Counterpossibles

COUNTERPOSSIBLES

A Dissertation Presented

by

BARAK L. KRAKAUER

Submitted to the Graduate School of the

University of Massachusetts Amherst in partial fulfillment

of the requirements for the degree of

DOCTOR OF PHILOSOPHY

February 2012

Department of Philosophy

COUNTERPOSSIBLES

A Dissertation Presented

by

BARAK L. KRAKAUER

Approved as to style and content by:

_____

Phillip Bricker, Chair

_____

Joseph Levine, Member

_____

Jonathan Schaffer, Member

_____

Angelika Kratzer, Member

_____

Hilary Kornblith, Department Head
Philosophy

ACKNOWLEDGEMENTS

I would like to thank many people who have offered their invaluable support during the completion of this project. The personal, professional, and philosophical help I received over the course of this dissertation was essential to its completion in its present form.

My advisor, Phil Bricker, was especially supportive and helpful. I am grateful for his patience and thoughtful criticism of the various drafts of this project as well as presentations of ideas connected to it. He has been incredibly helpful not only reading and commenting on the ideas presented, but in locating further avenues to explore and discuss. His ability to understand and evaluate arguments as well as locate interesting problems nearby was essential in shaping my views of counterpossibles, and his clarity and insight served as a paradigm of how to approach metaphysical issues.

Jonathan Schaffer was also extremely helpful, not only in reading and commenting on the work, but inspiring much of the project. He was generous with his time and attention, and the time spent studying with him and others at the Australia National University had a large impact on the project. His creativity and ability to survey philosophical terrain makes him an invaluable influence.

Joe Levine was also patient and helpful, reading and commenting on several drafts. The comments on dependence and mental properties that he gave me challenged me to rethink many aspects of how this proposal of counterpossibles can be of use in various disputes in philosophy. His comments and suggestions provide the clearest advice on how to elaborate and defend this project as it is further developed.

My development at UMass also owes much to the other graduate in the program, particularly those involved with the dissertation seminars. I would especially like to thank

Einar Bohn, Heidi Buetow, Sam Cowling, Jeff Dunn, Ed Ferrier, and Kelly Trogdon for their suggestions over the course of this project. I would also like to thank others who have read drafts of this project and gave helpful comments over the course of the writing of this dissertation, including Jodi Azzouni, David Chalmers, Ned Hall, and Stephen Yablo. Finally, I would like to thank audiences at the Brandeis University Graduate Philosophy Conference (2010), the Pacific APA (2009), the AAP (2009), and the Southeast Graduate Conference (2008) and Princeton/Rutgers Graduate Conference (2008) for comments on earlier versions of this work.

I would also like to thank my parents, Marcia and Randall Krakauer, as well as Amy Rose Deal, for their love and support.

ABSTRACT

COUNTERPOSSIBLES

FEBRUARY 2012

BARAK KRAKAUER, B.A., BRANDEIS UNIVERSITY

Ph.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Phillip Bricker

Counterpossibles are counterfactuals with necessarily false antecedents. The problem of counterpossibles is easiest to state within the "nearest possible world" framework for counterfactuals: on this approach, a counterfactual is true (roughly) when the consequent is true in the "nearest" possible world where the antecedent is true. Since counterpossibles have necessarily false antecedents, there is *no* possible world where the antecedent is true. On the approach favored by Lewis, Stalnaker, Williamson, and others, counterpossibles are all trivially true.

I introduce several arguments against the trivial approach. First, it is counter-intuitive to think that all counterpossibles are true. Second, if all counterpossibles were true, then we could not make sense of their use in logical, philosophical, or mathematical arguments. Making sense of the role of sentences like these requires that they not have vacuous truth conditions.

The account of counterpossibles I ultimately favor is an extension of the "nearest possible world" semantics discussed above. The Lewis/Stalnaker account is supplemented with the addition of *impossible* worlds, and the nearness metric is extended to range over these impossible worlds as well as possible worlds. Thus, a counterfactual is true when its consequent is true in the nearest *world* where the antecedent is true; if the counterfactual's antecedent is impossible, then the nearest world in question will be an impossible world. Once the framework of impossible worlds and similarity is in place, we can put it to use in the analysis of other philosophical phenomena. I examine one proposal that makes use of a theory of counterpossibles to develop an analysis of the notion of metaphysical dependence.

