some variable z.} \end{cases}$$

[62] Note: a relation symbol that occurs more than once in a formula can occur negatively and also occur non-negatively.

With the notion of negative occurrences in hand, we can propose that for a sentence S to be Liar-like, it must contain a negative occurrence of an alethic expression. This proposal allows us to accommodate Curry sentences. For example, *d*('If C is true then grass is purple') is '(~(C is true)) v grass is purple'. Clearly 'true' occurs under the scope of a single negation in the latter formula, so 'If C is true then grass is purple' contains a negative occurrence of 'true'. Thus, this Curry sentence meets the proposed necessary condition for being Liar-like, as pre-theoretically it should.

While the condition just proposed may be satisfactory as a necessary condition for being Liar-like, it is clearly not a sufficient condition, since plenty of unproblematic sentences satisfy it. For example, the sentence ''Grass is purple' is not true' contains an occurrence of 'true' under the scope of a single negation, but is not Liar-like. In response to this observation, a natural suggestion is that a sentence S is Liar-like if and only if it contains a negative occurrence of an alethic expression which is applied to a name for S. On this suggestion, a sentence must refer to itself in order to be Liar-like.

However, recall from previous sections that none of the Yablo, Cook, and Bad Pair paradoxes involves self-reference. Likewise, while Grelling's sentence U involves self-application of a predicate, it contains no term that refers to U itself. Finally, Cook's paradox involves no kind of referential circularity whatsoever.[63] Since we are counting all these paradoxes as Liar-like, we can conclude that a sentence or set of sentences need not involve self-reference, or any other kind of referential circularity, in order to give rise to a Liar-like paradox. That is an important lesson about what it takes to be a Liar-like sentence.

[63] For a general notion of referential circularity, see the notion of a *fixed point*, (Cook 2006).

Still, there is a general notion that includes both referential circularity and the patterns of reference involved in Grelling's, Yablo's, and Cook's paradoxes as special cases. To develop this notion, I must introduce the idea of *alethic reference*.[64] Very roughly, the idea is that a sentence S alethically refers to a formula φ if and only if there is an alethic expression *e* which occurs in S, and S refers to φ within the scope of that occurrence of *e*. Much less roughly:

(i) For any alethic expression *e*, natural number n, singular terms $a_1,\ldots,a_n$, and formula $\varphi(x_1,\ldots,x_n)$, any sentence that contains the formula $\ulcorner e(\langle a_1,\ldots,a_n\rangle, \varphi(x_1,\ldots,x_n))\urcorner$ *alethically refers* to the sentence $\ulcorner\varphi(a_1,\ldots,a_n)\urcorner$ *via* the occurrence of *e* in the subformula $\ulcorner e(\langle a_1,\ldots,a_n\rangle, \varphi(x_1,\ldots,x_n))\urcorner$.

(ii) Let *e* be an alethic expression, z be a variable that ranges over indices of formulas, *f* be a function that maps the index of each formula to the number of free variables in that formula, let a_i be a name and x_i be a variable for all natural numbers i, and let ψ be a formula that contains the formula $\ulcorner e(\langle a_1,\ldots,a_{f(z)}\rangle, \varphi_z(x_1,\ldots,x_{f(z)}))\urcorner$. Then for all values p of z, the formula $\ulcorner\Box z\ \psi\urcorner$ *alethically refers* to the sentence $\ulcorner\varphi_p(a_1,\ldots,a_{f(p)})\urcorner$ *via* the occurrence of *e* in the subformula $\ulcorner e(\langle a_1,\ldots,a_{f(z)}\rangle, \varphi_z(x_1,\ldots,x_{f(z)}))\urcorner$.

(iii) For any sentence S, natural number n, singular terms $a_1,\ldots,a_n$, and formula $\varphi(x_1,\ldots,x_n)$, S *alethically refers to* $\ulcorner\varphi(a_1,\ldots,a_n)\urcorner$ if and only if there exists an alethic expression *e* and an occurrence E of *e* in S such that S alethically refers to $\ulcorner\varphi(a_1,\ldots,a_n)\urcorner$ via E.

A sentence S *negatively alethically refers* to a formula $\varphi(x_1,\ldots,x_n)$ if S alethically refers to $\varphi(x_1,\ldots,x_n)$ via, and only via, a negative occurrence of an alethic expression.

To get a grip on this definition, let's look at some examples. Recall the Strong Liar sentence, 'L* is not true'. By clause (i) above, the sentence L* alethically refers to itself. Modulo some minor cosmetic differences, L* is '~satisfies(L*)', which is the result of substituting all zero of the names in the (unique) 0-tuple for all zero of the variables in L*. L* negatively alethically refers to itself, because the only occurrence of an alethic expression by which L*

[64] This notion of alethic reference is modeled after Lavinia Picollo's eponymous notion, defined in (Picollo 2017). Many thanks to Lavinia for her helpful suggestions.

alethically refers to itself is a negative occurrence (falling within the scope of a single negation operator).

In a slightly different way, it follows from clause (i) above that the Grelling sentence U—that is, the sentence '~satisfies ('~satisfies (x, x)', '~satisfies (x, x)')'—alethically refers to itself. U contains an occurrence of the alethic expression 'satisfies' applied to the pair <'~satisfies (x, x)', '~satisfies (x, x)'>. Therefore, U alethically refers to the sentence that results from substituting the name ''~satisfies (x, x)'' for the variable 'x' in the formula '~satisfies (x, x)'. But that sentence is just U itself. Therefore, U alethically refers to itself. U negatively alethically refers to itself, since the unique occurrence of 'satisfies' by which U alethically refers to itself occurs under the scope of a single negation operator.

It is also easy to check that by clause (i), each Cook sentence negatively alethically refers to all subsequent ones. Modulo cosmetic differences, given any natural number n, for each m > n the Cook sentence S_n contains the formula 'satisfies (S_m)' (establishing alethic reference) under the scope of a single negation operator (establishing negative alethic reference).

Lastly, clause (ii) above allows us to say that each Yablo sentence alethically refers to all subsequent Yablo sentences. Modulo cosmetic differences, for each natural number n the Yablo sentence S_n is '$\forall$y > n, ~satisfies (S_y)'. Here *f* is the constant function whose value is 0, since for each m, S_m is a sentence, containing no free variables. Since S_n contains the formula 'satisfies (S_y)' and the variable 'y' ranges over the indices of all subsequent Yablo sentences, S_n alethically refers to all subsequent Yablo sentences. This alethic reference is negative, since the formula 'satisfies (S_y)' occurs in the scope of a single negation operator.

From these examples, we can observe that all the Liar-like sentences we have seen involve negative alethic reference. To build toward a definition of 'Liar-like sentence', we now

need only a few further notions. Let an *alethic referential chain* be a countable sequence of (not necessarily distinct) sentences, each of which alethically refers to at least one of the subsequent sentences in the sequence. (So, every alethic referential chain is infinitely long.) An alethic referential chain C is a *bad referential chain* if every alethic referential subchain of C includes infinitely many instances of negative alethic reference.

To get the idea, let's look at a few examples. For starters, the sequence whose every element is L* is a bad infinite referential chain with L* at its head. Each sentence alethically refers to all subsequent ones, and all of these infinitely many instances of alethic reference are instances of negative alethic reference, so any alethic referential subchain contains infinitely many instances of negative alethic reference. For similar reasons, the sequence whose every element is the Grelling sentence U is a bad referential chain. As for the Yablo sentences, for each such sentence, the sequence that consists of all of the subsequent Yablo sentences, in order, is a bad referential chain. That is easy to see, given that each Yablo sentence negatively alethically refers to all subsequent ones. Similarly for Cook's sentences.

The case of the Bad Pair is a little more interesting. Sentence A—that is, 'B is not true'—is the first element of the bad referential chain

'B is not true', 'A is true', 'B is not true', 'A is true',…,

and B—that is, 'A is true'—is the first element of the bad referential chain

'A is true', 'B is not true', 'A is true', 'B is not true'…..

Let's take a moment to verify that both of these sequences are bad referential chains. Both are alethic referential chains, for both sequences are infinitely long, and in both sequences, each sentence refers to the next sentence in that sequence. Moreover, for both sequences, the only alethic referential subchains are ones that contain infinitely many Bs. (Suppose we are given a

subsequence C that contains only finitely many Bs. Then after some number n of places in the sequence, C contains only As. Since the sentence A refers only to B and not to itself, the copy of A in place n+1 fails to refer to any subsequent sentences in C. Therefore, C is not an alethic referential chain.) Thus, any alethic referential subchain contains infinitely many Bs. But B negatively alethically refers to A. So, any alethic referential subchain contains infinitely many instances of negative alethic reference.

With the notion of a bad referential chain in hand, we can now propose the following as a necessary and sufficient condition for a sentence's being Liar-like:

(ABS_1) A sentence S is Liar-like if and only if S is the first element of a bad referential chain[65]

Here 'ABS' stands for the abstraction strategy for identifying all and only the Liar-like sentences; we will soon consider some definitions that emerge from the definitions strategy. We will soon have reason to refine (ABS_1) even further, but for now let us set the abstraction strategy aside and examine the definition strategy.

5.1.2. The Definition Strategy

In this section, I will explore the *definition strategy* for identifying all and only the sentences one needs to target in order to solve the Liar-like paradoxes. The motivating idea here is that whatever it is that explains why the Liar-like paradoxes arise, the most immediate task for a diagnosis is to target all and only the sentences that give rise to such paradoxes. For that purpose, one can simply identify the Liar-like sentences to be all and only the sentences that give rise to Liar-like paradoxes. The task of identifying all such sentences then reduces to that of

[65] The 'via, and only via' clause prevents the sentence 'This sentence is either true or not true' from counting as negatively alethically referring to itself. Thus, the infinite sequence each of whose elements is that sentence fails to count as a bad referential chain, and so that sentence fails to be Liar-like, as pre-theoretically it should.

adequately defining 'Liar-like paradox' and 'gives rise to'. In what follows, I will develop the idea that something is a *Liar-like paradox* if it is a paradox that makes ineliminable use of (Sat-intro) and (Sat-elim), or, what comes to the same, (T-intro) and (T-elim).[66] And I will propose that a sentence S *gives rise to* a Liar-like paradox P if (a) S cannot be removed from P, and (b) P contains a subderivation either of S or of S's negation.

5.1.2.1. Liar-like Paradoxes

Say that a rule of inference is *eminently plausible* if it so strongly strikes us as acceptable as to be virtually immune to rejection. Similarly, a sentence is *eminently plausible* if what it says is so plausible as to be virtually undeniable.[67] Likewise, a sentence is *eminently implausible* if what it says is so unacceptable as to be virtually immune to acceptance. With these definitions in place, I will move toward a definition of 'paradox'. To start with, I will define a closely related notion, that of a *formally-presented paradox*. As a first pass, say that something is a *formally-presented paradox* if it is a formal derivation which meets the following conditions:

1. Each sentence in the sequence follows from the others by rules of inference that are:
    i. eminently plausible, but for the existence of the derivation, and
    ii. licensed by the syntax of the sentences to which they are applied.
2. The premises are eminently plausible, but for the existence of the derivation.
3. The conclusion is eminently implausible.[68]

However, this definition of 'formally-presented paradox' will not do. Consider the following derivation:

[66] Recall the discussion of Grelling's Paradox, Section 5.6.
[67] Here I ignore issues of context-sensitivity.
[68] In personal communication, Volker Halbach raised the concern that this definition is vague, since it is a vague matter what sentences and rules of inference are eminently plausible. In fact, though, the vagueness in the definition should be embraced as a virtue, since the phenomenon the definition is attempting to capture—namely, the property *being a paradox*, as pre-theoretically understood—is itself vague in the same way.

1. C = ‘If C is true then there are infinitely many primes’ (definition of ‘C’)
2. |C is true (assumption)
3. |‘If C is true then there are infinitely many primes’ is true (Substitution, (1), (2))
4. |If C is true then there are infinitely many primes ((T-out), (3))
5. | there are infinitely many primes (cond. elim., (2), (4))
6. If C is true then there are infinitely many primes (cond. intro., (2)-(5))
7. ‘If C is true then there are infinitely many primes’ is true ((T-in), (6))
8. C is true (Substitution, (1), (7))
9. There are infinitely many primes (cond. elim., (6), (8))

Let’s call the above derivation ‘D’. D’s conclusion is not eminently implausible. In fact, it is eminently plausible! Still, however, D is highly problematic; the Curry-like reasoning that D exemplifies is not an acceptable way to reach the conclusion that there are infinitely many primes. One reasonable way to explain what is wrong with D is to emphasize that reasoning of this same kind could be used to derive any declarative sentence one likes. For example, replacing the clause ‘there are infinitely many primes’ with ‘grass is purple’ throughout yields a derivation that still meets conditions (1) and (2), and whose conclusion is absurd.[69]

In light of cases like D, then, I propose to define ‘formally-presented paradox’ as follows. Something is a *formally-presented paradox* if it is a formal derivation which meets the following conditions:

1. Each sentence in the sequence follows from the others by rules of inference that are:
    iii. eminently plausible, but for the existence of the derivation, and
    iv. licensed by the syntax of the sentences to which they are applied.
2. The premises are eminently plausible, but for the existence of the derivation.
3. If one replaced the conclusion with one that is eminently implausible, making at most minor adjustments throughout, then (1) and (2) would still hold.

[69] Here is how that would go:

1. C = ‘If C is true then grass is purple’
2. |C is true (assumption)
3. |‘If C is true then grass is purple’ is true (Substitution, (1), (2))
4. |If C is true then grass is purple ((T-out), (3))
5. |Grass is purple (conditional elimination, (2), (4))
6. If C is true then grass is purple (conditional introduction, (2)-(5))
7. ‘If C is true then grass is purple’ is true ((T-in), (6))
8. C is true (Substitution, (1), (7))
9. Grass is purple (conditional elimination, (6), (8))

This definition takes account of derivations like D, enabling them to count as paradoxical in virtue of clause (3).

One last remark before moving on. Clauses (1)-(3) serve only to define 'formally-presented paradox', not the more general term 'paradox'. That is because not every bit of reasoning that we would want to call a paradox is a formal derivation. For my purposes, the most relevant obstacle is that of missing steps. For example, assume that truth is the most philosophically interesting property. Then the sentence 'This sentence lacks the most philosophically interesting property' generates a version of the Liar paradox. But to get from that sentence to a contradiction in a formal derivation, one needs the premise 'Truth is the most philosophically interesting property'. That is because the predicate 'lacks the most philosophically interesting property' is not ordinarily governed by (T-intro) and (T-elim); one needs to substitute in 'is not true' in order to apply these rules.

With these reflections in mind, let us say that something P is a *paradox* if P is a sequence of sentences, and either:

A) P is a formally-presented paradox,

or

B) Inserting some extra, eminently plausible steps into P results in a formally-presented paradox that uses all of the sentences in P to arrive at its conclusion.[70]

For any sequence of sentences P, if P is a paradox that is not formally-presented, then anything that is a result of applying the operation described in (B) to P is *a formalization of* P. If P is a formally-presented paradox, then P is its own (unique) formalization.

[70] The clause about using all of the sentences in P prevents a great many pre-theoretically unparadoxical sequences of sentences from counting as paradoxes. For example, take the one-line formal derivation of 'a = a' and append to it a formally presented paradox. This is now a formally-presented paradox. Without the clause requiring that all the lines be used in deriving a contradiction, the original derivation (that is, the single line consisting of the sentence 'a = a') would count as a paradox.

With the notion of a paradox in place, I will now begin explaining the notion of a paradox's *making ineliminable use of* a rule of inference. Let us say that a derivation D *involves an ineliminable use* of an inference rule R if D involves an instance of R, and in addition there is no sequence of transformations of D that preserve syntactic consequence and whose end result is a derivation with the same conclusion as D that contains no instances of R. (A transformation T *preserves syntactic consequence* if and only if for every derivation E, if the conclusion of E is a syntactic consequence of the premises of E then the conclusion of T(E) is a syntactic consequence of the premises of T(E).)

Now for the rules of inference that make a paradox Liar-like. It is tempting simply to identify these with (T-intro) and (T-elim), since these rules have featured in most of the paradoxes we have seen so far. However, we saw in connection with Grelling's Paradox that (T-intro) and (T-elim) can be viewed as merely special cases of (Sat-intro) and (Sat-elim), modulo some trivial cosmetic differences. So, to be satisfyingly general, our definition of 'Liar-like' should focus on (Sat-intro) and (Sat-elim).[71]

Here, then, is my definition of 'Liar-like paradox':[72]

a) A formally-presented paradox P is *Liar-like* if for some $n \geq 0$, P makes ineliminable use of (Sat-intro) and (Sat-elim), perhaps modulo some trivial cosmetic differences.

b) A paradox is *Liar-like* if all of its formalizations are Liar-like.

[71] See (Scharp 2013) chapter 2 and Section 3.6 for justification for a focus on (T-intro) and (T-elim). Given the connections between these rules and the rules (Sat-intro) and (Sat-elim), the same considerations justify a focus on the latter.

[72] This definition rules out several paradoxes that some philosophers might feel to be pre-theoretically Liar-ish, such as Berry's paradox, Richard's paradox, and the paradox of the name. However, I myself simply do not find these paradoxes to be Liar-ish, so I am unmoved by their exclusion. Note, though, that the definition does apply to paradoxes that turn on the notion of definability, given that definability is understood in terms of satisfaction: to be *definable* by a predicate E is to be the unique satisfier of E.

This definition is fairly easy to support. As the presentations given in earlier sections suggest, it is highly plausible that any formalization of the Strong Liar, Yablo, Cook, Bad Pair, Curry, and Grelling paradoxes will involve ineliminable applications of (Sat-intro) and (Sat-elim), or of (T-intro) and (T-elim). Thus, these paradoxes all count as Liar-like according to my definition. That agrees with most philosophers' pre-theoretical reactions to these paradoxes.

What is less immediately obvious is that even in cases in which the offending sentence does not make explicit reference to truth or satisfaction, once the paradoxical reasoning is formalized it emerges that (Sat-intro) and (Sat-elim) are playing an ineliminable role in arriving at a contradiction. To see the point, recall the Classic Liar Sentence, L—that is, 'L is false'. One can easily check that both (T-intro) and (T-elim) are used in the associated paradox. Enough failures to represent the reasoning in a formal way without them have convinced me that they are ineliminable.

In my discussion of the No-No "paradox", I left open whether or not aberrationism applies to it. That is because I take it to be an open question whether or not the No-No is indeed a paradox. As I have defined 'paradox', the matter hinges on whether or not the presupposition that the No-No sentences must receive the same status is eminently plausible, but for the fact that using standard rules of inference one can derive a contradiction from this presupposition. If the presupposition is indeed eminently plausible, then we have a genuine paradox. Indeed, in that case, we have a Liar-like paradox, since it is highly plausible that (T-intro) and (T-elim) play ineliminable roles in the derivation. In that case, the No-No sentences are Liar-like, and so one-off aberrations diagnoses apply to them. According to these diagnoses, a one-off aberration occurs in connection with these sentences. In Section 3.4, I laid out the best available ways of developing this claim: one can, as I prefer, hold that the occurrences of 'true' in both sentences

witness aberrations, or one can claim that there is a single aberration and it is indeterminate which occurrence witnesses it.

For my own part, I am strongly inclined to believe that the No-No paradox is indeed a genuine paradox, and to say that both of the sentences involved witness aberrations. However, other philosophers, such as Nicholas Smith, might be more willing to take the lesson to be that some deeply similar sentences have different semantic statuses. In that case, the No-No sentences do not present us with a genuine paradox, and so there is no need to wheel in any kind of solution, let alone aberrationism. Moderate aberrationists such as myself will find this situation hard to accept, since we are committed to positing strong, explanatory links between a sentence's semantic status and the kinds of non-semantic properties that the No-No sentences obviously have in common. However, only moderate aberrationists are committed to Semantic Regularity; radical aberrationists will find an asymmetric diagnosis perfectly acceptable. The fact that each account of the No-No "paradox" is consistent with some version of aberrationism demonstrates that the No-No poses no threat to the general idea of aberrationism.

5.1.2.2. Liar-like Sentences

Again, the motivating idea in this subsection is that a sentence is Liar-like if and only if it gives rise to a Liar-like paradox. Now that the notion of a Liar-like paradox is in place, the only remaining work is to define 'gives rise to' in a precise, satisfying way. To motivate the definition on which I will settle, it will be helpful first to consider two simpler attempts.

An initial idea is to say that a sentence S gives rise to a paradox P if S occurs in every formalization of P. But many sentences satisfy this condition without being pre-theoretically Liar-ish. For example, if P is a formally-presented paradox that contains the sentence '1 = 1' as

an extraneous step, then because P is its own unique formalization, every sentence in P, including '1 = 1', will occur in every formalization of P. That would render the sentence '1 = 1' Liar-like. And it would do likewise for a great many other entirely unproblematic sentences which serve as extraneous steps in other formally-presented paradoxes. Clearly this is unacceptable. We need to pin down the informal sense in which a sentence can be *responsible* for a paradox, as opposed to merely occurring in it.

Thinking along these lines, one might next propose that a sentence S is Liar-like if there is a Liar-like paradox from which S cannot be removed. More precisely, if S is a sentence and D is a formal derivation, say that S *cannot be removed from* D if there is no transformation T of D such that

1. T consists only of removing steps from D, other than the conclusion, and
2. T preserves syntactic consequence, and
3. T removes S everywhere from D except possibly the conclusion.

Given this definition of 'cannot be removed', we can propose that

(DEF_1) For all sentences S, S is Liar-like if there is a Liar-like paradox P such that P contains S, and for every formalization D of P, S cannot be removed from D.

Even (DEF_1) however, fails to state a sufficient condition for being pre-theoretically Liar-ish, as many pre-theoretically unparadoxical sentences fit the above description. For example, the sentence 'Grass is purple' bears this relation to the Curry-type paradox exhibited below:

1. C = 'If C is true then grass is purple'	
2. |<u>C is true</u>	(assumption)
3. |'If C is true then grass is purple' is true	(Substitution, (1), (2))
4. |If C is true then grass is purple	((T-out), (3))
5. |Grass is purple	(conditional elimination, (2), (4))
6. If C is true then grass is purple	(conditional introduction, (2)-(5))
7. 'If C is true then grass is purple' is true	((T-in), (6))
8. C is true	(Substitution, (1), (7))
9. Grass is purple	(conditional elimination, (6), (8))

The sentence ‘Grass is purple’ cannot be removed from the derivation, because of the important role it plays in the derivation of step (6). Yet although this sentence is false, it is not pre-theoretically Liar-ish. Moreover, and perhaps more controversially, I claim that the same holds of the sentence ‘C = ‘If C is true then grass is purple’’, written on line (1). That is, although this sentence is ineliminable from the derivation, it is not Liar-ish. The paradox arises, I urge, not from this sentence, which is used to define the name ‘C’, but rather from C itself.[73] Again, the point is, not every sentence that occurs in a Liar-like paradox is responsible for that paradox in the way that pre-theoretically Liar-ish sentences appear to be.

What, then, distinguishes pre-theoretically Liar-ish sentences from others that are bound to occur in their associated paradoxes? Examining all the paradoxes we have seen, we find something striking: in each case, the sentence that we pre-theoretically take to be most responsible for the paradox is entered into the derivation via a sub-derivation. For instance, in the case of the Strong Liar, from the undischarged assumption of ‘L* is true’ we derive a contradiction. Discharging the assumption, we then use this sub-derivation to enter L*—that is, ‘L* is not true’. Similarly, in the Curry paradox, from the undischarged assumption of ‘If C is true then grass is purple’ we derive ‘Grass is purple’. Discharging the assumption, we then use this sub-derivation to enter C—that is, ‘If C is true then grass is purple’.

However, as it happens there are some Liar-like paradoxes that contain no sub-derivation of their offending sentences. Consider, for example, the following:

[73] Given the existence of the lexical constituents of C and of the ways of combining them afforded by English grammar, the existence of C is incontrovertible. And since this sentence exists and we have the letter ‘C’ available as a name, we can simply stipulate that ‘C’ will henceforth name the sentence ‘If C is true then grass is purple’. Even aberrationists, who claim that stipulations as to the reference of terms cannot always succeed (recall Section 3.3), can perfectly well allow that the stipulation that defines ‘C’ succeeds. And in the interest of minimizing the ways that our linguistic stipulations can go awry, they should allow this.

1.	L* = ‘L* is not true’	definition of ‘L*’
2.	|L* is not true	(assume for *reductio*)
3.	|‘L* is not true’ is true	(2), (T-intro)
4.	|L* is true	(1), (3), substitution
5.	|contradiction	(2), (4), contradiction introduction
6.	Not (L* is not true)	(2)-(5), negation introduction
7.	L* is true	(6), double negation elimination
8.	‘L* is not true’ is true	(1), (7), substitution
9.	L* is not true	(8), (T-elim)
10.	Contradiction	(7), (9), contradiction introduction

This formally-presented paradox does not involve a subderivation whose conclusion is L*. Still, it accomplishes this only by including a subderivation whose conclusion is L*’s negation.

In light of this example, then, I propose that a sentence S *gives rise* to a Liar-like paradox P if and only if for every formalization D of P, D includes either a subderivation whose conclusion is S or a subderivation whose conclusion is S’s negation, and S cannot be removed from D. Plugging this definition into the definition of ‘Liar-like sentence’, we get the following:

(DEF_2) For all sentences S, S is *Liar-like*$_{DEF}$ if and only if there is a Liar-like paradox P such that

for every formalization D of P,

a. D contains, as a sub-derivation, a derivation either of S or of S’s negation,[74]

and

b. S cannot be removed from D.

Although (DEF_2) may stand in need of yet further refinements, I will leave my exploration of the definitions strategy here.

5.1.3. Accommodating the Maximal Expressiveness of Natural Languages

Some of aberrationism’s motivations impose significant constraints on which sentences its advocates can take to be Liar-like. In this subsection, I will explore whether (ABS_1) and

[74] Given that Curry’s paradox does not involve negation, one might worry that there is some Curry-like paradox that involves a Curry sentence but no subderivation of either it or its negation. However, the most straightforward attempt to achieve this ends up including a subderivation of the sentence after all:

(DEF_2) can meet these constraints. To see the constraints in question, it will help to introduce the notion of a *Russellian proposition*. For our purposes, it is safe to think of Russellian propositions as ordered pairs of objects and properties. Whenever an object *o* and a property P exist, there exists the Russellian proposition <*o*, P>, which attributes the property P to the object *o*.

One might reasonably want to hold that for every Russellian proposition, there is some sentence of English which expresses that proposition. Likewise, for every Russellian proposition, there is some sentence of English which negates that proposition. At least, these claims are part of the pre-theoretically plausible notion that natural languages are maximally expressive (recall Section 3.5)—in particular, that anything which can be said at all can be said in a single natural language, perhaps supplemented with some further vocabulary. As I explained in Section 3.3, the idea that natural languages are maximally expressive constitutes part of the motivation for aberrationism.[75]

To illustrate the problem that I want to discuss, I will consider two examples. Look first at the Strong Liar sentence, L*. Aberrationists claim that what L* fails to say about itself can be said by some other sentence. In particular, in Section 3.3, I claimed that L' will do the job:

(L') 'L* is not true' is not true

---

1. |<u>If C is true then (if C is true then grass is purple)</u>
2. ||<u>C is true</u>
3. ||If C is true then grass is purple
4. ||Grass is purple
5. |If C is true then grass is purple
6. If C is true then (If C is true then grass is purple)
7. |<u>C is true</u>
8. |If C is true then grass is purple
9. |Grass is purple
10. If C is true then grass is purple
11. 'If C is true then grass is purple' is true
12. C is true
13. Grass is purple

So, Curry's paradox does not furnish us with a counterexample to the proposed definition.

[75] See also the discussion of incompleteness in Chapter 1, Section 5.

However, this suggestion raises the immediate question of whether L' is Liar-like in turn. If it is, then by aberrationists' own lights, its unquoted occurrence of 'true' witnesses a referential aberration. In that case, L' fails to negate the Russellian proposition <L*, truth>. But that is a problem, given that L' was introduced precisely in order to negate that proposition!

Similar remarks hold for the sentence L_1 from above:

(L_1) ~(L_1 has the most philosophically interesting property)

Relative to the assumption that truth is the most philosophically interesting property, L_1 is a Liar sentence. Yet one might hope that the Russellian proposition <L_1, truth> can be negated in some other way. A straightforward suggestion is L'' below:

(L'') L_1 is not true

As we had with L' above, this suggestion immediately raises the question whether L'' is Liar-like. If it is, then L'' fails to negate the Russellian proposition <L_1, truth>. But that was why L'' was introduced in the first place!

Now for a response. If one is positing aberrations only in Liar sentences, then it is easy to avoid positing them in sentences like L' and L''. Start with L'. L' is not a Liar sentence; at least, not by the definition I gave in Section 1.1. (In short: a Liar sentence talks about itself, whereas L' talks about a different sentence.) Admittedly, despite its not being a Liar sentence, one can derive a contradiction from L', given the indisputable premise 'L* = 'L* is not true'' and the rule of substitution. Here is how that would go:

1. L* = ‘L* is not true’ (definition of ‘L*’)
2. |‘L* is not true’ is true (Assume for *reductio*)
3. |L* is true (Substitution, (1), (2),)
4. |L* is not true ((2), (T-elim))
5. |Contradiction (Contradiction Introduction, (2), (4))
6. ‘L* is not true’ is not true (Negation Introduction, (1)-(4))
7. L* is not true (Substitution, (1), (6))
8. ‘L* is not true’ is true ((T-intro), (7))
9. Contradiction (Contradiction Introduction, (6), (8))

Still, aberrationists can simply reject the instance of substitution that gets us from L’ in step (6) to L* in step (7). For if it is just Liar sentences that undergo aberrations, then L* but not L’ witnesses an aberration. In that case, the two sentences say different things, and thus we have principled grounds for rejecting the instance of substitution that gets us from L’ to L*. All the same points can be made regarding L_1 and L’’: as long as we are only positing aberrations in Liar sentences, we have no problem finding a way to say of L_1 that it is not true. L’’ is not a Liar sentence, for precisely the same reason that L’ is not; and so long as we are only targeting Liar sentences, we can exploit this fact when positing referential aberrations.

However, aberrationism aims to diagnose all Liar-like sentences, not just Liar sentences, and the problem is harder in that more general setting. Since sentences can be Liar-like in many different ways, the mere fact that a sentence fails to be a Liar sentence does not automatically guarantee it against witnessing an aberration. Aberrationists need a definition of ‘Liar-like’ that is inclusive enough to apply to all the sentences that give rise to Liar-like paradoxes, but exclusive enough to leave room for the possibility of expressing, negating, etc., the Russellian propositions that these sentences fail to express, negate, etc. In particular, they need a definition that applies to L* and L_1, but not also to the sentences that they take to negate the propositions <L*, truth> and <L_1, truth>.

Whatever its other virtues may be, the definition (ABS_1) will not fit this bill. It is easy to see that both L' and L'' lie at the heads of bad referential chains. L' lies at the head of the chain

''L* is not true' is not true', 'L* is not true', 'L* is not true', …,

and similarly, L'' lies at the head of the chain consisting of itself, followed by infinitely many copies of L_1. An initial suggestion is that

(ABS_2) A sentence S is *Liar-like*$_{ABS}$ if and only if

(a) it lies at the head of a bad referential chain

and

(b) not all of the singular terms in the scope of S's negative occurrence of its alethic expression are quote names of sentences

Clause (b) of ABS_2 exonerates L', rendering it non-Liar-like. And while (b) does not exonerate L'' from counting as Liar-like—this owing to the fact that L'' contains the name 'L_1', which is not a quote-name—clause (b) does exonerate the sentence L''' below:

(L''') '~(L_1 has the most philosophically interesting property)' is not true

So, we have a way to negate the proposition <L_1, truth>, as we need.

Grelling's sentence U is a trickier case. As it pre-theoretically should, ABS_2 counts U as Liar-like: although all the singular terms in the scope of U's negative occurrence of 'satisfies' are quote names, they are quote names of formulas, not sentences. Still, a challenge remains to find a sentence which negates the proposition <'~satisfies (x, x)', self-satisfaction> without itself counting as Liar-like. While I am hopeful that there is a way for aberrationists to improve on (ABS_2), I will leave that investigation here, turning now to the case of (DEF_2).

Recall that to be Liar-like$_{DEF}$, a sentence S must give rise to a Liar-like paradox P. For that, S must be non-removable from every formalization D of P. Now consider again the paradox associated with L':

1. L* = ‘L* is not true’ (definition of ‘L*’)
2. |‘L* is not true’ is true (Assume for *reductio*)
3. |L* is true (Substitution, (1), (2))
4. |L* is not true ((T-elim), (2))
5. |Contradiction (Contradiction introduction, (2), (4))
6. ‘L* is not true’ is not true (Negation Introduction, (1)-(4))
7. L* is not true (Substitution, (1), (6))
8. ‘L* is not true’ is true ((T-intro), (7))
9. Contradiction (Contradiction introduction, (6), (8))

The only role of L’ in the derivation is to introduce the sentences ‘L* is true’ (in step (3)) and L* (in step (7)), via the principle of substitution. This derivation can be easily transformed into the Strong Liar Paradox by simply removing L’ and making a few other minor changes. By contrast, L* clearly cannot be removed, either from the derivation above or from the Strong Liar Paradox. So, L’ does not count as Liar-like, whereas L* does. I’ll spare the reader the details, but it is easy to check that similar remarks apply to L’’; the only role of L’’ in getting a contradiction is to reintroduce L_1, and so we can safely count L’’ as not being Liar-like.[76]

5.1.4. *Being Liar-Like* and Revenge

Now that I have moved toward making aberrationism more precise, the question arises once again whether it suffers from revenge. In particular, one might wonder whether either of the definitions of ‘Liar-like sentence’ that I explore gives rise to a revenge problem. The standard way to create such a problem would be with the sentences LL_{ABS} and LL_{DEF} below, considering each under the assumption that Liar-like-ness is identical with the property that that sentence appears to attribute to itself:

[76] Still, though, one might worry about Yablo’s and Cook’s sentences. For each of these sentences, one can obtain an almost identical version of the paradox just by considering all the subsequent ones in the list. Does it follow that each sentence is eliminable? If so, then (DEF_2) needs to be revised. However, the fact that all the sentences are nearly identical, plus the fact that the paradox cannot be run without using at least some or other of these sentences, provide an undeniable sense that the sentences are close to being non-removable. E.g., it is impossible to remove all sentences which are of the form ˹For all m > n, S_m is not true˺ for some natural number n. Refining (DEF_2) so as to incorporate this observation is a project for future research.

(LL_{ABS}) LL_{ABS} is Liar-like$_{ABS}$

(LL_{DEF}) LL_{DEF} is Liar-like$_{DEF}$

For both these sentences, however, there is a quick answer to the revenge worry. If the sentence is indeed Liar-like, according to its own definition of 'Liar-like', then according to aberrationism, the key occurrence of its alethic expression undergoes an aberration. Thus, the sentence fails to say what it appears to say, and so paradox is blocked.

A complication here is that neither 'Liar-like$_{ABS}$' nor 'Liar-like$_{DEF}$' is intersubstitutable with 'satisfies', either in general or for any particular natural number n. So, neither LL_{ABS} nor LL_{DEF} contains any alethic expressions; and so, strictly speaking, aberrationism does not allow either of these sentences to witness an aberration. However, the obvious response to this problem is to think of the term 'Liar-like$_{ABS}$' not as a primitive predicate but rather as an abbreviation of the phrases to define it; that is, to imagine that instead of 'Liar-like$_{ABS}$', we simply had the definition written out in full. Then we consider whether the resulting sentence would witness an aberration.

Even with the quick argument from above in place, it is worth considering whether LL_{ABS} and LL_{DEF} even count as Liar-like according the definitions of 'Liar-like' that they presuppose. If each turns out to be simply false, then there was no cause for concern to begin with. Start with LL_{ABS}. Again, a sentence S is Liar-like$_{ABS}$ if and only if S lies at the head of a bad referential chain and not all of the singular terms in the scope of S's negative occurrence of its alethic expression are quote names of sentences. Clearly LL_{ABS} contains no quote names, so the important question here is whether LL_{ABS} lies at the head of a bad referential chain. Since LL_{ABS} refers to itself, if there is a bad referential chain in the vicinity, then it is the infinite sequence each of whose elements is LL_{ABS} itself. To decide whether this is a bad referential chain, we have

to see whether every alethic referential subchain contains infinitely many instances of negative alethic reference. Since all subchains will be identical with the original chain, the question is simply whether LL_{ABS} negatively alethically refers to itself; that is, whether it alethically refers to itself via, and only via, a negative occurrence of an alethic expression. Since we are pretending that LL_{ABS} contains the exhaustively-spelled-out definition of 'Liar-like$_{ABS}$', the thing to do is look through this spelled-out definition and see whether we find the name 'LL_{ABS}' within the scope of a negative occurrence of an alethic expression. On reflection, the definition does contain an alethic expression. The definition of 'Liar-like$_{ABS}$' includes the term 'bad referential chain', whose definition in turn includes the term 'alethic expression', whose definition in turn includes the term 'intersubstitutable'; and intersubstitutability here is intersubstitutability *salva veritate*, which is defined in terms of truth. (Recall: for all expressions x and y, x is *intersubstitutable* with y if and only if: for any sentence S_1 containing x, if S_1 is true then the result of substituting y for x in S_1 is true; and for any sentence S_2 containing y, if S_1 is true then the result of substituting x for y in S_2 is true.) The point is, given the involvement of intersubstitutability, we may treat LL_{ABS} as "containing" an alethic expression. Still, for LL_{ABS} to negatively alethically refer to itself, the name 'LL_{ABS}' would have to occur in the scope of at least one negative occurrence of this expression. That does not hold.

To see why, plug the definition of 'intersubstitutable' into the definition of 'alethic expression', plug the result into the definition of 'bad referential chain', and then restate LL_{ABS} using the result. Omitting some irrelevant clauses, that gives us roughly the following:

(LL_{ABS}*) LL_{ABS}* is the first in an infinite sequence of sentences, each of which contains a name for some subsequent sentence in the sequence, within the scope of a negative occurrence of an expression *e* which is such that:

a) there is an expression x that is governed by (Sat-intro) and (Sat-elim), and

b) for any sentence S_1 containing x, if S_1 is true then the result of substituting *e* for x in S_1 is true; and for any sentence S_2 containing *e*, if S_1 is true then the result of substituting x for *e* in S_2 is true

The occurrence of 'LL_{ABS}' in LL_{ABS}* does not fall within the scope of any occurrences of 'true', let alone the negative ones. Therefore, LL_{ABS} is not Liar-like$_{ABS}$. So, it is not paradoxical; it is simply false.

Now consider LL_{DEF}. Let us begin by asking whether the predicate 'is not LL_{DEF}' is an alethic expression. Recall the definition: for a sentence to be Liar-like$_{DEF}$ is for it to give rise to a Liar-like paradox. Being a Liar-like paradox involves making ineliminable uses of (Sat-intro) and (Sat-elim). Admittedly, a derivation that makes ineliminable use of (Sat-intro) and (Sat-elim) must itself contain some alethic expressions. But the description of these rules does not itself employ any alethic expressions; rather, the description merely mentions the word 'satisfies' without using it or any other, intersubstitutable term. Moreover, no trouble arises from the notion of an inference rule's being ineliminable from a derivation. The term 'ineliminable' is defined in terms of syntactic consequence. (Recall: a derivation D makes an ineliminable use of an inference rule R if there is no sequence of transformations of D that preserve syntactic consequence and result in a derivation with D's conclusion and that does not involve R.) But syntactic consequence is not defined in alethic terms. It is defined by simply listing the rules of inference that are standardly regarded as acceptable. For a conclusion to be a syntactic consequence of a set of premises just is for there to be a sequence of inferential steps from the premises to the conclusion in which every step is an instance of one of the rules listed in the definition. Again, the point of these remarks is that 'Liar-like paradox' is not an alethic expression. Similar remarks apply to 'giving rise to', but I will spare the reader the details. Because 'gives rise to a Liar-like paradox' is not an alethic expression, there is no way that

LL_{DEF} can give rise to a derivation that makes ineliminable uses of (T-intro) and (T-elim). So, LL_{DEF} is simply false; there is no revenge problem here.

5.2. Monster-Barring and the Classification Problem

5.2.1. Introducing Monster-Barring

In (Scharp 2013), Kevin Scharp criticizes a kind of diagnosis that he calls "monster-barring." Approaches of the kind Scharp has in mind target paradoxical sentences simply on the grounds that they are paradoxical. It should be fairly clear that if it uses either of my proposed criteria for applicability—that is, either definition of 'Liar-like sentence'—aberrationism is bound to do this. For aberrationism singles out certain sentences as being subject to a distinctive kind of semantic aberration, simply on the grounds that these sentences give rise to Liar-like paradoxes. Note that this criticism applies even when the criterion used is *being Liar-like*$_{DEF}$, since ultimately the motivation for positing aberrations in sentences that are Liar-like$_{DEF}$ is the fact that they give rise to paradoxes. So, although Scharp himself does not discuss aberrationism, his criticisms of monster-barring can be applied to it. In this subsection, then, I will present Scharp's criticisms and explain how aberrationists can elude them.

As Scharp defines them, *monster-barring* approaches hold that

> it is illegitimate to use the sentences in the reasoning involved in aletheic paradoxes. These approaches seek to find something wrong with the paradoxical sentences themselves. There are several versions of this approach: *syntactic* (paradoxical sentences are not syntactically well-formed), *semantic* (paradoxical sentences are meaningless), *pragmatic* (paradoxical sentences cannot be asserted), and *inferential* (paradoxical sentences cannot be supposed in hypothetical reasoning). (p.21)

On the best interpretation of Scharp's remarks here,[77] monster-barring approaches are views according to which the reasoning goes wrong because of "something wrong with the paradoxical sentences themselves." On this interpretation aberrationism counts as monster-barring. Part of what these approaches claim is that due to the facts about what Liar-like sentences mean, the seemingly-contradictory steps in the inferences involving them do not genuinely contradict. For example, recall the reasoning associated with L*:

1. L* = 'L* is not true'	(definition of 'L*')
2. | L* is true	(Assume for *reductio*)
3. | 'L* is not true' is true	(substitution, (1), (2))
4. | L* is not true	((3), (T-elim))
5. | Contradiction	(Contradiction introduction, (2), (4))
6. L* is not true	(Negation Introduction, (1)-(4))
7. 'L* is not true' is true	((T-intro), (6))
8. L* is true	(substitution, (1), (7))
9. Contradiction	(Contradiction introduction, (6), (8))

According to aberrationism, L* fails to say of itself that it is not true, whereas the sentence in steps (2) and (8) does say of L* that it is true. Thus steps (2) and (4) do not genuinely contradict, and neither do steps (6) and (8); and so the applications of the rule of Contradiction Introduction in steps 5 and 9 are unjustified. That amounts to saying that because of what L* means, "it is illegitimate to use" L* in "the reasoning involved" in the Strengthened Liar paradox.

Scharp's criticisms of monster-barring approaches focus on variants of these approaches that take paradoxical sentences to be in some sense meaningless. In particular, he addresses the views of (Goldstein 2009) and (Armour-Garb and Woodbridge 2013). Goldstein argues that Liar

[77] For any Liar-like paradox, the reasoning associated with that paradox is not truth-preserving—that is, unless absolutely every declarative sentence is true. So, anyone who wants to resist the claim that absolutely every declarative sentence is true has to hold that "it is illegitimate to use the sentences in the reasoning involved in aletheic paradoxes." Illegitimate, that is, at least in the sense that such uses of those sentences do not constitute reasoning that is truth-preserving. However, by that criterion, every solution to the Liar paradox that involves retaining the concept of truth counts as a monster-barring approach. Scharp clearly takes the monster-barring approaches to be some narrower, more specific class of solutions, and the quoted passage strongly suggests the definition that I adopt in the main text.

sentences and their ilk fail to express propositions. For their part, Armour-Garb and Woodbridge distinguish two types of meaning that declarative sentences can have—meaning$_1$ and meaning$_2$—and argue that paradoxical sentences lack meaning$_1$.[78] Scharp sums up his criticisms of these views in the following statement:

> a successful monster-barring strategy cannot be just a made-up condition on meaningfulness or expressing a proposition or making a statement (or whatever) that paradoxical sentences fail to meet. One has to motivate it—show that sentences that meet it are intuitively meaningful (or whatever) and sentences that do not meet it (other than the paradoxical ones) are intuitively meaningless. Moreover, simply saying that paradoxical sentences lead to paradoxes is not a legitimate justification—the fact that they have been thought to be paradoxical is actually evidence that they *are* grammatical, contentful, assertable, and supposable. Otherwise, no one would think they pose any kind of problem. (p.61, italics his, formatting mine)

In what follows, I will consider Scharp's remarks as they relate to aberrationism.