TABLE OF CONTENTS

INTRODUCTION

Counterpossibles are counterfactuals with impossible antecedents. The aim of this dissertation is to provide an analysis of counterpossibles that makes sense of their use in natural language as well as philosophical practice. Below is a summary of the various chapters of this dissertation. In the remainder of this introduction, I explain what kinds of statements are properly called 'counterfactuals,' and briefly discuss some different kinds of ways in which a statement is impossible.

1. Summary

In chapter 1, I discuss the "nearest possible world" analysis of counterfactuals developed by Lewis and Stalnaker: a counterfactual is true (roughly) if the consequent is true in the "nearest" or "most similar" possible worlds where the antecedent is true. While I am sympathetic with the "nearest possible world" account, and assume that such an account is appropriate for at least most counterfactuals, it cannot straightforwardly accommodate counterpossibles. The problem of counterpossibles can be seen clearly enough: how can we determine whether a counterpossible is true if there is no possible world where the antecedent is true, and thus no possible world at which we may evaluate the truth of the conditional?

According to the approach favored by Lewis, Stalnaker, and others, *all* counterpossibles are true; they are said to be *vacuous*, since they all express the necessary truth. There is no difference in content among counterpossibles, or even between counterpossibles and other necessary truths.

I introduce several of arguments against the vacuous approach. First, it is simply counter-intuitive to think that all counterpossibles are true; consider, for example, 'If nine were prime, then there would be fewer kangaroos.' Second, if all counterpossibles were true, then we could not make sense of their (non-vacuous) use in logical, mathematical, or philosophical arguments. Consider counterpossibles such 'If Platonism were true in mathematics, then we would have no way of knowing about numbers,' or 'If there were finitely many prime numbers, then there would be some number *n* equal to the product of every prime number + 1.' Making sense of the role of sentences such as these requires that they not be vacuous.

The behavior of 'might' counterpossibles is also problematic for the defender of the vacuous approach. According to Lewis, 'might' counterpossibles are all false. In addition to introducing a new family of counterpossibles with counter-intuitive truth values, the falsity of 'might' counterpossibles undercuts many of the defenses of the truth of all 'would' counterpossibles. The falsity of 'might' counterpossibles also introduces widespread violations of some plausible principles about the behavior of counterfactuals, such as the principle that a 'would' counterfactual entails its corresponding 'might' counterfactuals. Ultimately, only a non-vacuous account of counterpossibles is capable of making sense of our linguistic intuitions and philosophical practice, and only a non-vacuous account of counterpossibles is capable of explaining the failure of the logical principle related to the behavior of 'might' and 'would' counterfactuals.

In chapter 2, I attempt to formulate several versions of a non-vacuous approach to counterpossibles that addresses at least some of the worries presented above in chapter 1, without the addition of impossible worlds.

One such approach is the "nearest possible proposition" account of counterpossibles. On such an account, a counterpossible is evaluated not at the nearest possible world where the antecedent is true, but at the nearest possible world where the nearest possible antecedent is true. That is, a similarity metric is applied not only to possible worlds, but to propositions as well; the impossible antecedent of some counterpossible is nearby to some possible proposition that could serve as the antecedent of some counterfactual. We then evaluate whether the counterpossible is true by evaluating whether this counterfactual (whose antecedent is this relevantly similar, possible proposition) is true. Of course, it is difficult to say just what the nearest possible proposition might look like, and it is difficult to see why we will get the intuitively correct truth conditions for counterpossibles by making use of other, nearby counterfactuals.

Another such approach is to analyze a counterpossible such that it is true when its consequent can be derived, in some non-classical logic, from the antecedent and other background propositions. Such an approach, however, faces embarrassing questions related to how one chooses the logical system, and whether one could ever make sense of counterpossibles whose antecedent express some impossibility relative to the logic of the counterfactual conditional.