5.2.2. Meaninglessness

One of Scharp's categories of monster-barring approaches is the "semantic" approaches. One might naturally think that aberrationism falls under this heading, since the centerpiece of any one-off aberrations diagnosis is a semantic claim—namely, that the occurrence of the key alethic expression in any Liar-like sentence fails to co-refer with that term, and that this causes the sentence not to say what it appears to say. However, it is important to notice that advocates of these approaches need not hold that paradoxical sentences are meaningless. My own preferred view is that they are far from meaningless. I explained in Section 1.2 that according to the moderate one-off aberrations diagnosis, the reference of the occurrence of 'true' in any Liar-like sentence is significantly influenced by how we use the word 'true'. Although nothing we do short of changing how we use other words could force every occurrence of 'true' to refer to truth,

[78] For present purposes the definitions of meaning$_1$ and meaning$_2$ are irrelevant. For details see (Scharp 2013) p.59.

our use of ‘true’ in accordance with (T-intro) and (T-elim) guarantees that any occurrences which cannot refer to truth nonetheless come close to doing so. In particular, my own view is that these occurrences are indeterminate in reference as between ascending truth and descending truth. So, Liar-like sentences are far from meaningless; what they mean is strongly influenced by how we use their component words and the manner in which these are combined. And since the key occurrence(s) of its alethic expression(s) comes close to co-referring with that term, any Liar-like sentence will come close to saying what it appears to say. This picture contrasts sharply with views that simply insist, with no further explanation, that Liar-like sentences are meaningless, or that they fail to express any propositions.

Still, Scharp might respond by insisting that it is just unbelievable that Liar sentences fail to say what they seem to say. Of course, barring some other solution, it would follow that the reasoning associated with any Liar-like paradox really is truth-preserving, and therefore absolutely every sentence is true. But Scharp can take this line, since he is ultimately out to show that (at least for purposes of thinking rigorously about language and thought) we must get rid of the concept of truth. So much the better for his view, then, if it turns out that truth is a trivial property, in the sense that absolutely every declarative sentence has it.

However, it is unacceptable to hold that absolutely every declarative sentence is true. For one thing, making this claim involves making an unacceptable use of the term ‘true’—a use that licenses its application to every declarative sentence whatsoever. Implicit in our use of the word ‘true’ are norms that restrict its application in significant ways. Speakers try to conform to these norms even once they have been made fully aware of the Liar paradox. To insist that the Liar paradox licenses the profligate application of ‘true’ is to insist on a change in use that the rest of

the linguistic community has resisted, even in the face of the tremendous pressure exerted by the paradox.

Still, since Scharp holds that the norms for the use of 'true' are incoherent,[79] he would deny that we have any good reasons to abide by them. But this diagnosis of our norms is radically uncharitable, as it ignores the linguistic community's struggle to resist incoherence. Before jumping ship, it is worth exploring the prospects of a response to the Liar paradox that involves a more charitable interpretation of the behavior of English speakers. Such a response is precisely what aberrationism promises to provide.

5.2.3. Targeting Paradoxical Sentences Because They Are Paradoxical

A familiar but more serious challenge emanates from some of Scharp's other remarks in the passage quoted above:

> simply saying that paradoxical sentences lead to paradoxes is not a legitimate justification—the fact that they have been thought to be paradoxical is actually evidence that they *are* grammatical, contentful, assertable, and supposable. Otherwise, no one would think they pose any kind of problem. (p.61, italics his, formatting mine)

Scharp is right to point out that Liar sentences and their kin appear to be perfectly meaningful. And relatedly, we have very strong pre-theoretical impressions about what they say. These facts call out for explanation, and it is bad for a solution to claim that Liar-like sentences do not mean what we think they mean—whether because they are meaningless or because they mean something else—without giving any further explanation.

At the same time, however, any solution to the Liar paradox will involve violating some of our pre-theoretical impressions, and our pre-theoretical impressions about what our sentences mean are no more immune from revision than any others that one might reasonably take the

[79] See Chapter 2 of (Scharp 2013).

paradox to threaten. One cannot rule out prior to investigation that giving up these impressions is what affords us the overall best solution. The most that Scharp's remarks show is that one who does decide to give up these impressions must be able to explain them away. But as I argue in Chapter 1, this is precisely what my own version of moderate aberrationism does. In that respect, it improves markedly over its radical cousins that reject Semantic Supervenience or Semantic Regularity. Unlike these, my approach allows that Liar-like sentences come close to saying what they seem to say, and allows that what they say is significantly influenced by how we use their component terms and the manner in which they are combined. It is an important theoretical discovery, a surprising lesson of the Liar paradox, that these sentences cannot say what they appear to say. That explains why this was not obvious at the outset.

6. Concluding Remarks

In this essay, I argued that aberrationism can successfully avoid revenge problems, and that it provides a compelling solution(s) to all prominent Liar-like paradoxes, not just the Strong Liar Paradox that I used to introduce it. With respect to revenge, we saw that competing approaches generally face a trilemma: they must either fail to apply to the language in which they are stated, assert the incoherence of the phenomena to which they themselves centrally appeal or that are closely related, or else face a revenge problem. Aberrationism, by contrast, avoids all of these bad alternatives. It can allow for reference to the phenomena that it invokes—viz., component contexts, reference by occurrences, a sentence's being Liar-like, indeterminacy, and ascending truth and descending truth (when it comes to my own preferred view)—because by aberrationist lights, reference to these phenomena cannot in turn be used to construct new, problematic sentences. With respect to the Contingent Liar, Curry, Bad Pair, Yablo, Cook, and

Grelling paradoxes, we saw that aberrationism provides us with a straightforward, appealing, and effective way to avoid contradictions.

In the course of discussing the No-No paradox and certain putative revenge problems for aberrationism, an interesting phenomenon emerged. These paradoxes confronted us with *parity problems*—pairs of profoundly similar sentences that cannot consistently be assigned the same classical truth value, but to which one can consistently assign distinct classical truth values. While all aberrationists are able to solve these parity problems, it became clear that different such theorists will approach these problems in interestingly different ways. It is my hope that the further development of these solutions will become a topic for fruitful research.

In the final section, I made some headway on the important task of formulating a precise criterion of applicability for aberrationism, one that can be used to generalize the approach beyond the Strong Liar. I developed two approximate definitions of 'Liar-like sentence', and then illustrated the appeal of these definitions. I showed that adopting these definitions would allow aberrationism to diagnose and solve a variety of different pre-theoretically Liar-like paradoxes, including the Contingent Liar, Curry, Grelling, Bad Pair, Yablo, and Cook paradoxes, as well as the No-No "paradox," provided that the latter counts as a paradox.

REFERENCES

1. Armour-Garb, Bradley, and Woodbridge, James. (2013) “Semantic Defectiveness and the Liar.” *Philosophical Studies*, Vol. 164, pp.845–863.
2. Barnes, Elizabeth J., and Williams, J. Robert G. (2011) “A Theory of Metaphysical Indeterminacy.” In Bennett, K. & Zimmerman, Dean W. (eds.), *Oxford Studies in Metaphysics*, Volume 6, pp.103-148. Oxford University Press.
3. Brandom, Robert. (1994) *Making it Explicit*. Harvard University Press, Cambridge, MA.
4. Burge, Tyler. (1979) “Semantical Paradox.” *The Journal of Philosophy*, Vol. 76, No. 4, pp.169-198.
5. Feferman, Solomon. (1984) “Toward Useful Type-Free Theories. I” *The Journal of Symbolic Logic*, Vol. 49, No. 1, pp.75-111.
6. Glanzberg, Michael. (2004a) “A Contextual-Hierarchical Approach to Truth and the Liar Paradox.” *Journal of Philosophical Logic*, Vol. 33, No. 1, pp.27-88.
7. Glanzberg, Michael. (2004b) “Truth, Reflection, and Hierarchies.” *Synthese*, Vol. 142, pp.289-315.
8. Goldstein, Laurence. (2009) “A Consistent Way with Paradox.” *Philosophical Studies*, Vol. 144, pp.377–389.
9. Grim, Patrick. (1995) “Book Review: Universality and the Liar: An Essay on Truth and the Diagonal Argument.” *The Philosophical Review*, Vol. 104, No. 3, pp.467-469.
10. Heck, Richard. (2003) “Semantic Accounts of Vagueness.” In J. C. Beall (ed.), *Liars and Heaps*, pp.106-127. Oxford University Press.

11. Icard, Thomas. (2012). Surface Reasoning Lecture 4: Negative Polarity and Antitonicity [PowerPoint slides]. Retrieved from http://www.nasslli2012.com/files/courses/icard-slides-4.pdf
12. Kripke, Saul. (1975) "Outline of A Theory of Truth." *The Journal of Philosophy*, Vol. 72, No. 19, Seventy-Second Annual Meeting American Philosophical Association, Eastern Division, pp.690-716.
13. Kripke, Saul. (1982) *Wittgenstein on Rules and Private Language*. Cambridge, MA: Harvard University Press.
14. Lewis, David. (1970) "How to Define Theoretical Terms." *The Journal of Philosophy*, Vol. 67, No. 13, pp.427-446.
15. McGee, Vann and McLaughlin, Brian. (2004) "Logical Commitment and Semantic Indeterminacy: A Reply to Williamson." *Linguistics and Philosophy*, Vol. 27, pp.123–136.
16. McGee, Vann. (2005) "Inscrutability and Its Discontents." *Noûs*, Vol. 39, No. 3, pp.397–425.
17. McLaughlin, Brian and Bennett, Karen, "Supervenience," *The Stanford Encyclopedia of Philosophy* (Spring 2014 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/spr2014/entries/supervenience/>.
18. Priest, Graham. (1990) "Boolean Negation, and All That." *Journal of Philosophical Logic*, Vol. 19, No.2, pp.201-215.
19. Quine, Wilard Van Orman. (1960) *Word and Object*, Cambridge, Massachusetts: MIT Press.
20. Scharp, Kevin. (2013) *Replacing Truth*. Oxford, U.K., Oxford University Press.

21. Simmons, Keith. *Universality and the Liar*, Cambridge, U.K., Cambridge University Press, 1993.
22. Smith, Nicholas. "Semantic Regularity and the Liar Paradox." *The Monist*, Vol. 89, No.1, (2006), pp.178-202.
23. Sorensen, Roy. (2001): *Vagueness and Contradiction*, Oxford: Oxford University Press.
24. Tarski, Alfred. (1935) "Der Wahrheitsbegriff in den formalisierten Sprachen," *Studia Philosophica*, Vol.1, pp.261–405.
25. Tarski, Alfred. (1944) "The Semantic Conception of Truth: and the Foundations of Semantics." *Philosophy and Phenomenological Research*, Vol. 4, No. 3, pp.341-376.
26. Wilson, W. Kent. (1990). "Some Reflections on the Prosentential Theory of Truth." In J. M. Dunn & A. Gupta (eds.), *Truth or Consequences*, pp.19-32. Dordrecht: Kluwer Academic Publishers.

CHAPTER 3

INTENTIONALITY AND PSYCHOLOGICAL EXPLANATION

1. ***Introduction***
2. ***Field's Pilot Case***
3. ***Dynamical Approaches to Cognition***
4. ***Intentionality in the Explanandum***
5. ***Concluding Remarks***

1. Introduction

1.1. Definitions

In this essay, I will defend the claim that intentionality plays an ineliminable role in psychological explanation, both when it comes to explanantia and when it comes to explananda. With explanantia, a central case in point will be the talk of beliefs and desires that occurs in everyday life, the humanities, and the social sciences. I will argue that when we explain someone's behaviors by saying what she believes and desires, these explanations also shed light on how she would react to various counterfactual circumstances. This kind of modal informativeness is a distinctive virtue of our folk-psychological explanations, which has not yet been replicated in explanations that are framed in entirely non-intentional terms. Until we have found a way to replicate this virtue, it is premature to insist that intentional talk, and in particular talk of beliefs and desires, is dispensable when it comes to explaining people's behaviors. Although it concerns our everyday explanations, this point is also relevant to the status of intentional talk in scientific explanations, insofar as the distinctive modal informativeness afforded by intentional talk is a feature that also matters for scientific explanation.

As with explanantia, one may likewise wonder whether intentional talk can be excised from all characterizations of explananda in psychology. There are, I will argue, important reasons to think it cannot. Explananda in psychology are overwhelmingly characterized in intentional terms, and so as long as intentionality continues to resist attempts at reduction, switching to non-intentional characterizations of the explananda will amount to simply changing what is being explained. So, the status of intentional talk in psychology is at least as secure as the propriety of its current explananda. Moreover, I will show that a prominent explanatory program

that is embraced by many would-be skeptics of the necessity of intentional talk itself indulges in intentional characterizations of its explananda.

Such are my positions and the arguments for them, in broadest strokes. To give a more precise sense of what I want to say, I'll begin by clarifying the meanings of 'intentionality' and 'psychological explanation'. Following common use among philosophers, call a thing *intentional* if it has or purports to have[1] a *subject matter*: something that it is of, is about, or represents. And similarly, a characterization of an entity is *an intentional characterization*, or equivalently, *characterizes its subject matter in intentional terms*, if it presents that entity as having a subject matter.[2] The examples of intentional entities and intentional characterizations are ubiquitous. For instance, when someone believes that snow is white, her belief is about snow and whiteness; it has snow and whiteness as its subject matter. The belief itself represents snow as being white, and so does the believer, in having that belief. In addition to beliefs, one can give a long list of familiar things that are intentional.[3] And beyond familiar things, the mental representations invoked in cognitive science[4] are also intentional; or at least, they are standardly characterized in

[1] I want to count as intentional items like the name 'Santa Claus', whose purported subject matter doesn't exist. For brevity, I'll simply speak of such items as having a subject matter.

[2] Of course, intentional characterizations are also intentional in the straightforward sense given a moment ago: they themselves have a subject matter (that being the thing(s) of which they are characterizations). Since this is obvious, in the rest of the paper when I say that a characterization is intentional I will mean that it presents its subject matter as having a subject matter. I only raise the point here in order to observe that on the most straightforward interpretation of 'intentional characterization'—namely, the interpretation given in the main text—one who claims to want to get rid of intentional characterization has, in making this claim, herself indulged in it.

[3] Here are some more examples. Concepts are intentional in my sense of the term. E.g., the subject matter of anyone's concept of Barack Obama is Barack Obama. Perception is another case: when someone perceives something, the subject matter of her act of perception is the thing(s) she perceives, and, if applicable, the properties that she perceives it as having or the relations in which she perceives it as standing. Desires are also intentional: the subject matter of anyone's desire that Hillary Clinton be president of the U.S. consists of Hillary Clinton and the property *being president of the U.S.* Certain bits of language are another case: declarative sentences, imperative sentences, and words are intentional in much the way that beliefs, desires, and concepts, are, respectively. Maps and pictures are also intentional in my sense of the term: their subject matter consists of the objects, properties, and relations that they depict.

[4] To give just one example, (Churchland 2012) claims that people have "neural maps" of the spectrum of humanly visible colors. Never mind the details about what exactly these maps are and how they work; the point is that if this claim is correct, then our neural maps have these colors and the relations between them as their subject matter.

intentional terms. For these entities, their (purported) subject matter consists of the things they represent and the ways they represent these things as being. Overall, my point here is that intentionality and intentional characterization are ubiquitous; a great many familiar entities and descriptions are thoroughly saturated with it, and it has shown up in cognitive science as well.

Now I'll say a little about 'psychological' and 'explanation'. *Psychological explanation* is the explanation of people's and sophisticated animals' behaviors and cognitive capacities. The range of these behaviors and capacities is quite broad, including basic capabilities such as the ability to perceive the distances of external objects, but also complex behaviors like political decision-making. Within psychological explanations we can distinguish those that are familiar from everyday life, those that appear in current cognitive science, and those that would appear in a cognitive science that was ideal in various respects.

Within explanations more generally, we can distinguish between *communicative explanation* and *ontic explanation*.[5] *Communicatively explaining* is something that people do. More specifically, it is a type of speech act—that is, a type of action that one can perform by uttering words or displaying images. For example, when we say that Rebecca explains how action potentials work, we are identifying something that Rebecca does, by, e.g., giving a lecture or displaying a diagram. That act which Rebecca performs is communicative explanation. Similarly, I will take *communicative explanations* to be what is said when someone performs an act of communicatively explaining something. Thus, I will take communicative explanations to be arguments that consist of propositions. As a convenient shorthand, I will say that a communicative explanation *mentions* (or invokes, or refers to, or posits, etc.) an object, property, or relation if giving that explanation involves mentioning (invoking, referring to, positing) that

[5] For more on this distinction see (Craver 2014), p.30. Craver also introduces two further kinds of explanation, which it will be safe to ignore here. I owe the example in the main text to Craver.

object, property, or relation. Throughout, I will use 'explains' and 'explanation' interchangeably with 'communicatively explains' and 'communicative explanation', except where I explicitly say otherwise.

Whereas Rebecca might explain how action potentials work to her students, it also seems correct to say that the working of action potentials is explained by the flux of ions across neuronal membranes. This latter kind of explanation is *ontic explanation*; it is a relationship that can hold between phenomena "out in the world," so to speak, rather than just between a phenomenon "out in the world" and a human communicating about that phenomenon. An illustrative example is the *mechanistic conception of ontic explanation*, according to which what ontically explains something is what produces or constitutes it. On that view, what ontically explains how action potentials work is not Rebecca but rather the flux of ions across neuronal membranes, since ion flux is something that helps to constitute the working of action potentials.

The issue of the relationship between communicative and ontic explanation will loom large in what follows, in the form of two different conceptions of communicative explanation. Both conceptions take (communicatively) explaining to be a kind of speech act, but they differ in what they require of a speech act for it to count as an act of explaining. The *metaphysical conception of communicative explanation*[6] takes ontic and communicative explanation to be tightly connected. It says that to count as an act of explaining, an act must involve description of the entities and activities that ontically explain the explanandum. This view can take different forms, depending on one's views about ontic explanation. One particularly clear example, which I will return to frequently in what follows, is *the mechanistic conception of communicative explanation*, which combines mechanistic conception of ontic explanation with the metaphysical

[6] A more standard term here is 'ontic conception', but I use 'metaphysical' since I am already using 'ontic' in a different, (also standard) way, in the phrase 'ontic explanation'.

conception of communicative explanation. This view says that to count as an act of explaining, an act must involve description of the entities and activities that produce or constitute the explanandum.[7] On this view, Rebecca can succeed in (communicatively) explaining how action potentials work only if she describes the entities and activities that give rise to (that is, ontically explain) the working of action potentials.

Importantly, on the metaphysical conception of communicative explanation, what makes a speech act an act of explanation is the nature and properties of the entities that the speaker posits. By contrast, on the *epistemic conception of communicative explanation*, what makes an act one of explaining is not the nature of the entities posited but rather the virtues of the ways these entities are characterized. For instance, these characterizations might aid in producing understanding,[8] help us make useful predictions,[9] or facilitate manipulation and control of the explanandum.[10,11] In what follows I will stay neutral between the metaphysical and epistemic conceptions, exploring the consequences of each.

[7] This position is also commonly phrased as the claim that communicative explanations must specify the structures and causes at work in the systems whose behaviors they explain. Variants of the mechanistic conceptions of communicative explanation are advocated in (Machamer et al. 2000), (Bechtel and Abrahamsen 2005), (Craver 2006), (Craver 2007), (Bechtel 2008), (Eliasmith 2010), (Kaplan and Craver 2011), and (Hochstein 2012).
[8] See (Achinstein 1983), (Waskan et al. 2014), and (Braverman et al. 2012).
[9] See (Batterman 2001), (Batterman 2002), (Hochstein 2012), and (Rice 2013).
[10] For this last view see (Woodward 2000) and (Woodward 2003).
[11] Of course, one's views about what it takes to enhance prediction, understanding, or manipulation may in turn simply lead one back to the metaphysical conception. For example, one might claim that understanding consists in having true, relevant beliefs about the entities and activities that produce the explanandum—so that then something counts as an explanation only if it correctly identifies the entities and activities that produce the explanandum, as on the mechanistic conception of communicative explanation, a version of the metaphysical conception. Or, with a similar result, one might claim that the descriptions that best enable us to make useful predictions simply are those that identify the structures and causes of the systems they describe, because such descriptions allow us to directly intervene in the workings of the system to determine how it behaves in various circumstances. In both cases above, supplementing the epistemic conception of explanation with certain further claims leads back to the metaphysical conception, by shifting the emphasis back onto the entities described rather than the descriptions themselves. However, since these further claims (about prediction, understanding, manipulation, etc.) are themselves controversial, until they are decided, we do best to allow ourselves a distinction between the metaphysical and the epistemic conception of communicative explanation.

The metaphysical and epistemic conceptions yield, in turn, two different senses in which a communicative explanation could be said to involve intentionality. On the metaphysical conception, a communicative explanation is an *intentional explanation* if its explanandum or explanans is something which is, or which reduces something that is, intentional, regardless of whether the explanation employs any intentional characterizations. Here we need not concern ourselves too much with what exactly reduction is; what will matter for my purposes is that for all X and Y, if X reduces to Y then in any situation in which Y obtains, so does X.

On the epistemic conception, by contrast, an *intentional explanation* is one that employs an intentional characterization of its explanandum or explanans—that is, again, the language used in giving the explanation presents the explanandum or the explanans as having a subject matter. It is easy to see why these definitions are appropriate. On the metaphysical conception, what matters for explanation are the nature and properties of the entities described, as opposed to the characterizations employed. So, it is fitting that on this conception, an explanation is intentional (or not) depending on the intentionality (or not) of the entities it describes. By contrast, on the epistemic conception what matters for explanation are only the virtues of the characterizations employed. Fittingly, then, on this conception an explanation is intentional (or not) depending on the intentionality (or not) of the characterizations used in describing the explanantia and explananda.

1.2. Skepticism About the Explanatory role of Intentionality

However one understands 'intentional explanation', it is clear that intentional explanations are a staple of everyday discourse, the humanities, and many social sciences. For example, we quite often explain why someone behaves as she does, or how she is capable of

doing something, by saying what she believes and desires. And many explananda in psychology, such as the capacity to perceive the distances of external objects, to add numbers, and to recognize human faces, are standardly characterized in intentional terms.

Despite their ubiquity, however, intentional explanations have been subject to some skepticism. The following is frequently taken for granted by numerous philosophers:

> (Elim) Intentional explanations can always, in principle, be eliminated in favor of non-intentional explanations, and performing such eliminations would almost always yield better explanations.

Below I will present and assess some arguments for (Elim), considering how it fares when interpreted, respectively, under the metaphysical and the epistemic conceptions of explanation.

Advocates of (Elim) are to that extent skeptical of the explanatory importance of intentionality, so throughout I will refer to them as *skeptics*. Still it is worth noting straightaway that skeptics are not *ipso facto* committed to the doctrine that intentionality does not exist.[12] Moreover, there are philosophers who endorse something like (Elim) for some but not all kinds of intentional explanations. For example, (Churchland 2012) only targets explanations that involve *propositional representation*.

Propositional representation will figure frequently in what follows, so it is worth taking a moment to understand. *Propositional representation* is representation in which what is represented can be expressed using a *that-phrase*, such as the phrase 'that snow is white' as it occurs sentences like 'Tamar believes that snow is white'. Believing and desiring are the most familiar varieties. Similarly, *propositional representations* are intentional items, such as beliefs and desires, whose contents can be specified using that-phrases. And *propositional explanations*

[12] In earlier work, such as (Churchland 1979), (Churchland 1981), Churchland goes so far as to deny the existence of beliefs and desires. (Churchland 2012) is less explicit on this matter. At any rate, in (Churchland 2012), Churchland countenances some intentionality, since he advocates the existence of map-like mental representations that represent various aspects of the external world. (See Section 4 below.) For a similar view see also (Eliasmith 2013).

are explanations that involve propositional representation, either in the explanans or the explanandum. Folk-psychological explanations of people's behaviors in terms of (among other things) their beliefs and desires are the most familiar propositional explanations. Throughout his career, Churchland has persistently raised concerns about all of the above.

In addition to the propositional / non-propositional distinction, two further distinctions will be important in what follows. One is the distinction between explanation in everyday life and in science. Most skeptics claim only that intentionality can be eliminated from science, not that it can be eliminated from everyday discourse.[13] Similarly, because psychology as it currently stands is saturated with intentionality (see Sections 3 and 4), skeptics are best conceived as claiming that an ideal science would eschew intentional explanation. Lastly, different skeptics might variously advocate (Elim) for explanations that involve intentionality in the explanans, for those that involve it in the explanandum, or for all intentional explanations whatsoever.

Given that the available skeptical positions exhibit such variety, one wonders which ones I will be addressing. In Section 2, I will discuss an argument due to Hartry Field that works under the metaphysical conception of communicative explanation, and that, as stated, targets the invocation of (instances of) propositional representation as explanantia. Despite its focus on propositional representation, however, we will see that if Field's argument is good then it easily generalizes to threaten all invocation of intentional facts as explanantia. Moreover, although Field's argument targets everyday intentional notions like that of belief, the argument's failure bears significantly on the appropriateness of intentional discourse in science. Everyday intentional explanations, I will argue, are informative in a distinctive way. If this kind of informativeness matters in science, then the fact that no one has found a way to achieve it

[13] (Stich 1983) and (Dennett 1987) can be read in this way. See also my discussion of Hartry Field in Section 2.1 below.

without involving intentionality gives us a reason to believe that (Elim) is false as concerns scientific explanations, as well as our everyday ones.

Whereas Field works under the metaphysical conception of explanation, in Section 3, I will discuss a skeptical position which also targets explanantia but embraces the epistemic conception. The focus will be on (Chemero 2009), who can be interpreted as arguing for (Elim)[14] by extrapolating from empirical, scientific studies rather than by invoking the assumption that is central to Field's argument, (Jerry Rig) below. Then in Section 4 I will turn to the matter of intentionality and explananda in psychology. I will examine, respectively, skeptical positions under this heading that adopt the metaphysical (4.2) and the epistemic (4.3) conceptions of communicative explanation.

2. Field's Pilot Case

2.1. Background

In (Field 2001), Hartry Field gives an argument for (Elim). Field's argument is an especially sophisticated instance of a more general kind of argument for (Elim) that many different philosophers have given, whose characteristic feature is reliance on the following premise:

(Jerry Rig) Multiple radically different and incompatible sets of intentional descriptions can always be jerry-rigged *ad hoc* to provide equally good explanations of the same phenomena.[15]

[14] I am uncertain whether Chemero actually embraces (Elim). His primary aim is to argue against the invocation of *internal representations* in psychology—that is, he is concerned to show that the manifestation of psychological capacities emerges from dynamic, closely-coupled interactions between the agent and the world, rather than from the manipulation of internal items (that is, *internal representations*) that serve as intermediaries between the agent and the external world. This position can easily take on an anti-intentional appearance, since the most obvious bearers of intentionality are internal items such as concepts and states of belief and desire. If, however, Chemero does not embrace (Elim), then one should interpret my discussion of him as being directed toward philosophers who would use his views to justify (Elim).

[15] I thank an anonymous reader for help with the formulation of (Jerry Rig). (Jerry Rig) also admits of a linguistic

While (Jerry Rig) owes its origins to Chapter 2 of (Quine 1960), many different philosophers continue to endorse it in one form or another.[16] However, focusing on Field's argument will be especially fruitful and informative. Operating against the backdrop of his reductionist orientation and metaphysical conception of communicative explanation will bring out a number of interesting points, and will shed light on the kinds of phenomena to which we should expect to appeal in attempting to explain intentional facts of various sorts.

2.2. Field's Argument

In the postscript to (Field 1994) in (Field 2001), Field writes:

> It is perfectly obvious that in explaining how [any given[17]] pilot manages to land a plane safely with some regularity, one will appeal to the fact that she has a good many true beliefs: beliefs about her airspeed at any moment, about whether she is above or below the glideslope, about her altitude with respect to the ground, about which runway is in use, and so forth. (None of these beliefs need be based very directly on observation; she might be flying in bad weather with some of her instruments not working, so that she must rely on complicated cues). (p.153)

From these remarks one might be tempted to infer that talk of beliefs plays an ineliminable role in some explanations of how some people are able to do certain things. But then Field tells us

---

version, which readers of (Quine 1960) will recognize:

(Interpretations) When interpreting a language, there are always multiple, radically different interpretations that serve equally well for explaining the behaviors of the language users.

Much of I want to say applies to (Interpretations) as well as (Jerry Rig), but for ease of exposition I'll focus on (Jerry Rig).

[16] In addition to (Quine 1960) Chapter 2, some especially prominent and influential examples include (Quine 1992) Chapter 3, (Churchland 1979) p.94, (Stich 1983), (Dennett 1987) p.342, and, as we will see, (Field 2001) pp.153-156. It is also worth noting that some claims from (Kripke 1982) have a similar flavor, although there the role of interpretation in the explanation of behavior is not explicitly mentioned.

[17] Given the context, it is plausible that Field takes his argument to show that, quite generally, reference to intentional phenomena (or to reduction bases for some such phenomena) is never essential in order to explain anything. Evidence of this view is abundant throughout Field's other writings; for some examples see (Field 1986) p.84 bottom paragraph, (Field 1994) the paragraph straddling p.254 and p.255, and (Field 2001) pp.153-155. So, we should read Field as saying something about *all* pilots, not just some.

that "talk of representation is serving a merely heuristic role in the explanation of the pilot's ability."

Although Field is not explicit in making this claim, here the phrase 'merely heuristic' carries with it a strong suggestion that whatever falls under this classification has no place in a truly scientific psychology. For on the most common way of understanding 'heuristic', something serves a heuristic role if it serves merely to facilitate further investigation, if it merely enables people to discover or learn something for themselves, as opposed to being, itself, directly or explicitly informative. Presumably, then, anything that serves a merely heuristic role can under more opportune circumstances be eliminated,[18] replaced by something more directly or explicitly informative. Indeed, elsewhere Field tellingly refers to intentional explanations as being "second-class" explanations (see Section 2.4.2 below), and considers non-intentional ways of replacing intentional talk.[19]

In particular, right after presenting the standard intentional explanation of how a pilot is able regularly to land a plane, he goes on to sketch a generic non-intentional explanation of the pilot's ability, an explanation which would, in an ideal psychology, presumably replace the everyday one:

> put without the heuristic, the explanation…involves the existence of some class C of internal representations[20] in the pilot and two subclasses C_1 and C_2 of C such that
>
> (Behavioral Effects) when she believes a representation in C_1 she slows the plane and when she believes one in C_2 she speeds it up, and
>
> (Relation to Airspeeds) there is a 1-1 function *f* from C to a certain set of real numbers such that

[18] The claim I attribute here is also evidenced by Field's remarks elsewhere, e.g., (Field 1986) p.84, where Field considers a non-intentional way of replacing intentional talk in an explanation.
[19] See, respectively, (Field 2001) p.55, (Field 1994) pp.254-255, (Field 1986) p.84, and immediately below in the main text.
[20] Given the dialectical context, Field must be using 'representations' and 'believes a representation' to describe things that are non-intentional. See below in the main text.

a) C_1 is that subclass of C that is mapped into numbers above a certain threshold and C_2 is that subclass of C that is mapped into numbers below a certain (slightly lower) threshold, and
b) she tends to believe a representation *r* in C when the airspeed in knots is approximately $f(r)$.

(Field 2001, pp.153-155, formatting and labels mine)

(The "certain thresholds" are the threshold above which the pilot's airspeed is too fast for her safely to land, and below which it is too slow.)

An immediate reaction is that since Field's explanation involves positing a class C of internal "representations" that the agent "believes," it invokes some intentional facts. Given the terminology, the reader could hardly be blamed for supposing that the class C represents the pilot's possible airspeeds, that each of its elements represents some particular airspeed, and that for each element *r* of C, there is a number *n* such that when the pilot "believes" *r,* she thereby represents the plane as having an airspeed of *n*. If all that is correct, then Field's example does not show that talk of propositional representation, let alone all talk whatsoever of intentional facts, can in principle be safely excised from our explanations.

Clearly, then, the most charitable interpretation of Field's remarks is that at the very least, intentionality cannot simply be read off of Field's description. That is, talk of an agent's "believing" a "representation", as he understands it, does not amount to intentional characterization; it is not to be taken for granted from the beginning that Field's "representations" are representations of anything in particular, or that "believing" a "representation" involves representing anything(s) as having any properties or as standing in any

relations.[21] Rather, we should assume, if indeed "believing" a "representation" involves intentionality then that is something that will have to be shown.

Of course, the anti-skeptic can be expected to insist immediately that it can be shown. That is, even if Field's explanation does not employ any intentional characterizations (hence the quote marks around 'representations' and 'believes'), the anti-skeptic could still insist that Field's explanantia are in fact intentional. In particular, she could insist that the properties that Field invokes reduce the intentional properties attributed in the original intentional explanation: *having a class C of internal "representations" that satisfies (Behavioral Effects) and (Relation to Airspeeds)* reduces *having available a set of beliefs about one's possible airspeeds*, and *"believing" a "representation" in C* reduces *tokening a belief about one's airspeeds*.

It is worth noticing that this kind of anti-skeptical move can be repeated against any advocate of (Elim). Any advocate of (Elim) will hold that all intentional talk in our explanations could in principle and should wherever possible be replaced by some non-intentional analog(s). But in response, and whatever the specifics of the skeptic's proposal, the question can then be raised whether this analog appeals to the reduction bases for some intentional properties. For skeptics who adopt a metaphysical conception of communicative explanation, this is an important question, since a positive answer would render the purportedly non-intentional explanation, by their lights, intentional "in fact if not in name." (Here I use Field's phrase.[22]) So,

[21] In addition to Field's stated argumentative goal (Field 2001 p.153), there is some further evidence supporting this non-intentional interpretation. See (Field 1978) (Field 1994) p.254 where he talks of sentences, or "sentence-analogues," as being objects of belief. Also, in (Field 1978) p.18, he says that representations have "the same kind of meaning or content which sentences have." In (Field 1978), Field proposes to explain belief (the ordinary, intentional relation) by "factoring" it into two further relations, both definable in non-intentional terms: a dispositional relation, belief*, which one can bear to a representation, and a relation R which a representation can bear to a proposition. Throughout, he suggests that believing* a sentence involves having certain dispositions with respect to it (seep.13 and p.17). So, there is reason to think that belief* can be understood in non-intentional terms. (Field 1978) would understand (Field 2001)'s talk of believing representations as talk of believing* representations, and he would understand (Relation to Airspeeds)(b) to be a description of an instance of R.

[22] (Field 1994) bottom p.254.

even if one does not like Field's particular non-intentional explanation, the same kind of response is available against other skeptics.

(On the other hand, it is worth noting, a skeptic who adopts the epistemic conception of communicative explanation would hold that an act of explanation owes its quality not to what entities it describes, but rather to the descriptions themselves. Thus, she could insist that the explanation given in non-intentional terms is superior, even though the phenomena it invokes can be accurately described in intentional terms. Insofar as Field is working under the metaphysical conception (recall Section 2.1) this move is not available to him.)

Field anticipates that the anti-skeptic might claim his explanantia as a reduction base for something intentional, and he has a reply.[23] He grants that "part [(Relation to Airspeeds)(b)] of the explanation uses an indication relation,"[24] but insists that

> The function [*f*] mapping internal representations into airspeeds needn't…give the intuitive truth conditions of [the pilot's] representations in all cases: one could tell a story in which the pilot's beliefs about what she was doing were so weird that it would be natural to assign quite different truth conditions to her representations. (Perhaps she believes she isn't in an airplane at all, but is using the controls to direct U.S. ground forces on a foreign mission). (Field 2001, p.154)

Two important observations about this reply of Field's. Firstly, Field's reply amounts to an invocation of (Jerry Rig), since it assumes that, as far as intentional explanations go, one could get by just as well by attributing an extreme delusion. This is what I take Field to be getting at when he says "one could tell a story"; there would be no reason to point out that one could tell

[23] The fact that Field responds to this reaction evidences that he is working under the metaphysical conception of communicative explanation. For as we just saw, it is only under that conception that the appeal to the reduction base for something intentional counts as a problem for the skeptic.

[24] *Indication* has been understood in different ways in the literature. (Stampe 1977), (Dretske 1981), and (Stalnaker 1984) p.18 have appealed to indication in trying to explain intentionality. The most common theme in definitions of indication is that one state of affairs indicates another only if they are reliably correlated. For instance, if the cross section of a tree trunk has 54 rings, then that indicates that the tree is 54 years old.

such a story, unless telling that story would amount to giving an explanation that was just as good as the original.

Secondly, although Field's primary concern is with propositional representation,[25] his remarks here commit him to a strong claim, one that concerns all intentionality. The key observation here is that when Field's story shifts the beliefs attributed to the pilot, it does so in a way that shifts the subject matter attributed to the pilot's ideation (airspeeds vs. U.S. ground forces). This move of shifting the subject matter can be applied to any intentional explanation, since all it requires is that the explanans have a subject matter; one simply replaces the original explanans with something that has a different subject matter, and then claims that the result is an equally good explanation, as far as intentional explanations go.[26] It is because Field's reply to the anti-skeptic can be applied to any intentional explanation that I take him to be committed to (Elim) in its most general form, targeting all intentionality rather than just propositional representation. Field is not alone in this. To date, all extant arguments that invoke (Jerry Rig) share the wide scope of Field's.[27]

Before moving toward a response, let us briefly review the dialectic. Earlier on, Field claimed that describing such facts as those specified in (Behavioral Effects) and (Relation to Airspeeds) is all that one needs to do in order adequately to explain any pilot's ability to land a plane. Conceding this, the anti-skeptic then proposed that *"believing" a representation which*

[25] This is because propositional representation involves truth-conditions, and Field's ultimate concern is to motivate what he calls a "deflationary" attitude about truth. See (Field 1994) p.271 as well as (Field 2001) pp.153-156.
[26] On the other hand, imagine that Field's story kept fixed that the pilot had beliefs which were about her airspeeds, and shifted only what these beliefs said about these airspeeds. Stated thus, the argument would apply only to things like beliefs, which not only have a subject matter but, moreover, say something about that subject matter. It would be harder to generalize the argument to things like words and concepts, which have a subject matter (a referent) but need not say anything about it.
[27] Compare the argument from (Quine 1960) Chapter 2, which shifts the subject matter from rabbits to undetached rabbit parts or rabbit-stages. Similarly, (Kripke 1982) and (Field 1994) shift not the truth-conditions ascribed to a sentence, but rather the referent ascribed to a word, 'plus' and 'or', respectively.

belongings to a set C that satisfies (Behavioral Effects) and (Relation to Airspeeds) reduces *having a belief about one's airspeed*. In reply, Field now produces a counterexample. In general, for A to reduce B, A must be necessarily sufficient for B; and Field's delusional pilot is meant to give us a possible counterexample to sufficiency.[28] Field is claiming that for some possible pilots who have a set of "representations" that satisfy (Behavioral Effects) and (Relation to Airspeeds), their "representations" do not represent their airspeeds; rather, these "representations" represent movements of U.S. ground forces.[29] Due to this failure of reduction, it cannot be said that Field's explanation in terms of "representations" ends up being intentional in fact if not in name.[30] So, if this explanation is indeed just as good as the everyday intentional one, then that counts in favor of the view that invocation of intentional facts can be harmlessly excised from psychological explanations.

At this point, one might reasonably be puzzled about the role of (Jerry Rig) in Field's argument. After all, the claim that a pilot with an extreme delusion could have a set of

[28] In this connection, note the significance of the fact that the delusion is an extreme one. Field needs an extreme delusion, in order to show that the non-intentional facts that (ontically) explain the pilot's ability do very little to constrain the subject matter of her beliefs. For example, suppose the pilot merely believed that her altimeter was off by two feet. That would not be different enough from the original, sane belief to make Field's point. It is like the situation with indeterminacy of translation: recognizing that there is some modest degree of indeterminacy doesn't support (Quine 1960)'s strong conclusions.

[29] More precisely: for some possible worlds w_1 and w_2 and individuals x in w_1 and y in w_2, x in w_1 is indiscernible from y in w_2 with respect to satisfaction of (Behavioral Effects) and (Relation to Airspeeds), but in w_1, the representations of x which are in (x's version of) C represent x's airspeed, whereas in w_2 the representations of y which are in (y's version of) C do *not* represent y's airspeed. Rather, they represent movements of U.S. ground troops.

[30] Now, technically, as long as *"believing" a representation which belongings to a set C that satisfies (Behavioral Effects) and (Relation to Airspeeds)* reduces *some* intentional property, the explanatory importance of intentional properties is in no serious trouble; all that is in trouble is the idea that the intentional properties that we ordinarily posit are essential for explaining anything. For all that conclusion shows, there might be intentional properties of some other kinds that are indeed essential. However, admittedly it is hard to see what intentional properties *"believing" a representation which belongings to a set C that satisfies (Behavioral Effects) and (Relation to Airspeeds)* could reduce, if it could underlie *having a belief about one's airspeed* in one pilot and underlie *having a belief about U.S. ground troops* in another. In particular, it is hard to see what could be the subject matter of states that in one pilot would count as beliefs about airspeeds and in another would count as beliefs about ground troops. So, I'll grant that there is a genuine threat to the explanatory importance of intentionality here.

"representations" that satisfies (Behavioral Effects) and (Relation to Airspeeds) is not identical with the claim that as far as intentional explanations go, the attribution of an extreme delusion works just as well as an explanation as the attribution of normal beliefs. However, these two claims are rendered equivalent by an assumption that is shared by all parties to the debate at this point in the dialectic: namely, that the only property that is relevant to explaining a pilot's landing ability is the property *having a set of "representations" that satisfies (Behavioral Effects) and (Relation to Airspeeds)*.[31] Since both a sane and an extremely delusional pilot could have this property (in the latter case, as long as the delusion has certain very specific properties,) Field's idea is then that both belief-attributions can serve equally well in the heuristic role of guiding us to this property, provided that the attribution of delusional beliefs is suitably tweaked. Given Field's assumption that intentional explanations serve a heuristic role, the claim that both belief attributions serve equally well as guides to the same, ontically-explanatory non-intentional property is identical with the claim that both intentional explanations are equally good, as far as intentional explanations go.

2.3. A Characteristic Virtue of Intentional Explanations

Field is right that (Behavioral Effects) and (Relation to Airspeeds) do not reduce *having beliefs about one's airspeeds*. But he is wrong to suggest that one could profitably replace the original intentional explanation with the non-intentional one that he provides. There may be some kinds of explanation of a pilot's ability reliably to land a plane that need only describe the

[31] Field makes this assumption when he claims that the attribution of beliefs to the pilot serves as a heuristic for the attribution of this property. And the anti-skeptic implicitly grants the assumption when she proposes this property as a reduction base for something intentional. What she needs to show is that the property that ontically explains the pilot's capacity to land the plane reduces some intentional property. So, there is no need for her to consider the property that Field identifies unless she grants that it really is the property that ontically explains the pilot's capacity.

likes of (Behavioral Effects) and (Relation to Airspeeds), for instance, any kind of explanation whose only purpose is to say something about the pilot's "representations" which entails that the pilot tends to land the plane. Perhaps cognitive science has or would ideally have some uses for such explanations. However, both in everyday life and in the sciences, we frequently want our explanations to do more than just entail the occurrence of the phenomenon being explained. Below I'll describe in detail one important further thing that everyday intentional explanations do.

Let's give the name 'folk psychology' to our everyday conception of humans as sometimes-rational agents who act on the basis of beliefs, desires, fears, suspicions, hopes, traits of character, etc. For our purposes here, it will ease exposition to assume that folk psychology is a theory that traffics in *ceteris paribus* generalizations which relate beliefs, desires, and other intentional states, often together with traits of character, to behaviors.[32] An example of such a generalization in the pilot case might say something like this: *ceteris paribus*, a pilot who wants to land the plane she is flying, who believes her airspeed to be approximately *n* knots and her height to be approximately *m* meters, who believes that airspeeds of *n* knots at heights of *m* meters are too fast for landing the plane, and who believes that doing *X* will slow the plane, will do *X*.

A common virtue of many good psychological explanations is that they provide modal information about the agents whose behaviors they explain.[33] In particular, define the

[32] The idea that folk psychology involves some such generalizations is a common one. For some advocates, see (Stich 1983) pp.130-133 and (Fodor 1987) pp.1-6 and 10. However, (Anscombe 1971) and (Miller 1987) deny that explanation in folk psychology generally involves appeal to any *ceteris paribus* generalizations. So, the assumption that I make here is controversial. However, while there is not space to attempt it here, I am confident that the main points I want to make can be put without embracing this assumption.

[33] Here I am merely saying that modal informativeness of a certain kind is a good-making feature of many explanations. It is worth distinguishing this claim from versions of the epistemic conception of communicative explanation which insist that providing such modal information is required for something to count as an explanation.

counterfactual spread[34] of a psychological explanation to be the information that it provides about how the agent would behave under various counterfactual circumstances. Put in these terms, my main point is that everyday intentional explanations enjoy a distinctive, wide counterfactual spread. This spread is *wide* in the sense that it contains information about a wide variety of different scenarios.[35]

The counterfactual spread of an everyday intentional explanation arises from the interaction of one salient part of the explanans—in the pilot case, this would be propositions about the pilot's beliefs, desires and perceptions that are directly relevant to her capacity to land the plane—with background propositions (*ceteris paribus* generalizations) that form part of folk psychology. In various alternative circumstances, many of those most relevant states would still obtain (e.g., those states that are most relevant to how to fly an airplane), and the same folk-psychological background propositions would be in play. However, some salient parts of the explanans would undergo replacement by attitudes that the pilot doesn't actually have. The background propositions tell us how the pilot would behave in those scenarios. In this way, an intentional explanation of an episode of behavior provides some information about how the agent would behave under various circumstances other than the circumstances of that episode.

We just saw how everyday intentional explanations acquire wide counterfactual spreads by employing concepts that figure in the background theory of folk psychology. In principle, this can also happen with other kinds of explanations and other theories. Say that a theory about minds (or some minds) is *holistic* if, for a wide variety of cognitive and affective states, the

Someone who makes only the claim I make here could, for instance, embrace the mechanistic conception of communicative explanation, maintaining that what makes something an explanation is that it describes the objects and activities that produce the explanandum.

[34] I owe this phrase to Harold Hodes.

[35] Field recognizes this point in (Field 2001, p.78).

theory describes a wide variety of relations that these states bear to one another, to the state-bearers' external environments (including, perhaps, the states and behaviors of other cognitive agents), and to the state-bearers' behaviors, dispositions, and capacities. In general, a psychological explanation can acquire a wide counterfactual spread by attributing states (types) that figure in a holistic background theory.

To see an example outside of everyday intentional explanations, let us look again at Field's "representations", letting $b_0,\ldots,b_n$ be the "representations" described in (Behavioral Effects) and (Relation to Airspeeds). Then a holistic theory about the pilot would be one that not only describes how $b_0,\ldots,b_n$ are implicated in the pilot's capacity to land planes, but also connects the very simple regularities described in (Behavioral Effects) and (Relation to Airspeeds) with a much broader set of regularities governing the ways that $b_0,\ldots,b_n$ function in the pilot. These broader regularities might include some causal links between tokenings of $b_0,\ldots,b_n$ and tokenings of various further "representations" of the pilot, as well as some of the relations in which these further "representations" stand to the pilot's bodily movements and to her environment, and maybe even some of the relations that tokenings of $b_0,\ldots,b_n$ bear to tokenings of various further "representations" used by other members of the pilot's linguistic community. A theory that described relationships such as these would be holistic, in the sense that it would situate $b_0,\ldots,b_n$ and the pilot's capacity to land planes in a bigger picture. When given against the backdrop of such a theory, an explanation that described the contributions of $b_0,\ldots,b_n$ to the pilot's capacity to land planes would have a wide counterfactual spread, since some of the propositions in the explanans—such as propositions to the effect that the pilot tokens $b_0,\ldots,b_n$—would also feature in a wide variety of further propositions about how the pilot would react to various counterfactual circumstances, these background belonging to the holistic theory.

It is worth noting that in both everyday explanations and in the Fieldian one just imagined, an explanation can achieve a wide counterfactual spread without itself including many propositions from the background theory. In both cases, the explanation does not explicitly describe how the pilot would react to various alternative circumstances. Indeed, perhaps that would simply fill the explanation with clutter.[36] Rather, for a wide spread to be achieved, it suffices that some of the propositions in the explanans figure in many further propositions that form part of a holistic theory.

I have claimed that everyday intentional explanations enjoy a distinctive, wide counterfactual spread. But exactly what modal information do they provide? This can be brought out nicely by contrasting the normal explanation and the aberrant explanation in the pilot case. Suppose, for example, that while landing, a pilot tokens a "representation" *r*, which on a normal attribution of truth-conditions has the truth-condition that the pilot's speed is approximately *n* knots and the altitude is approximately *m* meters, where a speed of *n* knots is too fast for landing that plane from *m* meters. In accordance with (Behavioral Effects) and (Relation to Airspeeds), *r* tends to cause the pilot to make a bodily motion that slows the plane. Suppose also that according to the delusion Field sketches, *r* has the truth-condition that U.S. ground troops are being attacked from the west, and suppose that any pilot with this delusion thinks that the bodily motion that is typically caused by tokening *r* amounts to an order to the effect that the troops make a certain defensive maneuver. Now, if we attribute this delusion to the pilot, what answer

[36] This remark raises a question: if detailed modal information about the subject is irrelevant for explaining the pilot's capacity, then how could the failure to provide such information count against Field's non-intentional explanation? As I will soon explain in the main text, the information contained in the counterfactual spreads of everyday intentional explanations is useful in everyday life, the humanities, and the social sciences, for at least the reason that it helps us predict how the agent might likely react to various circumstances. So, at least when prediction is a going concern, one should resist the temptation to say that this information is completely irrelevant to psychological explanation. At the same time, however, it does seem plausible that an explicit spelling out of this modal information would count as clutter.

do we get to the question, "Imagine the president were to announce that enemies have begun *pretending* to attack U.S. ground troops from the west, but that these appearances were certainly not to be trusted. Would tokening *r* still cause the pilot to slow the plane?" The answer is 'no' for the situation as so-far described. And it remains 'no' unless we meticulously tweak the delusion, inventing some reason why the pilot wouldn't believe the announcement. By contrast, if we attribute to *r* the truth-condition that the airspeed is approximately *n* knots, we get no such answer; other things being equal, a sane pilot would not inhibit the belief that her airspeed is approximately *n* knots from causing her to slow the plane just because she heard an announcement about U.S. ground troops. Rather, for such a pilot this announcement would simply be irrelevant to how her beliefs about her speed affected her control of the plane. So, if indeed a pilot would react in a normal way to such announcements, the delusion-attribution fails to entail this unless it is meticulously tweaked, whereas the belief-attribution entails it without any tweaks.

Similarly, if we attribute the delusion, what answer do we get to the question, "Suppose the pilot were to hear a service announcement over the radio, to the effect that speedometers in that type of plane were defective, readings being typically 100 knots faster than the true speed. In that case, would seeing a speedometer reading of *n* knots still cause *r*?" We get a bad answer unless we tweak the delusion in some quite specific ways, for instance, by stipulating that—and how—the pilot interprets service announcements as code-talk for events taking place in the battle on the ground.[37] Unless the pilot has some such specific delusion, such a service announcement wouldn't change the typical causes of *r* in a delusional pilot.[38] By contrast, a sane pilot would

[37] I owe this observation to Richard Boyd.

[38] Or, at any rate, even if there is such a delusion when we only look at reactions to service announcements, it is highly implausible that there is a delusion so thoroughly tweaked that a pilot thus deluded would behave in all the ways predicted by the attribution of beliefs about airspeeds. See Section 2.4.1 below.

react to such an announcement by inhibiting a speedometer reading of *n* knots from causing *r*—that is, by inhibiting such a reading from causing her to believe that her speed is approximately *n* knots. Again, unless the delusion is tortuously tweaked, only the normal belief-attribution explains why the pilot would react in some such way. So, it is just false that the normal belief-attribution could be replaced by a suitably tweaked, radically aberrant attribution, with no loss in the quality of the explanation.

Returning to the big picture, the point is that to describe the likes of (Behavioral Effects) and (Relation to Airspeeds) is not to give an explanation that is informative in all the same ways as the original intentional one. Everyday intentional explanations have, and are supposed to have, a certain counterfactual spread; they provide certain modal information about the agents whose behaviors they explain. This information is useful in everyday life, the humanities, and some social sciences, because it helps us anticipate people's behaviors in a wide range of different circumstances. To replace everyday intentional explanations without any loss of explanatory goodness, the skeptic must come up with non-intentional explanations whose counterfactual spreads are comparable to those of everyday intentional explanations. In the pilot case, an equally good non-intentional explanation would have to connect to a background theory which described a variety of facts about the pilot that would be more complex than the likes of (Behavioral Effects) and (Relation to Airspeeds). This theory would at least have to shed some light on how those of the pilot's "representations" that are implicated in landing the plane are in turn linked to some of her dispositions to react to radio announcements. Arguably, it also would have to say something about the circumstances under which these radio announcements would have different effects. E.g., if the pilot believed she were landing the plane in a battle then she might indeed respond to announcements about enemy troops. Furthermore, the theory would

need to go beyond describing the pilot's dispositions to respond to radio announcements; it would need to describe any other dispositions about which the ascription of beliefs about her airspeeds provides information.

But now suppose that the skeptic could give a non-intentional explanation that provided similar information about these dispositions, by employing explanantia that figured in a holistic theory applicable to the pilot. Then the anti-skeptic could respond by claiming that the skeptic's explanantia form a reduction base for some intentional phenomena. For instance, a simple version of this kind of response would claim that having "representations" that are describable by (Behavioral Effects) and (Relation to Airspeeds) *and* are such that i) the effects of tokening these "representations" would be uninfluenced by announcements about enemy troop movements, and ii) the effects of tokening them would be changed by the service announcement, is a reduction base for having beliefs about one's airspeeds. If that claim turned out to be true, then the skeptic's explanation would turn out to be an intentional explanation, at least as the latter is construed by the metaphysical conception of communicative explanation.

2.4. Replies from the Skeptic and Anti-Skeptical Responses

2.4.1. Tweaking Again

In a moment, I will consider how the skeptic might reply to my remarks about counterfactual spread by tweaking her aberrant intentional explanation. But let us first pause a moment and remind ourselves of the skeptic's argumentative strategy and the role that tweaking plays in it. So far, the skeptic's tactic has been to try to produce an intentional description of a deluded pilot whose ability to land the plane would be ontically explained by non-intentional facts equivalent to those that ground the ability of a sane pilot, this showing that the non-

intentional facts which ontically explain both pilots' abilities do not reduce any intentional facts. The skeptic tinkers with—*tweaks*—her description of the deluded pilot, to make it compatible with non-intentional facts that ontically explain the ability of a sane pilot. Equivalently, the skeptic can be seen as trying to concoct a competing, equally good intentional explanation of an apparently sane pilot's behaviors and abilities, an explanation that depicts this pilot as severely deluded. Under that description of the skeptic's strategy, tweaking is a matter of modifying the competing, aberrant explanation to make it as good as the non-aberrant one.

It is clear how a skeptic who persists with this kind of approach would reply to my remarks about counterfactual spread: she would say it is possible for a pilot with a suitably tweaked, extreme delusion to satisfy (Behavioral Effects) and (Relation to Airspeeds) and have *all* the behavioral dispositions about which the attribution of normal beliefs informs us. (Equivalently, she would say that the behaviors and capacities of a sane pilot can be equally well explained by attributing a suitably tweaked, extreme delusion, one that is consistent with the pilot's dispositions to react to various counterfactual scenarios.) Let's say that such a suitably-tweaked delusion is *specifically tailored* to these dispositions. If indeed there is an extreme delusion specifically tailored to the dispositions about which the attribution of normal beliefs informs us, then the skeptic can insist that the non-intentional facts underlying these dispositions fail to reduce any intentional facts. So, non-intentional explanations that mimic intentional explanations in providing information about these dispositions need not appeal to the reduction bases for any intentional facts, and so can count as non-intentional explanations.

To assess the plausibility of these claims, let us remind ourselves what they amount to. Imagine a pilot who is able to land a plane reliably, and for whom we would normally explain this by saying that she has various true beliefs about her airspeeds. The skeptic claims that the

facts about the pilot that (ontically) explain her ability and that connect it with her dispositions to react to various counterfactual scenarios do very little to constrain the subject matter of her beliefs—that is, these facts do not come anywhere near to determining that her beliefs are about her airspeeds. (Compare: if these facts are compatible with the pilot's beliefs' being about her airspeeds times 0.999, then the skeptic is not vindicated. To make Field's point, the delusion has to be an extreme one.) In particular, for example, she claims that all of these facts could hold for an extremely delusional pilot who has beliefs about U.S. ground troops.

But this is simply unbelievable. If the skeptic's claim were merely that for all the dispositions in some small range, (e.g., dispositions to react to radio announcements,) there could be an extreme delusion specifically tailored to these, then perhaps that could be granted; though I think even this rather unlikely. But, if the relationship between beliefs and behaviors is anything like what we normally take it to be, then it is just utterly implausible that there could be an extreme delusion specifically tailored to all of the dispositions about which the attribution of beliefs about airspeeds informs us. Rather: if someone is able to land planes because she has true beliefs about how fast she is going, then it is overwhelmingly plausible that there be some circumstances in which she would be disposed to behave differently from someone who is able to land planes because she believes she is controlling U.S. ground troops. Our beliefs constrain our behaviors in at least this minimal way.

Admittedly, in making this claim about how our beliefs constrain our behaviors, I am relying on our ordinary understanding of beliefs. This understanding is not beyond dispute; in principle, a skeptic could respond to the foregoing by denying that beliefs are connected with behavior in anything like the way we ordinarily take them to be. However, surely the burden of proof would be on the skeptic here. Our ordinary understanding of belief, including of the ways

it is connected with behavior, has been honed continuously over millennia of everyday use.[39] It would be quite surprising if it turned out to be wrong in as radical a way as what the skeptic suggests.

The two most important points here are as follows. Firstly, just because a belief's role in the production of behavior is always mediated by various desires, traits of character, and other beliefs, it does not follow that a given belief can, if suitably supplemented by these other factors, give rise to any pattern of behavior whatsoever; rather, our beliefs place some constraints on our behavioral dispositions. Secondly, for this reason, it is not the case that for any belief, attributing it can, if suitably supplemented by attribution of desires, traits of character, and other beliefs, serve to (communicatively) explain any behavior or capacity whatsoever.

While these points most obviously concern our everyday explanations, they are also potentially relevant to explanation in a scientific psychology. As we saw, in the pilot case, a virtue of the ordinary intentional explanation is its counterfactual spread: it provides certain information about some of the pilot's dispositions to react to various counterfactual situations. So, if there are scientific contexts in which this kind of modal informativeness is important, then in these contexts, a non-intentional explanation that can apply equally well to a sane and to a delusional pilot would be no substitute for the everyday explanation that attributes beliefs about airspeeds. Now of course, for all that I have said, there may be a variety of non-intentional explanation that shares the counterfactual spread of everyday ones; but then these non-intentional explanations will not be applicable to wide a range of agents whose beliefs have wildly different subject matters. For such explanations, if someday they are produced, there will be a better case for the claim that they describe reduction bases for some intentional facts than there was for

[39] To mention just two out of many ancient sources, strongly belief-like notions can be found in Plato's *Theaetetus* and the Hebrew Bible.

Field's explanation that invokes only (Behavioral Effects) and (Relation to Airspeeds) in the absence of a holistic background theory.

2.4.2. A Different Skeptical Reply

In the "delusion-tweaking" line of argument just considered, the skeptic argues that there are some explanations that enjoy the counterfactual spread of everyday intentional explanations but do not appeal to any intentional properties, or to the reduction bases for any such properties. In response, I argued that this skeptic illicitly stretched the concept of belief to the breaking point: Field's claim that a radically delusional pilot could have the same dispositions as a sane pilot (and thus that the facts relevant to explaining the pilot's ability to land the plane do not reduce any intentional facts) ignores the significant constraints that our beliefs impose on our behavior.

However, in personal communication, Field has described an alternative approach that a skeptic might take. Instead of stretching the concept of belief to accommodate ever more intricately-deluded pilots, she offers a new, unsympathetic account of everyday intentional explanations. On this account, these explanations are *second-class explanations*—that is, they involve tacit reference to the person giving the explanation (see below for a full definition). But, the skeptic insists, second-class explanations would have no place in a truly scientific psychology.

To understand Field's line of thought in full detail, I will begin Section 2.4.2.1 by defining 'second-class explanation'. Then I will present a representative passage in which Field claims everyday intentional explanations to be second-class. After that, I will discuss some reasons why one might deem second-class explanations unsuitable for science. Finally, I will

present (2.4.2.2 - 2.4.2.3) and evaluate (2.4.2.4 - 2.4.2.6) Field's reasons for taking everyday intentional explanations to be second-class.

2.4.2.1. Background: Second-class Explanations and Science

Define *first-class explanations* to be those that proceed via context-insensitive reference to (or, if one prefers, attribution of) some explanatorily-relevant properties. By contrast, *second-class explanations* operate by effecting a rough comparison to the person giving the explanation, this comparison serving to fix the reference of some of the expressions used on some relevant properties.[40] It need not be fully determinate to which properties reference is fixed (see Field's remarks later in this subsection), but presumably for the explanation to be useful, there must be some non-trivial degree of determinacy.

In second-class psychological explanations, the comparison being made is between the speaker and the agent whose behaviors or capacities are being explained. The objective property on which reference is fixed is, or is similar to, an objective property that the speaker herself instantiates; that explains how comparison to the speaker can succeed in fixing reference to this property. For example, on one straightforward version of how everyday intentional explanations might be second-class explanations, in any explanatory context in which the predicate 'believes that snow is white' is applied to an agent, the use of this predicate triggers a comparison to the speaker, which in turn fixes the reference (relative to that context) of the predicate 'believes that snow is white' on an objective property shared by the speaker and the agent.

[40] I am grateful to Harold Hodes for mentioning the idea of reference-fixing in this setting. Field makes no explicit mention of reference-fixing, but it is a plausible interpretation of passages like the one immediately below in the main text. Moreover, after several attempts to understand second-class explanations in alternative ways, I have come to the conclusion that this is the only plausible account of how they operate.

Field endorses something closely akin to this account of belief-attributions in a number of remarks throughout his work since the 1980s. For instance, he writes:

> When explaining a person's behavior (say the raising of his gun) in terms of his belief that there is a rabbit nearby, what I am in effect doing is explaining the behavior in terms of his believing* a representation that plays a role in his psychology rather similar to the role that 'There are rabbits nearby' (or the mental representation associated with it) plays in mine…Such an explanation is still basically non-intentional: truth conditions play no real explanatory role. Of course, there is a sense in which my sentence 'There are rabbits' plays an explanatory role here: obviously not as a causal factor in the explanation, but as a device we use in picking out the agent's internal representation (which is a causal factor). (Field 2001, p.78)[41]

Here belief* is a non-intentional relation that an agent can bear to an internal "representation". For purposes of illustration, it will be harmless to identify believing* a "representation" with tokening it.

There is room for disagreement about what Field means by 'picking out', but one straightforward possibility is that in any explanatory context C, the use of the predicate 'believes that there is a rabbit nearby' interacts with the explanatory interests at hand to make salient some role X that a "representation" can play in an agent's psychology, a role possessed, in particular, by some of the speaker's own "representations". This fixes the reference, in C, of the predicate 'believes that there is a rabbit nearby' on the property *believing* a "representation" which plays role* X.

Now that we have a clear account of what second-class explanations are, we are in a position to see why one might be uneasy about their suitability for science,[42] or at least for an

[41] For similar remarks, also (Field 1978), p.47, (Field 1986) p.82, and (Field 2001) p.78 and p.155. This picture of belief-attribution echoes a similar one from (Quine 1960) p.219, and both philosophers display affinity for the *simulation theory of folk psychology*, prominently stated in (Gordon 1986) and cited in both (Field 1986) p.87 and (Field 2001) p.78. According to the simulation theory, in folk psychology "we represent the mental states and processes of others by mentally simulating them" (Gordon 2009). For example, we predict others' behavior by putting ourselves in their shoes and asking how we would act, rather than by deducing propositions about how they will behave from generalizations that relate beliefs, desires, and traits of character to behaviors.

[42] That said, it is worth noting that second-class explanations will count as genuine explanations under the

ideal science. The most obvious concern has to do with indexicality. Arguably, a distinctive feature of science is a kind of independence from the inquirer of the product of her inquiry, an independence which is hampered by indexical reference. For example, if one tries to fix reference to a temperature property by saying 'hot to the touch', the inquirer-relativity of that concept makes an explanation that employs that phrase unscientific. That is because another inquirer might differ in her judgments about what is hot to the touch, if (e.g.,) the first just had his hands in front of a fire, whereas the second just had her hands in ice-water. For this reason, one might reasonably think, good science avoids indexical reference, e.g., by preferring measurement (50 degrees Fahrenheit) to qualitative descriptions (hot). This consideration counts against the inclusion of second-class explanations in an ideal science.

In addition to their reliance on comparison to the explanation-giver, second-class explanations might be unsuitable for science in virtue of being affected by contextual shifts in the relevance of different respects of comparison. Consider, for example, the attribution to an ancient Greek of the belief that there is a thunderstorm nearby.[43] If all we care about is the Greek's ability to keep safe in storms, then the fact that a state leads the Greek to seek shelter may license the description of it as being a belief that there is a thunderstorm nearby. But in contexts in which we also care about explaining her attempts to control the weather by offering sacrifices to (what she takes to be) the gods, it might be better to describe her as having the belief that Zeus is hurling thunderbolts rather than as believing that it is thundering. For our belief that it is thundering is dissimilar to the Greek's state, in that it does not lead us to make sacrifices; so, in this context, likening the Greek to ourselves will be misleading. Again, the point is that if

metaphysical conception of explanation, whenever the properties whose attribution they involve ontically explain their (the explanations') explananda.

[43] This example is due to (Field 2001), p.80-81.

indeed everyday intentional explanations are second-class, then the appropriateness of any given belief-attribution is liable to vary in response to contextual shifts in what we are out to explain. As with indexicality, one might take this feature of context-shiftiness to make belief-attributions inappropriate for purposes of science.

Finally, the reference-fixing involved in second-class explanations may be vulnerable to indeterminacy. Two plausible sources of indeterminacy are the indexicality and sensitivity to the explanatory interests at hand that these explanations involve. Indeed, for example, Field takes the reference-fixing (allegedly) involved in everyday intentional explanations to be indeterminate:

> A natural view is that when we use representational concepts or properties[44] to explain facts described in nonrepresentational terms, the representational concepts or properties just code for conceptual or functional role properties: we specify the functional role property by specifying the representational property. This is certainly very plausible, if we don't take it as committed to there being any very uniform account of what conceptual role properties a representational property codes for, *and if we don't suppose that even on a given occasion there is a very precise conceptual role property that is coded for*. (Field 2001, p.77, italics mine)

The point, again, is that one might reasonably take indeterminate reference to make those second-class explanations that suffer from it unsuitable for an ideal science.

By now, we have seen Field claiming that everyday intentional explanations are second-class, and we have seen some reasons why an ideal science might not include any second-class explanations. Still, what reasons do we have to take everyday intentional explanations to be second-class? As I will explain in the next subsection, Field arrives at this conclusion via an inference to the best explanation.[45]

---

[44] Field's use of 'representational concepts or properties' rather than simply 'belief and desire' shows that his claim concerns not only everyday intentional explanations but all propositional explanations whatsoever, and, indeed, perhaps all intentional explanations, insofar as, for instance, reference by words and concepts, which is non-propositional, counts as a kind of representation.

[45] To my knowledge, Field never explicitly makes this inference. Rather, I attribute it to him as part of my best attempt to develop his position, based on his remarks in personal communication.

2.4.2.2. Outline of Field's Inference to the Best Explanation

To explain Field's reasoning, I will need to discuss two important notions. The first we have already seen; it is the notion of a theory's being *holistic*. A theory about minds (or about some minds) is *holistic* if, for a wide variety of cognitive and affective states, the theory describes a wide variety of relations that these states bear to one another, to the state-bearers' external environments (including, perhaps, the states and behaviors of other cognitive agents), and to the state-bearers' behaviors, dispositions, and capacities. As we saw, a psychological explanation can acquire a wide counterfactual spread by attributing states (types) that figure in a holistic background theory.

The other notion I need to discuss is that of a theory's being *widely applicable*. Roughly, the idea here is that a theory is widely applicable if the patterns of explanation that it embodies can be applied to many different phenomena. To introduce the idea more precisely, I will borrow some terminology from (Kitcher 1989), helpfully explained in (Woodward 2017):

> A *schematic sentence* is a sentence in which some of the nonlogical vocabulary has been replaced by dummy letters. [E.g.,] the sentence 'Organisms homozygous for the sickling allele develop sickle cell anemia' is associated with a number of schematic sentences including 'Organisms homozygous for A develop P' and 'For all X if X is O and A then X is P'. *Filling instructions* are directions that specify how to fill in the dummy letters in schematic sentences. For example, filling instructions might tell us to replace 'A' with the name of an allele and 'P' with the name of a phenotypic trait in the first of the above schematic sentences. *Schematic arguments* are sequences of schematic sentences. *Classifications* describe which sentences in schematic arguments are premises and conclusions and what rules of inference are used. *An argument pattern* is an ordered triple consisting of a schematic argument, a set of sets of filling instructions, one for each term of the schematic argument, and a classification of the schematic argument. The more restrictions an argument pattern imposes on the arguments that instantiate it, the more *stringent* it is said to be. Roughly speaking, Kitcher's guiding idea is that explanation is a matter of deriving descriptions of many different phenomena by using as few and as stringent argument patterns as possible over and over again. (Woodward 2017, Section 5.1, italics mine)

In what follows, I will assume that part of the business of theories is to furnish explanations, and that for many theories, there are specific argument patterns that are associated with, and characteristic of, those theories. I will speak of such argument patterns as *belonging to* these theories. An explanation *belongs to* a theory if it instantiates an argument pattern that belongs to the theory and uses terminology that is characteristic of (or, as I will also sometimes say, *belongs to*) the theory.[46] In what follows, I will assume that everyday intentional explanations instantiate argument patterns that are distinctive of folk psychology. E.g., the schematic sentences that these patterns involve might be ones such as 'For all X, Q, and W, if X desires Q and believes that doing W will bring about Q then X does W'.

The *range of application* of a theory is the range of phenomena descriptions of which can be derived from the theory's argument patterns. A theory is thus *widely (narrowly)* applicable if the argument patterns that belong to it can (respectively, cannot) be used to derive descriptions of many different phenomena. Similarly, an explanation that forms part of a theory is *widely (narrowly) generalizable* if the argument pattern(s) that it instantiates can (respectively, cannot) be used to derive descriptions of a wide range of phenomena in addition to the explanandum. An explanation of the behavior or capacities of one agent can be *generalized* to a different agent if the argument pattern(s) that the explanation instantiates can be used to derive a description of the other agent's behavior or capacities.

Using the notions of holism and wide applicability, I can now explain how Field arrives at his claim that everyday intentional explanations are second-class. The idea is that although pre-theoretically, folk psychology appears to be quite holistic and also quite widely applicable, it

[46] We need not assume that in general the terminology that is characteristic of a theory is implicitly defined by its role in that theory. E.g., perhaps the terminology of a genetic theory is ultimately defined not by any theory in the discipline of genetics but rather in some theories of organic chemistry.

is in fact indeterminate on both counts. Moreover, as one might expect, this indeterminacy infects everyday intentional explanations. When we learn that a pilot has beliefs about her airspeeds, it is indeterminate what we learn about her dispositions to react to various alternative circumstances, and which the circumstances are such that we learn something about how the pilot would react in them. And for any everyday intentional explanation, it is indeterminate to which other agents the explanation can be generalized. That is, fixing the folk-psychological argument pattern(s) that this explanation instantiates, it is indeterminate for which other agents this pattern can be used to derive descriptions of their behavior.

In turn, Field's idea is that it makes sense that the counterfactual spread and generalizability of everyday intentional explanations would be indeterminate, if indeed giving such an explanation consists in inviting a comparison with oneself in unspecified but contextually salient respects. As we saw, second-class explanations are vulnerable to an indeterminacy as to which objective property is attributed.[47] Thus, Field arrives at the claim that everyday intentional explanations are second-class by an inference to the best explanation: this is the hypothesis that best explains their indeterminacy in the two respects just described.

By contrast, on Field's view, the background theories that would be employed in a truly scientific psychology would not be indeterminate, either with respect to holisticness and the width of the counterfactual spread of their explanations, or with respect to their range of applicability and the degree of generalizability of their explanations. However, he asserts, for truly scientific theories, holism and wide applicability are conflicting properties; no such theory

[47] Recall also Field's claim that although talk of beliefs "codes for" some "conceptual role properties," on any particular occasion on which we attribute a belief, it is indeterminate which such property is "coded for." (Put in my terms, Field is claiming that it is indeterminate to what property reference is being fixed by the comparison to the explanation-giver.) If everyday intentional explanations indeed involve indeterminacy in this way, then that would explain the (alleged) indeterminacy as to their counterfactual spread and generalizability.

can have both properties in the proportions that folk psychology purports to. Rather, any truly scientific explanations with the counterfactual spread of folk psychological explanations would be less generalizable than the folk ones; and any truly scientific explanations that were as generalizable as folk psychological explanations would be less holistic than the folk ones.

In the next section (2.4.2.3), I will flesh out these claims by describing in detail the two kinds of background theories that Field thinks would be available for a scientific psychology, and comparing them both to folk psychology. Then in Sections 2.4.2.4 - 2.4.2.6, I will respond to Field's inference to the best explanation.

2.4.2.3. Two Types of Non-Intentional Background Theories

What I will call *type 1 theories* are quite widely applicable. Explanations that are based on type 1 theories provide information about only rather simple properties and relations of the agent's "representations", such as those properties and relations mentioned in (Behavioral Effects) and (Relation to Airspeeds). The regularities invoked in such an explanation, such as those described in (Behavioral Effects) and (Relation to Airspeeds), are ones that could be instanced by the states of many different agents, including many agents to whom we would, pre-theoretically, ascribe radically different beliefs. As Field points out, an explanation along the lines of (Behavioral Effects) and (Relation to Airspeeds) could apply to both a sane pilot and a delusional pilot. Moreover, perhaps some argument patterns in some type 1 theories could even be applied both to genuine cognitive agents and also to some non-cognitive but still minimally environmentally-responsive systems, such as thermostats. After all, just like the representations in (Behavioral Effects) and (Relation to Airspeeds), the states of the bimetal strip in a thermostat

can systematically trigger "appropriate"[48] behaviors in response to changing environmental conditions.[49]

At the same time, however, a type 1 theory about, for example, pilots, would provide little information about how the "representations" implicated in landing the plane are related to the agent's other "representations", or to those of other individuals in the pilot's community. Thus, any explanation based on a type 1 theory would have a quite limited counterfactual spread, shedding little light on how the agent would react to a variety of other possible circumstances. For this reason, by themselves, the facts invoked in explanations that are based on type 1 theories are unsuitable to serve as reduction bases for any intentional facts, familiar or otherwise. Accordingly, as I granted in Section 2.4.1, the facts described in (Behavioral Effects) and (Relation to Airspeeds), taken by themselves, do not reduce any intentional facts.

So much for type 1 theories. *Type 2 theories* differ from type 1 theories in that they are quite holistic. For instance, when it comes to agents like Field's pilot, a type 2 theory would describe many quite complex relations in which the agent's "representations" stand. These would include relations that connect the "representations" which are directly involved in landing the plane (that is, $b_0,\ldots b_n$) with the representations that would be tokened in response to various radio announcements, as well as a whole host of other circumstances. Explanations based on type 2 theories have wide counterfactual spreads, comparable to the spreads of everyday intentional explanations.

[48] Of course, these behaviors are not "appropriate" in the sense that they satisfy desires or intentions that the thermostat has, but rather only in the sense that they tend to give rise to the capacity being explained (namely, the capacity to bring and maintain the local environment at a specified temperature).
[49] A simple thermostat contains a bimetal strip which bends to different degrees of bent-ness in response to different ambient temperatures. When the strip is bent to a certain degree of bent-ness (set by the control knob), it closes an electronic circuit that triggers a furnace.

However—and this is Field's crucial point—type 2 theories are not applicable to a wide range of different agents in the way that folk psychology is supposed to be. For example, pre-theoretically, it appears that the attribution of beliefs about airspeeds could adequately explain the landing capacities of both a Democrat pilot and a Republican pilot, of both a Sikh pilot and a Zoroastrian pilot, of both a Dutch pilot and a South Korean pilot, etc., despite the differences in these pilots' other beliefs. The explanation is thus widely applicable in at least these ways. But the same does not hold of type 2 theories. Any type 2 theory describes state-state, state-environment, and state-behavior relationships that are so specific and detailed that they can only be instanced by the states of highly similar cognitive agents, falling within a narrow range. Indeed, perhaps each type 2 theory applies only to a single agent at a single point in her life. As we have with our familiar folk-psychological argument patterns, few if any argument patterns belonging to a type 2 theory could apply both to a sane pilot with beliefs about her airspeeds and also to a severely deluded pilot with beliefs about U.S. ground troops. But the range of application of the argument patterns belonging to type 2 theories is much narrower than that. *Unlike* with folk psychology, for instance, no argument pattern from a type 2 theory could be used to derive descriptions of both a Westernized 21st century person and an ancient Greek, even if in many circumstances, we would ordinarily want to describe both people as having the belief that it is thundering. Field's idea is that the Greek's psychology is sufficiently different enough from our own that no reasonably detailed scientific theory could apply to both.

Folk psychology appears to occupy a middle ground between type 1 theories and type 2 theories, enjoying the virtues of both. Its argument patterns can be applied to a wide variety of different agents (Dutch, South Korean, Sikh, Zoroastrian, etc.), yet it also manages to be quite holistic, giving rise to explanations which have wide counterfactual spreads.

However, Field's view (personal communication) is that folk psychology is in fact indeterminate on both counts. Any everyday intentional explanation is indeterminate, both as to the extent of its counterfactual spread and as to its range of generalizability. And on Field's view, the best way to explain this indeterminacy is to hold that everyday intentional "explanations" are second-class.[50] Moreover, he thinks, no theory that employs only first-class explanations can achieve the balance between holistic-ness and wide applicability that folk psychology (misleadingly) appears to achieve; and likewise, no variety of first-class explanations can achieve the balance between wide counterfactual spread and wide generalizability that everyday intentional explanations (misleadingly) appear to achieve.

2.4.2.4. First Response: Middle Ground and the Metaphysical Conception

As an immediate reaction to the foregoing, it is worth making the following observations. On the metaphysical conception of communicative explanation, most of an explanation's virtues derive from how well it serves as an accurate depiction of, or guide to, what ontically explains the explanandum. Some virtues, such as brevity, simplicity, and accessibility to non-specialists, might be exceptions; but all of these have obviously to do with ease of communication. Most other virtues, by contrast, do, on the metaphysical conception, owe themselves to the explanation's being, in one way or another, tightly connected to what ontically explains its explanandum.

This view about where explanations get their virtues affects the picture of second-class explanation that emerges under the metaphysical conception. In particular, on this conception, it is plausible that when a second-class explanation has a wide counterfactual spread, it owes this to

[50] See (Stich 1983), and (Churchland 1981) for similar views.

the properties being attributed, those on which reference is fixed by the comparison to the explanation-giver. Similarly, on the metaphysical conception, it is plausible that if a second-class explanation is widely generalizable—that is, if the argument pattern(s) that it instantiates can be used to derive descriptions of many different agents—then it owes this wide generalizability to the properties being attributed. That, again, is because wide counterfactual spread and wide generalizability are virtues that have to do with informativeness and not merely with ease of communication.[51]

However, given these consequences of the metaphysical conception, the fact that folk psychology appears to be both holistic and widely applicable should count as compelling evidence that there is some scientific theory that is both holistic and widely applicable in much the same ways, and whose explanations (non-second-class, non-indexical, context-insensitive ones, we are assuming,) have wide counterfactual spreads and ranges of applicability, comparable to those of everyday intentional explanations. Otherwise, the appearance that folk psychology and its explanations possess these virtues, and our ability to employ folk psychology, with great success, as though it possessed these virtues, would be quite mysterious. But now notice that if indeed there is a scientific theory as just described, then the door is opened for the claim that some of the explanantia of the explanations belonging to this theory reduce some intentional facts.

All that said, however, if folk psychology really is indeterminate with respect to its degree of holism and range of applicability, and if no theory that trafficked in first-class

[51] This is easier to see in the case of counterfactual spread than it is in the case of generalizability. I take it that an explanation's degree of generalizability bears on its informativeness in that it bears on the degree to which the explanation is informative about other agents.

explanations would be similarly indeterminate,[52] then it is doubtful that there can be a scientific theory that furnishes us with materials for reducing any intentional facts. So, we must carefully consider the plausibility of Field's allegations of indeterminacy.

2.4.2.5. Second Response: The Impression of Indeterminacy is Illusory

When it comes to holism, I suspect that Field's impression of indeterminacy is illusory. It is a misinterpretation of the fact that, taken by themselves, attributions of single beliefs are often insufficient to pin down exactly how the agent under consideration would react to a variety of alternative scenarios. This in turn stems from the fact that individual beliefs do not produce behaviors on their own, but rather in concert with other beliefs, desires, and traits of character. For the sake of simplicity, I left many of these factors implicit when discussing Field's pilot example. For example, when I suggested that in hearing that the pilot has beliefs about her airspeed we also learn how she would react to certain radio announcements, I tacitly held fixed that the pilot wants to land the plane rather than crash. Obviously, a pilot who wanted to crash might react differently.

However, I hasten to add that just because a belief's role in the production of behavior is always mediated by various desires, traits of character, and other beliefs, it does not follow that a given belief can, if suitably supplemented by these other factors, give rise to any pattern of behavior whatsoever. For this reason, it does not follow that for any belief, attributing it can, if suitably supplemented by attribution of desires, traits of character, and other beliefs, serve to explain any behavior or capacity whatsoever. That was the point of my remarks at the end of Section 2.4.1; I will not repeat the argument here.

[52] Field simply assumes the second conjunct of this if-clause. Its plausibility is worth examining, although there is no space to do so here.

More to the present point, just because the effects of beliefs on behavior are mediated by various other factors, it does not follow that it is indeterminate what information attributions of individual beliefs provide about a subject's dispositions to react to a wide range of alternative scenarios. Surely there are patterns that govern how sets of beliefs, desires, and personality traits interact to produce behaviors; otherwise the attribution of beliefs, desires, and personality traits would not have survived. And the fact that a subject has a given single belief relates her to those patterns in a definite way. My working hypothesis that these patterns are describable as *ceteris paribus* generalizations provides a clear model of how this might go: any given belief will feature in the if-clauses of a wide variety of such generalizations, and so the attribution of a single belief allows us to apply any such generalization to the subject, conditional on the rest of the if-clause.

Here it will help to compare a similar example. Rain is caused by the condensation of water in the upper atmosphere. This condensation is influenced by many different factors, including temperature, pressure, air currents, and levels of humidity, in accordance with patterns that are, I take it, enormously complex and not yet fully understood—hence our quite-imperfect ability to predict the weather. If one describes the temperature in a given location at a given time, that alone may not give one's interlocutor enough information to explain why it would rain at that location under various counterfactual conditions which hold that temperature fixed. Nonetheless, however, the information that one provides does relate the case at hand to those patterns in a definite way: relative to the case at hand, it fixes the value of one of the variables that those generalizations relate to one another. This allows the interlocutor to make some inferences about what would happen in the scenarios described by those generalizations.

Now for concerns about indeterminacy in everyday intentional explanations' degree of generalizability. A first point to be made here is that insofar as these concerns emanate from concerns about indeterminacy of counterfactual spread, there is no cause for concern. On this way of motivating the generalizability worry, if it is indeterminate what an explanation tells us about how the agent would react to various alternative circumstances, then it is therefore also indeterminate to what extent the argument pattern the explanation instantiates can be used to derive descriptions of other agents, who may have different dispositions. However, I have just been arguing at length against Field's concerns about indeterminacy when it comes to the counterfactual spread of everyday intentional explanations. If those concerns are ill-founded, then they cannot generate any further concerns about range of applicability. All the same points can be put in terms of theories, holism, and range of applicability, rather than in terms of explanations, counterfactual spread, and generalizability.

Still, in practice it remains true that there are many agents to whom we are uncertain whether or not folk psychology can be applied. Some minimally-cognitive agents, such as cats, might be one example; another is cognitively-impaired humans, such as those who suffer significant amounts of dementia.[53] In such cases, the skeptic can be expected to insist that the best explanation of our uncertainty is that there is no fact of the matter—indeterminacy—whether or not folk psychology applies.

In response, a fervent anti-skeptic could insist that all that is needed to resolve our uncertainty is to learn more about these agents; folk psychology as we know it is precise enough to give a definite verdict, once we know enough about those to whom we would apply it. However, when taken fully generally, this is an extreme position; it is plausible that for at least

[53] See (Stich 1983) p.54-56 for exposition of a similar problem.

some kinds of agents, we may have to make at least some modifications to the concepts and generalizations of folk psychology in order to have a definite verdict on whether or not it applies to agents of that kind. The issue, in that case, is whether these modifications would count as genuine theory change and replacement of our folk concepts, or simply a refinement of the existing theory and the existing concepts. (Though note that the changes would not increase the plausibility of (Elim) unless the new concepts were non-intentional.) This complicated issue remains to be resolved. There is no space to resolve it here; but I see no reason to assume in the meantime that it will be resolved as the skeptic hopes.

2.4.2.6. Third Response: Other Intentional Properties

Now that I have brought up the matter of intentional explanation outside of folk psychology, let me briefly mention a different line of response to Field's claims about everyday intentional explanations. These explanations are a reasonable target for skeptics, since they are the most familiar and uncontroversial example of intentional explanations. Thus, the debate so far has centered around the possibility of first-class explanations that emulate everyday explanations, thereby having a claim to being intentional in fact if not in name. However, for all Field says, the facts described in type 2 psychological theories (holistic theories that are only narrowly applicable) might reduce some hitherto-unrecognized intentional facts. On this view, the unavailability of any scientific theory that is both holistic and widely applicable impugns only the intentional phenomena familiar from folk psychology, not all intentionality whatsoever; for it is only folk psychology that is required to be widely applicable.

To see the point, recall the example of the ancient Greek who is able to keep herself safe in thunderstorms. Suppose that we try to explain this ability by saying that when there is a

thunderstorm where the Greek is located, she believes that it is thundering. As we saw, Field thinks that such a description might not be fully satisfactory.[54] For example, taken by itself, it might fail to explain the Greek's attempts to control the weather by offering sacrifices to (what she takes to be) the gods. For that purpose, it might be more useful to describe her as having the belief that Zeus is hurling thunderbolts, rather than as believing that it is thundering. In this way, the explanation that ascribes the belief that it is thundering might not apply to the ancient Greek in the way that it applies to 21st century Westerners; there might be facts about the Greek that it doesn't explain, and in addition it might suggest things about the Greek that are only true of 21st century Westerners. To explain the Greek's behaviors, the idea is, we would need a quite holistic theory that restricts itself to agents who are very similar to the Greek.

Now as it happens, I think folk psychology does just fine in the case of the Greek. We can safely attribute the same content to both us and the ancient Greeks, as long as we posit lots of differences in collateral beliefs. For example, we both believe that it is thundering, but the ancient Greek also believes that Zeus is throwing thunderbolts, which we don't believe. Given that behavior is generated by collections of beliefs (strung together by inference, prior to action, for example), it's not surprising that members of the two groups behave differently, even though they share some of their beliefs.[55]

However, assuming for the sake of argument that folk psychological intentional notions are less applicable to the Greek than they are for 21st century Westerners, there nonetheless does seem to be some sense in which the Greek's mental state is about thunder, and also, for that matter, in some sense about Zeus. In that case, perhaps what we need is to posit intentional states of some other, hitherto-unrecognized kinds, instantiable not by both 21st century Westerners and

[54] See (Field 2001), p.80-81.
[55] Thanks to Robert Rupert for bringing this point to my attention.

ancient Greeks, but only by ancient Greeks. The anti-skeptic's suggestion would then be that the explanantia of a type 2 theory that applied to ancient Greeks would furnish the materials for a reduction of such states.

Of course, what I have provided here is just speculation. It could also turn out not make much sense to associate any subject matter with the Greek's states. The point is merely that until we have learned more about the explanantia in the holistic explanations that Field envisions (recall the second, highly detailed variety of first-class explanations described above), we need not conclude that they fail to reduce any intentional facts. At most what we can conclude is that they fail to reduce any intentional facts of the sorts that are familiar from everyday life. Given a metaphysical conception of explanation, Field's picture leaves open that intentionality has a role to play, in fact if not in name.

2.5. Concluding Remarks on Field

My discussion of Field's argument has three general lessons. The first is that if one succeeds in identifying some non-intentional facts that explain what appeal to some intentional facts is supposed to explain, then it is open to the friend of intentionality to claim that the former facts reduce the latter facts. If correct, that would provide intentional facts with an explanatory role after all, given a metaphysical conception of explanation.[56] As I emphasized in the introduction and as the dialectic of this section illustrates, this train of thought yields a distinctive approach to reducing intentional facts: to figure out what goes in one's reduction, try to identify some non-intentional facts that (ontically) explain what the intentional facts in question are supposed to explain, and in just as effective a way. As one refines one's proposed non-

[56] In fact, I owe this insight to Field himself, who recognizes it throughout his work. See again (Field 1986) p.84 bottom paragraph, (Field 1994) the paragraph straddling p.254 and p.255, and (Field 2001) pp.153-155.

intentional explanations, their explanantia provide the materials for increasingly plausible reduction proposals. A second lesson is that skeptics who endorse a metaphysical conception of explanation should be prepared to address this approach to reducing intentional facts; to make their strongest case, skeptics should address accounts of intentional facts that are developed with an eye toward their (purported) explanatory role. Most importantly, the final lesson is that if a skeptic is to replace our everyday intentional explanations with ones that replicate all of their virtues except brevity, simplicity, and accessibility in everyday life, then she must take care to replicate their characteristic counterfactual spread.[57] This point has obvious implications for attempts to reduce intentional facts of the sorts we posit in everyday life: the reduction base must be such as to give rise to this counterfactual spread.

Clearly, the metaphysical conception of explanation played a highly significant role in this debate. It motivated the idea that non-intentional explanations can serve as a source of materials for attempts to reduce intentionality, thus enabling the anti-skeptic to raise the concern that intentionality was being invoked in fact if not in name.[58] By contrast, in the next section we will see that these moves are unavailable against the backdrop of an epistemic conception of explanation. Indeed, that conception changes the dialectic on both sides.

3. Dynamical Approaches to Cognition

3.1. Background

[57] Of course, a skeptic could simply insist that there are virtues other than counterfactual spread which are more important in science, and for the sake of which it is acceptable to sacrifice counterfactual spread. My point here applies only to skeptics who are interested in retaining the informativeness of everyday intentional explanations. I leave it to scientists to decide how important such informativeness is in their fields.
[58] See above, top p.12, bottom p.17, top p.25.

Implicit in many of the skeptical arguments in the last section is the idea that intentionality is not ontologically respectable, and that an accurate description of the structures and causes that give rise to behavior will always yield better explanations. However, there is some empirical evidence against the second of these two claims.[59] In practice, an explanation that is highly simplified or idealized can often provide a better counterfactual spread than an explanation that more accurately describes the causes of the system's behaviors. Thus, even theorists who take intentionality to be ontologically disreputable can resist the conclusion that non-intentional explanations will therefore be guaranteed to have wider counterfactual spreads than intentional ones. What is more, a realist about intentionality can maintain that while intentionality is ontologically respectable, but that the explanatory virtues of intentional discourse do not depend on its being so; and so, it is for some other reason that many intentional explanations have distinctive, wide counterfactual spreads. Both the realist and the anti-realist positions just described abandon the metaphysical conception of explanation, since they abandon the idea that the quality of an explanation depends on the nature of the phenomena it describes.

All that said, in principle, intentional explanation can be attacked from the standpoint of the epistemic conception of explanation, as well as the metaphysical conception. Although I know of no such attacks, advocates of the *dynamical systems theory approach* to cognition seem

[59] See, for example, (Batterman 2001), (Batterman 2002), and (Rice 2013). One can make a similar point about Field's tweaked intentional explanations. Even if it is possible to tweak any intentional explanation to get it to explain all of a subject's counterfactual behaviors, this will invariably have to be done after these behaviors are already known. To see the point, consider again Field's aberrant belief-attribution in the pilot case. One doesn't know how to tweak the delusion until one knows how a sane pilot would react to radio announcements. By contrast, our everyday attributions of intentional properties can be made in advance of observing the behaviors that would arise in alternative scenarios, giving these attributions predictive power. This speaks in favor of our everyday attributions of intentional properties, and against aberrant stories like Field's: of the two, only the everyday attribution gives predictive power to our explanations. I mention this point here rather than in the previous section because it has to do with the virtues of an explanation's descriptions rather than the nature of the entities it describes, and so only someone with the epistemic conception of explanation need be moved by it. Field, recall, works under the metaphysical conception.

a likely potential source. The working hypothesis of the dynamical approach is that cognitive capacities are properly explained by appealing to appropriate sets of differential equations, typically equations that describe interactions between parts of the brain and its external environment.[60] Some authors take this idea to conflict with views of the mind as a representational or information-processing system.[61] More precisely, *mental representations* are causally effective items internal to the agent that represent objects, relations, or situations in the environment, and so act as intermediaries between the agent and the external world. Dynamicists resist the idea that cognition consists in the manipulation of mental representations; rather, they take cognition to emerge from direct, dynamic interactions between processes in the brain, parts of the body, and the external world.

The reason the dynamicist perspective easily lends itself to (Elim) is that some of the most obvious and familiar candidates for bearers of intentionality, such as beliefs and desires, are most straightforwardly understood as mental representations. However, it is worth noting that the dynamical approach to understanding cognition is not *ipso facto* committed to (Elim), even when it comes to propositional representation (though it *is* committed to eschewing talk of propositional representation*s*). For example, one might deny that believing something amounts to tokening an internal, causally effective state, and instead hold that it amounts to bearing a certain relation something in the external world, in virtue of engaging in various dynamic interactions with it. All that said, as of now, dynamicists have characterized their explanantia in non-

[60] This approach is advocated and developed in (Thelen & Smith 1994), (Van Gelder & Port 1995), (Port 2003), (Kelso 1995), (Chemero 2009), (Silberstein & Chemero 2011), and (Riley, Shockley & Van Orden 2011).
[61] See (Thelen & Smith 1994), (Van Gelder & Port 1995), (Kelso 1995), (Chemero and Silberstein 2008a), (Chemero and Silberstein 2008b), and (Chemero 2009).

intentional terms, in addition to avoiding mention of mental representations. For this reason, it is worth pointing out the phenomena that their research has yet to explain.

Dynamical theorists often leave open or even explicitly concede that many of their explanantia can be accurately characterized in terms of mental representations.[62] However, what they emphasize is that even if these descriptions are accurate, they contribute nothing to the explanations in which they figure. Rather, the best (communicative) psychological explanations are to be framed without any mention of mental representations. And one can easily imagine someone making a similar claim about intentional characterizations. This sort of position is most congenial to the epistemic conception of explanation, since it emphasizes the language in which explanations are framed rather than the nature of the entities they describe. Accordingly, even if an explanation's explanantia are in fact intentional, it can count as a non-intentional explanation as long as the explanantia are not characterized in intentional terms.

In fact, there are further reasons why the dynamical approach to cognition fits especially well with an epistemic conception of explanation. Dynamicists do not operate by describing in detail the mechanisms that physically produce the explanandum, though their accounts can be supplemented with such descriptions. Rather, these scientists model various aspects of cognitive systems with equations that relate some posited, coarsely described mechanisms to various other, further mechanisms, or to various elements in the systems' environments, and thereby shed light on some of the systems' behaviors. (See below for an example.) These equations are particularly useful for making predictions; they show exactly how some features of the system will change, given that others change in specified ways. Thus, dynamicists who take themselves to be providing complete explanations need to hold that modeling a system by a set of equations can

[62] See (Van Gelder and Port 1995) p.2 and (Chemero 2009) pp.67-68.

count as explaining the system's behaviors, as long as it enables one to make a suitable variety of accurate predictions about those behaviors. In particular, they need to hold that presenting such a model can count as giving a complete explanation, even if the model is not supplemented with descriptions of the mechanisms that actually produce the behavior. It is worth noting that this claim is highly controversial, even when restricted to cognitive science rather than taken as a general thesis about explanation.[63] However, properly disputing it is beyond the scope of this paper, so in this section I will grant it for the sake of argument.[64]

Overall, I will have two points to make against dynamicism as a route to (Elim). Firstly, so far, dynamical systems theorists have only shown that some kinds of behaviors and capacities can be explained without invoking representations, not that all of them can. To justify (Elim), dynamical systems theorists would have to give us reason to think they will someday be able to give dynamical, non-intentional explanations of the things that we paradigmatically explain by invoking intentionality. While these scientists have shown intentional explanation to be unnecessary in a surprising variety of cases, a large gap remains between the explananda in these cases and the explananda for which we paradigmatically invoke intentionality. Secondly, although dynamical explanations allow one to make more precise predictions than everyday intentional ones do, so far, the range of these predictions is significantly narrower: the predictions only describe the ways that the parameters that feature in the equations can influence one another, and not how they can be influenced by yet further phenomena. This point is highly

[63] See (Kaplan and Craver 2011) for some compelling objections, and (Chemero 2009) for some responses.
[64] (Hochstein 2012) is friendly to an epistemic, prediction-based conception of explanation, and uses it as a platform for touting the virtues of intentional explanations. However, unlike my main claims in this section, his do not obviously conflict with (Elim). Hochstein emphasizes that intentional explanations are "ideal for situations where statistical and dynamical models are unavailable and/or uninformative. Intentional models allow us to make predictions without having to quantify over features of the system that we may not know how to measure" (p.553). This claim is compatible with intentional explanations being gainfully eliminated as soon as the domain of statistical and dynamical models is sufficiently extended.

reminiscent of one of my criticisms of Field from Section 2.4: unless dynamical theorists can produce explanations that are as modally informative as everyday intentional ones, they will not have justified the elimination of the latter. A pair of explanations will help to bring out these points.

3.2. Syllable Placement vs. Political Decisions

Human speech tends to take on regular, periodic patterns, giving every episode of speech something that (Port 2003) calls a *basic period*. What is more, the stressed phonemes within words tend to occur at certain regular intervals within the basic period of an episode of speech, especially halves and thirds of the basic period (p.599). The task of (Port 2003) is to explain this fact. Port suggests that there are neural oscillations that take place during speech. According to Port, the oscillations attract our attention to the half and third intervals, and also influence the motor system to produce the stressed syllables at these times. Port models the interaction of the neural oscillation with our attention and motor system with a differential equation.

Challenging Port's widely celebrated results is well beyond the scope of this essay as well as my level of expertise. It is also unnecessary for the point I want to make. Suppose Port does a perfectly good job of explaining why we tend to place our stressed syllables on the halves and thirds of our speech periods. The question for us now is, does that support (Elim)? Perhaps some fans of intentionality might have hoped to explain our speech regularities by suggesting that we have some internal items that represent the half and third intervals of our speech periods, and that these items are causally efficacious in guiding the placement of our stressed syllables. That would be an intentional explanation, since it involves saying that these items represent—are about—these intervals. However, assuming that Port's explanation is a good one that strategy is

now closed off. Importantly, though, one must avoid getting carried away. The explanandum here is one out of very many for which we might hope to invoke intentionality, and not a paradigmatic one at that. It is not at all clear that the general unsuitability of intentional explanations for cognitive science can be extrapolated from this case. (To be fair to Port, that is not what he is out to do in the paper.)

The point is especially easy to see when we return our attention to everyday psychological explanations. The kinds of explananda for which we typically invoke familiar intentional phenomena like beliefs differ in innumerable ways from the speech patterns that Port considers. Furthermore, nothing in Port's paper rules out that the practitioners of a mature cognitive science will explain these explananda by invoking intentionality, either in its familiar varieties or in some hitherto unknown ones. For example, consider the following paradigmatic intentional explanation and hotly disputed piece of political analysis:[65]

(Background)

> Throughout his career in Israeli politics, Ariel Sharon had solid right wing credentials. At several points in his career, he vociferously opposed the idea of Israel's withdrawing its military and civilian establishment from any of the territories Israel acquired in 1967, including the Gaza Strip. Yet as Prime Minister of Israel in 2005, Sharon orchestrated a withdrawal from the Gaza Strip, removing Israeli forces and evacuating the Israeli settlements there. This fact cries out for explanation: why did he do it?

(One Explanation)

> Sharon's decision was caused by his (correct) belief of the following things:
>
> 1. Many Palestinians would credit Hamas, an Islamist militant group, with bringing about Israel's withdrawal.
> 2. That would increase popular support for Hamas.

[65] Ross Brann advocated something like this in his Fall 2015 course on the subject. The explanation is also described in (Tessler 2009).

3. The ascendancy of militants following a Gaza withdrawal would give Israel's government a persuasive argument for refusing to withdraw from any more territory.[66,67]

For our purposes here, it does not matter whether or not (One Explanation) is correct.[68] Whether or not it is, whoever endorses it is attempting to explain why Sharon orchestrated the withdrawal by saying something about what he believed—that is, by giving an (everyday) intentional explanation.

It is worth noticing that behaviors like Sharon's are of central interest in subjects like ethics, political theory, and historical analysis. They are not irrelevant outliers that scholars in these disciplines can afford to ignore. Cognitive science must eventually address behaviors of these kinds, if it has any hope of contributing positively to the foundations of such disciplines, which occupy themselves with behaviors that we pre-theoretically judge to be paradigmatically cognitive. It is therefore significant that the explanandum in Port's experiment is very different from the one in the Sharon case. An especially obvious difference is that placement of stressed syllables and complex political decisions are no doubt produced by cognitive processes of radically different kinds. For example, it is quite plausible that the production of these two different kinds of behavior requires significantly different amounts of conscious control. Now, Port holds that placement of stressed syllables is susceptible to some amount of conscious control; but that still allows that it is generally accomplished in a comparatively unconscious and

[66] According to this reason, Israeli territorial withdrawals simply serve to strengthen the hand of the Palestinian militants who threaten Israel's security. Therefore, one cannot reasonably ask Israel to cede any more territory than it already has, especially not tactically significant territory like the West Bank.
[67] Whether or not Sharon anticipated (1)-(3), these propositions accurately describe what took place following the withdrawal.
[68] While Sharon may have anticipated (1)-(3), for my own part I find it implausible that this was his primary motivation in effecting the withdrawal. As an alternative explanation, commentators have held variously that Sharon was caving to the financial and logistical difficulty of protecting the Israeli settlements in Gaza, that he was attempting to relieve the great international pressure on Israel to offer something to the Palestinians, and that he had genuinely come to believe that any lasting peace would require significant territorial concessions on Israel's part.

automatic way, as introspection strongly suggests. By contrast, the process that led to Sharon's decision undoubtedly involved much painstaking, conscious deliberation; that man in particular was known for his elaborate strategizing. At any rate, whatever exactly the difference between these behaviors consists in, that they are quite different is obvious and surely something Port would grant. So, for all that the case of stressed syllable placement shows, there is room for cognitive scientists to invoke intentionality in order to explain things like Sharon's political behavior.

3.3. A Dynamicist Response

Although I have not seen it raised in connection with political examples, this sort of objection is not new. (Chemero 2009) anticipates a similar point, which he attributes to (Clark and Toribio 1994):

> Much work in radical embodied cognitive science explores what is often called *minimally cognitive behavior*, such as categorical perception, coordination, locomotion, and the like….The focus on minimally cognitive behavior is also necessary…given the current state of analytical and computational tools available. (I would be remiss if I didn't point out, though, that these tools get better every day.) What cognitive science needs, so the objection goes, is an approach that can explain real cognition, and for this you need representations. To my knowledge, the first version of this kind of response to radical embodied cognitive science is by Clark and Toribio (1994)….They wonder whether…radical embodied cognitive science…can ever account for what they call *representation-hungry cognitive tasks*. There are certain tasks, Clark and Toribio claim, that simply cannot be accomplished without representations. How, for example, could one think about temporally and spatially distant objects and events without mental representations of them?

One response that Chemero identifies is essentially what I have been recommending in response to Port's example:

> [A]gree that nonrepresentational analyses may be appropriate for what Brooks (1991) calls "the bulkiest parts of intelligent systems," but [insist] that more advanced cognition—thinking about the past, the future, the distant environment—requires internal [what I have called *mental*] representation and computation. [Such a] compromise…seems to some to find support in evidence about the brain[69]….One could,

[69] Here Chemero cites (Milner and Goodale 1995) and (Norman 2002).

> of course, accept this and resign radical embodied cognitive science to vision for action, using computational approaches for "real" representation-hungry cognition. (pp.38-39)

However, that is not the response that Chemero himself recommends. "Another, less defeatist possibility," he writes, "is to use empirical work to show that radical embodied cognitive science has the resources to explain representation-hungry tasks" (p.40). Chemero describes a study in (Van Rooij, Bongers, and Haselager 2002), in which agents are presented with a series of sticks varying in length, and asked whether they could use the sticks to move a distant object. The authors present an equation that, they hypothesize, describes the agents' responses in terms of the distance of the object and a parameter *k*, which is determined by "the length of a rod on a particular trial, the rod length on previous trials, and the agent's response on previous trials" (p.41). The task is representation-hungry because

> the subjects are asked to predict the outcome of an imagined action, one that hasn't yet happened and so is not perceivable….It would seem to require a comparison of a judged distance with a judged combined stick-plus-arm length. Indeed, some would argue that judging the distance of the to-be-poked object also requires a mental comparison of the expected size of the object with its apparent size. (p.40)

Yet, Chemero insists, "the model accurately accounts for the imagination of the action without calling upon mental representations of the action" (p.42). The authors were able to use the equation to predict a number of things about the agents' responses, predictions which were then confirmed when the experiment was performed.[70]

I have two points to make about this example. I grant that the task in (Van Rooij et al 2002) is a representation-hungry one. Moreover, Chemero emphasizes that it is a conscious task, since it requires agents to "report on their imagination" of something that has not happened (p.42). In that respect, it is arguably closer to Sharon's political behavior than placement of stressed syllables in the speech period is. However, the stick task is still far less sophisticated

[70] See pp.41-42 for some of these predictions.

than the kinds of behaviors that we saw in the Sharon example. Sharon's decision involved making educated guesses about the interaction of comparatively quite abstract social and political forces, not just his own possible bodily movements.

Chemero implicitly recognizes the limited scope of the extant dynamical explanations when he writes, "it is still an open question how far beyond minimally cognitive behaviors radical embodied cognitive science can get. We will have to wait and see." Indeed, we will. But in the meantime, there is the issue of which working hypotheses we are justified in adopting, and here skeptics must take care not get ahead of themselves. While it is illuminating to attempt to do without intentionality as far as one can, and while scientists should surely continue to produce studies like (Van Rooij et al 2002), so far, no such studies have gone so far as to explain behaviors like Sharon's, which are not merely representation-hungry but paradigm candidates for intentional explanation. I conclude that no such studies have gone far enough to justify (Elim). (In principle one can make a similar point about (Churchland 2012)'s explanations. Due to space limitations, I won't try to develop that point here.)

My second point is that even if we assume that facilitating useful predictions can be sufficient to make something an explanation, there are important kinds of predictions that Van Rooij et al's equation by itself cannot be used to make. What would happen, for example, if the agent were told, "Now put down the sticks and leave the room"? Simply saying that the agent's answers conform to the equation in a given context does not account for the ways that conformance to the equation is mediated by other facts about the agent, such as her belief that she is looking at sticks. We saw a similar point in connection with Field's pilot case: folk psychological explanations situate the behaviors being or capacities explained in a broader network of other cognitive states, behaviors, and facts about the external environment; and like

(Behavioral Effects) and (Relation to Airspeeds), by itself Van Rooij et al's explanation does not do this.

Now to be fair, what Van Rooij et al purport to be offering is a scientific explanation, not an everyday one, and it might be that all extant scientific intentional explanations that could compete with the dynamical one would likewise fail to be informative in the same ways as everyday ones. My point is only that because people's behaviors are systematically linked to one another in ways we ordinarily take to be mediated by their beliefs, any science of behavior that aspires to provide explanations that are informative in all the same ways as everyday ones will inevitably need to account for these links somehow or other. Perhaps that can be done by appealing to yet further differential equations, but it has not yet been done. Until more is said to show that this can be done, we are not forced to give up on the idea that intentional descriptions of our explanantia play a distinctive, ineliminable role in psychological explanation.

4. Intentionality in the Explanandum

4.1. Explanantia and the Nature of Psychological Explanation

In Section 2 and to some extent Section 3, I emphasized the distinctive qualities of explanations that appeal to familiar intentional phenomena, such as beliefs and desires. But in contemporary cognitive science, we also find many less familiar explanantia that are intentional, or at least that are standardly characterized in intentional terms. For example, computational approaches model cognition as the manipulation of internal representations, which represent various entities or aspects of the subject's external environment.[71] E.g., perception of an external object might consist of interactions between elements in the visual system that represent the

[71] See, for example, (Marr 2010).

object's edges. Likewise, connectionist approaches model cognition as the disciplined transition between states of neural activation, where these states are conceived of as representing various aspects of reality that are relevant to the capacity being explained.[72] In both cases, cognitive capacities that are intentionally characterized to begin with, such as the capacity to multiply numbers, the capacity to perceive the distances of external objects, or the capacity to recognize human faces, are then also explained in intentional terms.

On one plausible, prominent conception of psychology, this appeal to intentionality in the explanantia is an inevitable consequence of the widespread characterization of explananda in intentional terms. On this conception, the business of psychology is to explain cognitive capacities by breaking them down into their components and then showing how these components are realized in physical systems.[73] But so long as intentionality proves to be difficult to reduce, some of the components are bound to be intentionally characterized; otherwise, the explanans will be expressed in a vocabulary that is un-relatable to the explanandum in any clear way.[74] It follows, then, that wherever the capacities being explained are intentionally characterized, so too will be some of the explanantia.

Still, one might think, if intentionality in the explanantia arises from intentionality in the explananda, then that simply gives us another reason to ask whether intentionality can be excised from the explananda. I will now address this question, arguing that there are important reasons to think it cannot be excised.

---

[72] For example, see (Churchland 2012), who describes tokenings of activation states as being analogous to indexings of a map—that is, to acts of pointing to a particular spot and saying 'we are here'.

[73] (Cummins 2000) observes that despite the lip-service frequently paid to the Deductive-Nomological account of explanation, "actual theory building and explanation" in psychology "takes place in frameworks…not designed for the elaboration of laws but rather…for the elaboration of functional analyses" (p.137). To explain a capacity by giving a *functional analysis* is to break down the capacity into sub-capacities, the organized manifestation of which constitutes manifestation of the capacity being explained.

[74] See (Egan 1995) p.189 for similar remarks.

4.2. Under the Metaphysical Conception of Explanation

At the beginning, I defined psychological explanations to be explanations of people's and sophisticated animals' behaviors and cognitive capacities. As (Burge 2010) forcefully emphasizes throughout, in actual practice many of the capacities that cognitive psychologists try to explain receive intentional descriptions, and indeed strongly appear to be themselves intentional. For example, a celebrated explanandum in perceptual psychology is two-eyed organisms' capacity to estimate the distances of objects in their environments using the disparities in the images received from their two retinas. But on this standard characterization the capacity is clearly intentional: any act of estimating the distance of a perceived object has that distance as its subject matter. On the metaphysical conception of explanation, those who think intentionality can be eliminated from the explananda of scientific psychology must make the obviously false claim that many of our standard characterizations of our explananda, including the one just given, are false: these explananda are in fact non-intentional.

In fact, given a metaphysical conception of explanation and a reasonable conception of perceptual psychology, the inevitable presence of intentional explananda in the latter can be seen from the armchair. As explained at the beginning (see footnote 3), perception necessarily has a subject matter—viz., the things that are perceived and the ways they are perceived as being—and is therefore intentional. If the business of perceptual psychology is to explain how agents manage to perceive things, then the explananda of perceptual psychology are bound to be things that are intentional, regardless of how they are described. That makes the explanations intentional too, given a metaphysical conception of explanation and the explananda-inclusive reading of 'intentional explanation'.

This argument applies whether one is doing mainstream perceptual psychology or *ecological psychology* (see (Gibson 1966) and (Gibson 1967)). This is an important point to emphasize, since many dynamical systems theorists are committed to some version of ecological psychology.[75] The difference is that while in standard psychology the objects of perception are familiar things like rocks, chairs, and colors, in ecological psychology they are *affordances*—environmental opportunities for behavior. But that is quite irrelevant to the anti-skeptical point I am making here. Whether perception is of familiar things or of affordances, there is something it is of; and thus, it has a subject matter, and so is intentional in my sense of the term. (Ecological psychologists who adopt the epistemic conception of explanation have a response to this, which I'll criticize in the next subsection.)

The observations I have been making apply also to (Churchland 2012), which envisions a psychology that assigns no role to propositional representation.[76] Rather, on Churchland's view, one can explain many of our cognitive capacities by positing certain non-propositional representations. These representations are map-like, in the sense that they represent features of the external world by being homomorphic to them. However, as I'll now explain, there is a risk that Churchland's explanation of how his map-like representations manage to represent will in turn furnish an explanation of one form of propositional representation.

To see the point, notice that in general there are aspects of the content of any map that can be specified by that-phrases. For example, a map of the United States can (incorrectly) represent that Washington, D.C. is situated directly on a coast. (In fact, D.C. sits somewhat inland, on a river.) Thus, even if the content of the map as a whole cannot be specified by a single that-phrase, some aspects of its content can be specified in this way; one need only fix on

[75] For another example, see (Turvey and Shaw 1979).
[76] See p.4, 24, and 49 for some representative passages.

specific features that are being represented, such as the location of Washington, D.C. The point is, once we have allowed that the map is a map of the United States that contains a representation of Washington, D.C., we have sufficient materials to explain what it is for the map to represent Washington, D.C. as being situated directly on a coast; and the claim that D.C. is located on a coast is propositional.

Moreover, on the metaphysical conception of communicative explanation, Churchland's picture arguably gives propositional representation a role as an explanans as well. For it is plausible that if an item's being a map of the U.S. can serve as an explanans, then its propositional features, such as representing that D.C. is situated on a coast, can also serve as explanantia. (For instance, that could explain why map users who take routes through D.C. systematically fail to arrive at the beach.) Churchland's mental representations, being map-like, are no exception in this regard; see pp.38-45 and p.85 for some examples.[77] If, as Churchland grants, his representations succeed in representing various objects, then it is hard to avoid the conclusion that they also succeed in representing these objects as having various properties and as standing in various relations to one another. And if non-propositional representation by map-like entities serves as an explanans in psychology as Churchland conceives of it, then it is hard to avoid the conclusion that the associated propositional representation serves as an explanans as well.

[77] One can make the same point again when it comes to momentary indexings of maps. When your companion points to a spot on the map in order to indicate where you are, she thereby commits to any number of propositions: e.g., that you are south of Oregon, west of the Mississippi, etc. To see why this act of pointing involves propositional commitments, consider that it would make sense to accuse your companion of communicating something false, based on her act of pointing. Similar things can be said about indexings of Churchland's map-like representations. For Churchland, undergoing certain kinds of momentary neural activation can amount to representing an object as instancing a particular combination of properties—thus placing the object at a particular location in the space of properties that the map represents. But then such momentary activations have propositional content: namely, that the object in question has the properties associated with that point in the property space. See his p.4 for an example.

4.3. Under the Epistemic Conception of Explanation

I just discussed skeptics who adopt the metaphysical conception of explanation. I will now consider the fate of skepticism about intentionality in the explanandum under the epistemic conception. An immediate observation is that *prima facie*, the epistemic conception renders the intentional nature of many of psychology's explananda irrelevant to the debate. For on the epistemic conception, the skeptic's claim does not concern the nature of the entities explained, but rather the descriptions employed in the explanations. Her characteristic thesis is that intentional description of these explananda contributes nothing to the explanation, and so can be profitably jettisoned in favor of non-intentional characterizations. For example, such a skeptic would hold that in perceptual psychology, the standard explanations of how agents manage to perceive the distances of objects in their environments can be profitably rephrased so that the explananda are described in non-intentional terms. Perhaps, for instance, she would take the proper explananda simply to be the patterns of behavior typically associated agents' perceptions of the distances of objects in the environment, rather than these perceptions themselves.[78]

The suggestion just fielded highlights a concern about the skeptical claim on offer: the skeptic would not only have psychologists modify the ways that they describe their explananda, she would have them explain different phenomena altogether. In the example just given, the standard, intentional explanation was of how agents perceive the distances of external objects. By contrast, the non-intentional explanation is of something else: certain patterns of behavior typically associated with such perceptions. I proposed the latter because I was unable to concoct a non-intentional description of the original explanandum. Now, if everything intentional can be

[78] This view is *methodological behaviorism*. Although it has fallen out of favor, from the late 1910s to the 1950s it was a very popular conception of the ideal towards which scientific psychology should strive.

fully described in non-intentional terms, then my failure was simply due to lack of ingenuity. In that case the skeptic can simply claim that standard explanations of how agents judge object distances should instead describe that explanandum—the very same one—in non-intentional terms.

However, my failure to come up with a non-intentional description of *perceiving the distances of external objects* may well not have been an accident. Many philosophers suspect that nothing intentional is reducible to anything non-intentional; and if that is the case then nothing intentional can be fully (correctly) described in non-intentional terms. Thus, in the many cases in psychology in which the explanandum at hand is described in intentional terms, the move to non-intentional description will render the explanation an explanation of something else. Therefore, insisting on non-intentional descriptions across the discipline amounts to changing the subjects of much of psychology. Of course, a skeptic could simply insist that this change is all for the good, since the non-intentional explananda are what psychologists should have been explaining all along. But one might reasonably be suspicious of such a radical, sweeping claim.

Another concern is that an insistence on non-intentional characterization of the explananda in psychology would hinder psychology's ability to contribute to other disciplines. To see why, note that a virtue we might reasonably ask of a scientific psychology is that some of its explanations shed light on disciplines such as epistemology, ethics, political theory, historical analysis, and sociology by explaining the cognitive phenomena in which these disciplines traffic. But the applicability of a non-intentional psychology to these disciplines is unclear, given their wholesale reliance on intentional talk. For example, given that we explain why politicians do what they do by saying what they believe, *prima facie* what historians and political theorists need from a purportedly more fundamental discipline like cognitive psychology is an explanation of

belief. It is not obvious that these scholars would be equally well served by an explanation of some other, non-intentional phenomenon. Likewise, in many disciplines the definitions of central concepts are given in intentional terms. E.g., in epistemology many philosophers have tried to understand epistemic justification in terms of "virtuous" processes of belief formation, and in turn to understand rationality in terms of such processes, together with the logical coherence of one's beliefs.[79] And even philosophers who do not want to understand these fundamental notions in terms of beliefs do understand them in an intentional way, in terms of non-propositional representations.[80] On first blush, at least, it is hard to see how rationality and epistemic justification could be understood in entirely non-intentional terms.

All that said, there may well be ways for a non-intentional psychology to shed light on other subjects that are currently saturated with intentionality. One rather modest way would be by formulating generalizations that relate psychology's non-intentionally-described explananda to the intentionally described explanantia of other disciplines. Due to the inclusion of intentional terms, a skeptic could not regard such generalizations as being themselves proper explananda of psychology. But nonetheless, once these generalizations are on the table, scholars in other disciplines might find some of them useful as working hypotheses. For example, one might assume as a working hypothesis that people who undergo certain non-intentionally described processes, these being explained by psychology, tend to believe that social justice will never be achieved in the country in which they live. If the non-intentionally described processes had clear, readily-identifiable causes, then the above generalization (if true) could provide sociologists with a useful means of predicting the behaviors associated with despair in social justice movements.

[79] See (Sosa 1985), (Zagebski 1997), and (McDowell 1994).
[80] See (Churchland 2001).

The skeptical line just discussed allows for intentional talk in other disciplines. But a different, more radical project would be to insist on the non-intentional reformulation of discourse in other disciplines outside psychology. Whatever the prospects of this project turn out to be, it is worth noticing just how radical the project is. It promises a sweeping reform of the central concepts of many different disciplines, wherever these concepts are standardly defined in intentional terms. E.g., in epistemology, the central notion of a justified belief would have to be replaced. Moreover, since many of these disciplines characterize their explanantia in intentional terms, the view under consideration would have to insist on replacing these characterizations. This project is thus much more ambitious than the one with which we initially began; the latter only proposed to eliminate intentional characterization of explananda, and that only in psychology.

So far, my anti-skeptical remarks in this section have been rather tentative and cautionary. However, skeptics who embrace ecological psychology, it is possible to raise a more biting criticism. Ecological psychology (recall Section 4.2) claims that perception is not of external objects, but rather of *affordances*, environmental opportunities for behavior. Ecological psychology is meant to serve as a *guide to discovery*: a working hypothesis that helps scientists to predict new phenomena and generate new experiments.[81] These are precisely the sorts of virtues that skeptics who embrace an epistemic conception of explanation deny of intentional characterizations when they accept (Elim). But then these skeptics must also deny that ecological psychology's characterization of perception has these virtues. That is because, as we saw in Section 4.2, ecological psychology describes perception as having a subject matter (viz., affordances).

[81] See (Chemero 2009) p.85 for this definition. And see (Chemero 2009) chapter 5, and (Turvey, Shaw, and Mace 1981) for endorsement of ecological psychology as a guide to discovery.

Given that it conflicts with a central presupposition of her approach—viz., that one can gainfully characterize perception as being of affordances—saddling an ecological psychologist with (Elim) might seem uncharitable. In fact, this observation leads us to an important point. Although this can be obscured by their sometimes-radical pronouncements against familiar varieties of intentionality, the most charitable conception of dynamical systems theorists is not that they are attempting to excise intentionality from cognitive psychology. Rather, it is that for them, intentional phenomena, even explicitly so-characterized, serve as central explananda. It is just that they regard intentionality as a property of agents which emerges from their dynamic interactions with their environments—that is, interactions that can be described by sets of *coupled*[82] differential equations—rather than as a feature of some objects or states that serve as intermediaries between the agent and the external world.[83] This process-focused conception of intentionality is evidenced in many places. For example, the slogan "perception is of affordances" speaks of perception, an activity, rather than, say, of perceptual states. Still the description is quite obviously intentional, as I have emphasized throughout, since in being of affordances, perception is thereby of something. As further evidence, (Chemero 2009) frequently identifies his adversary as the idea that "the main business of cognition is…*mental gymnastics*, the construction, manipulation, and use of representations of the world" (p.18). But one can hold that agents represent things in the external world without claiming that they do so by means of manipulating any internal items that serve as intermediaries or stand-ins for those things. I submit that in light of the difficulties that arise for skepticism about intentionality, this is

[82] A pair of differential equations is *coupled* if the variables in one serve as parameters in the other.
[83] That said, some dynamical systems theorists leave little doubt about their intention to eliminate intentionality from the explanantia of cognitive science. Hence my discussion of Chemero in Section 3.

precisely how we should most charitably characterize the outlook of most dynamical systems theorists, particularly those who embrace ecological psychology.

5. Concluding Remarks

In this essay, I have defended the claim that intentionality plays an ineliminable role in psychological explanation. The content of this claim is significantly influenced by one's conception of communicative explanation, and by whether one is targeting intentionality as an explanans or as an explanandum. Against skeptics who adopt a metaphysical conception of communicative explanation and focus on explanantia, it is possible to argue that the non-intentional explanations that these skeptics favor appeal to intentionality in fact if not in name, and so are intentional after all. The better these explanations approximate the virtues of intentional ones, the more plausible that claim becomes. A central case in point was everyday intentional explanations, which have a distinctive counterfactual spread that has yet to be replicated in non-intentional explanations.

On the other hand, skeptics who adopt an epistemic conception of explanation can resist the charge of appealing to intentionality in fact if not in name. Their point is that intentional description contributes little to the virtues of an explanation. However, I argued that at the time of writing, even the considerable achievements of a congenial branch of cognitive science do not go far enough to justify (Elim) as such skeptics construe it.

Whatever one thinks about intentionality in the explanans, it is hard to play down the role of intentional phenomena as explananda in much of cognitive psychology. This point is especially clear in the case of perception, which necessarily has as its subject matter the things that are perceived. Moreover, the theoretical orientation of many would-be skeptics, that of

ecological psychology, in fact commits them to intentional description of their explananda. They are most charitably understood as targeting not intentionality itself, but merely a conception of intentionality as necessarily involving the manipulation of internal intermediaries that stand in for external objects.

REFERENCES

1. Achinstein, Peter. (1983). *The Nature of Explanation*. New York: Oxford University Press.
2. Anscombe, Gertrude Elizabeth Margaret. (1971) *Causality and Determination*. Cambridge University Press, Cambridge, MA.
3. Batterman, Robert W. (2001). The Devil in the Details: Asymptotic Reasoning in Explanation, Reduction, and Emergence. Oxford: Oxford University Press.
4. Batterman, Robert W. (2002). "Asymptotics and the Role of Minimal Models." *British Journal for the Philosophy of Science*, 53: 21-38.
5. Bechtel, William, and Abrahamsen, Adele. (2005) "Explanation: A Mechanist Alternative." *Studies in History and Philosophy of Biological and Biomedical Sciences*, Vol. 36, pp.421–441.
6. Bechtel, William. (2008) *Mental Mechanisms*. Routledge, New York.
7. Braverman, Mike, Clevenger, John, Harmon, Ian, Higgins, Andrew, Horne, Zachary S., Spino, Joseph, and Waskan, Jonathan. (2012). "Intelligibility is Necessary for Explanation but Accuracy May Not Be." *Proceedings of the Thirty-Fourth Annual Conference of the Cognitive Science Society*.
8. Burge, Tyler. (1989). Individuation and Causation in Psychology. Pacific Philosophical Quarterly, 70(4), 303-22.
9. Burge, Tyler. (2010). Origins of Objectivity. Oxford, UK: Clarendon Press.
10. Chemero, Anthony. (2009). *Radical Embodied Cognitive Science*. Cambridge, Massachusetts: MIT Press.
11. Churchland, Paul. (1981). Eliminative Materialism. The Journal of Philosophy, 78(2), 67-90.

12. Churchland, Paul. (2012). Plato's Camera: How the Physical Brain Captures a Landscape of Abstract Universals. Cambridge, MA: The MIT Press
13. Craver, Carl. (2006). "When Mechanistic Models Explain." *Synthèse* 153 (3): 355-376.
14. Craver, Carl. (2007). Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience. Oxford: Clarendon Press.
15. Craver, Carl. (2014). "The Ontic Account of Scientific Explanation." In Kaiser, Scholz, Plenge, and Hütteman (eds.), *Explanation in the Special Sciences*. Springer.
16. Cummins, Robert. (2000). "'How does it work?' versus 'What are the laws?': Two conceptions of psychological explanation." In Frank C. Keil and Robert Wilson (eds) *Explanation and Cognition*. Cambridge, MA: MIT Press and also in his *The World in the* Head Oxford: OUP, 2010.
17. Dennett, Daniel. (1981) *Brainstorms: Philosophical Essays on Mind and Psychology*. Cambridge, Massachusetts: MIT Press.
18. Dennett, Daniel. (1987) *The Intentional Stance*. MIT Press, Cambridge, Massachusetts.
19. Dretske, Fred. (1981). Knowledge & the Flow of Information. Cambridge, Mass: MIT Press.
20. Dretske, Fred. (1988). Explaining Behavior: Reasons in a World of Causes. Cambridge, Mass: MIT Press.
21. Egan, Frances. (1995) "Computation and Content." *The Philosophical Review*. Vol. 104, No. 2, pp.181-203.
22. Eliasmith, Chris. (2010) "How We Ought to Describe Computation in the Brain." *Studies in History and Philosophy of Science*, 41, pp.313–320.
23. Eliasmith, Chris. (2013). *How to Build a Brain*. Oxford University Press.
24. Field, Hartry. (1978). "Mental Representation." *Erkenntnis* 13, pp.9-61.

25. Field, Hartry. (1986). “The Deflationary Conception of Truth.” In G. MacDonald & C. Wright (eds.), *Fact, Science, and Morality* (pp.55-117). Oxford: Blackwell.
26. Field, Hartry. (1994) “Deflationist Views of Meaning and Content,” *Mind*, New Series, Vol. 103, No. 411.
27. Field, Hartry. (2001). *Truth and the Absence of Fact.* Oxford: Oxford University Press.
28. Fodor, Jerry A. (1974) “Special Sciences: Or the Disunity of Science as a Working Hypothesis.” *Synthese*, 28, 97–115.
29. Fodor, Jerry A. (1987). Psychosemantics: The Problem of Meaning in the Philosophy of Mind. Cambridge, Mass.: MIT Press.
30. Fodor, Jerry A. (1990). *A Theory of Content*. Cambridge, Mass.: MIT Press.
31. Gordon, Robert M. (1986) “Folk Psychology as Simulation.” *Mind and Language*, Vol. 1, No. 2, pp.158-171.
32. Gordon, Robert M. (2009) “Folk Psychology as Mental Simulation”, *The Stanford Encyclopedia of Philosophy* (Fall 2009 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/fall2009/entries/folkpsych-simulation/>.
33. Hochstein, Eric. (2012). “Minds, Models, and Mechanisms: A New Perspective on Intentional Psychology.” *Journal of Experimental & Theoretical Artificial Intelligence*, 24 (4), 547-557.
34. Kaplan, David. M., & Craver, Carl F. (2011). “The Explanatory Force of Dynamical and Mathematical Models in Neuroscience: A Mechanistic Perspective.” *Philosophy of Science*, 78, 601-627.
35. Kelso, J.A. Scott. (1995), Dynamic Patterns: The Self-Organization of Brain and Behavior, Cambridge, MA: The MIT Press.

36. Kitcher, Philip. (1989) “Explanatory Unification and the Causal Structure of the World.” In *Scientific Explanation*, Kitcher, Philip and Salmon, Wesley, pp.410-505. Minneapolis: University of Minnesota Press.
37. Machamer, Peter; Darden, Lindley; and Craver, Carl F. “Thinking about Mechanisms.” (2000) *Philosophy of Science*, Vol. 67, No. 1, pp.1-25.
38. Marr, David. (2010) *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information.* MIT Press. (Original published by W.H. Freeman and Company, 1982.)
39. McDowell, John. (1994) *Mind and World*, Cambridge, MA: Harvard University Press.
40. McLaughlin, Brian and Bennett, Karen. (2014) “Supervenience,” *The Stanford Encyclopedia of Philosophy* (Spring 2014 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/spr2014/entries/supervenience/>.
41. Millikan, Ruth Garrett. (1984) Language, Thought, and Other Biological Categories: New Foundations for Realism. Cambridge, Mass.: MIT Press.
42. Morton, Adam. (1980) Frames of Mind: Constraints on the Common-Sense Conception of the Mental. Oxford: Clarendon Press.
43. Port, Robert F. (2003) “Meter and Speech”, *Journal of Phonetics*, 31, 599-611.
44. Quine, Willard Van Orman. (1960) *Word and Object*. Cambridge, Mass., MIT Press.
45. Quine, Willard Van Orman. (1992) *Pursuit of Truth*. Cambridge, Mass., Harvard University Press.
46. Rice, Colin. (2013). “Moving Beyond Causes: Optimality Models and Scientific Explanation.” *Noûs*, Vol. 49, No.3, pp.589-615.

47. Riley, Michael A., Shockley, Kevin, and Van Orden, Guy. (2011). “Learning from the Body about the Mind.” *Topics in Cognitive Science* 4 (1): 21-34.

48. Silberstein, Michael and Chemero Anthony. (2011). “Complexity and Extended Phenomenological-Cognitive Systems.” *Topics in Cognitive Science* 4 (1): 35-50.

49. Sosa, Ernest. (1985) “The Coherence of Virtue and the Virtue of Coherence: Justification in Epistemology,” *Synthese*, Vol. 64, pp.3-28.

50. Stalnaker, Robert. (1984) *Inquiry*. Cambridge, Mass.: MIT Press.

51. Stich, Stephen P. (1983) From Folk Psychology to Cognitive Science: The Case against Belief. Cambridge, Mass: MIT Press.

52. Stich, Stephen, and Nichols, Shaun. “Folk Psychology: Simulation or Tacit Theory?”. *Mind & Language*, Vol. 7, No. 1-2, pp.35-71.

53. Thelen, Ester, and Smith, Linda. (1994). A Dynamic Systems Approach to the Development of Cognition and Action. Cambridge, MA: MIT Press.

54. Van Gelder, Timothy, and Port, Robert. (1995). “It’s About Time: An Overview of the Dynamical Approach to Cognition.” In Port, Robert, and Van Gelder, Timothy (eds.), *Mind as Motion: Explorations in the Dynamics of Cognition*. Cambridge, MA: MIT Press. 1-43.

55. Waskan, Jonathan, Harmon, Ian, Horne, Zachary, Spino, Joseph & Clevenger, John. (2014). “Explanatory Anti-Psychologism Overturned by Lay and Scientific Case Classifications.” *Synthèse*, Vol. 191, pp.1013-1035.

56. Woodward, James. (2000). “Explanation and Invariance in the Special Sciences.” *British Journal for the Philosophy of Science*, 51: 197-254.

57. Woodward, James. (2003). *Making Things Happen: A Theory of Causal Explanation*. Oxford: Oxford University Press.

58. Woodward, James. (2017) “Scientific Explanation”, *The Stanford Encyclopedia of Philosophy* (Spring 2017 Edition), Edward N. Zalta (ed.), forthcoming URL = <https://plato.stanford.edu/archives/spr2017/entries/scientific-explanation/>.

59. Zagzebski, Linda Trinkaus. (1997) “Virtue in Ethics and Epistemology,” *American Catholic Philosophical Quarterly*, Vol. 71 (Supplement), pp.1-17.