In chapter 3, I present my favored account of counterpossibles, which is an extension of the "nearest possible world" semantics discussed above in chapter 1. The Lewis and Stalnaker account is supplemented with the addition of *impossible* worlds, or worlds where logical, mathematical, or metaphysical impossibilities obtain (and do not necessarily result in absurdity). The similarity metric between worlds is extended to range over these impossible worlds as well as possible worlds. Thus, a counterfactual is true when its consequent is true in the nearest world (possible or impossible) where the antecedent is

true; if the counterfactual's antecedent is impossible, then the nearest world in question will be an impossible world.

One advantage of this approach is that it is a straightforward extension of semantics of counterfactuals that has been highly influential. In particular, the similarity metric is one based on the "default" similarity metric Lewis discusses in his [1979]. On this view, we take Lewis's dictum that we should avoid widespread violations of law while maximizing the region of perfect match of fact to apply to logical law. We can determine the logical, mathematical, and metaphysical laws of a world in a manner analogous to how we determine the nomological laws of a world: we try to fit the world into a deductive system that best makes sense of what is true at the world. A world is similar, then, to the extent that it minimizes changes in law and maximizes matches of matters of fact.

In chapter 4, I discuss the nature of impossible worlds. Since the proposed solution makes use impossible worlds, a controversial extension of an already-controversial apparatus, it is necessary to explain what impossible worlds are and why they are ontologically cheap.

Possible worlds can be used to build structured propositions, which can in turn be used to build "structured worlds." An impossible world is simply a structured world: a set of structured propositions. Such a set need not be complete or consistent, so impossibilities can be represented straightforwardly by these worlds. Furthermore, these worlds are ontologically cheap, requiring belief merely in possible worlds and set-theoretic constructions of those worlds. In this chapter, I also discuss Lewis's argument against impossible worlds, and how to best understand what is true and false at these worlds.

In chapter 5, I apply the theory of counterpossibles defended in chapter 3 to the problem of metaphysical dependence. I hold that dependence is best understood in terms of

explanation and *counterpossible exclusion*. That is, A depends on B iff A explains B, and the following counterpossible is true: if one of A or B were to be false, then A would be false. The counterpossible exclusion claim adds the requisite hyperintensionality and asymmetry required to make the supervenience account plausible; it also suggests the right kind of modal relation between the grounding and grounded entities.

One can then use this analysis to evaluate claims of metaphysical dependence. I argue that this approach gives the intuitively correct results, albeit with a complication related to subject matters. In cases where the counterpossible is more difficult to evaluate, the analysis at least points toward where the difficulty comes from: evaluating these counterpossibles will ultimately be a matter of determining the logical and metaphysical laws of various impossible worlds, which in turn is a matter of determining what the best way of systematizing these worlds are. Cases where the dependence claims are more controversial are cases where we are less sure about how to understand facts related to explanation or the structure of the impossible worlds in question.

2. Counterfactuals

I have described counterpossibles as a kind of counterfactuals. Unfortunately, it is not entirely clear what counterfactuals are. They are often glossed along lines such as 'conditional sentences in the subjunctive mood.' If this is correct, then there is a clear syntactic test to determine whether a statement is a counterfactual: we can determine whether a sentence expressed a counterfactual by looking at its form.

These kinds of tests might be a guide for some cases, but it seems highly implausible to say that counterfactuals are usually expressed in this form. Philosophers and prescriptive grammarians often express counterfactuals in the subjunctive mood, but ordinary speech

generally eschews the subjunctive mood. Indeed, there are questions about what the subjunctive mood comes to, and whether this test would work across languages that do not seem to have anything rightly called a subjunctive mood or tense.

Indeed, many counterfactuals are not even expressed as conditionals. Consider Lewis's example, 'No Hitler, no atom bomb,' which expresses a counterfactual, even though one could not guess this merely by looking at the form of the sentence. One could multiply cases: 'Sure, but then they'd topple over,' is a counterfactual whose antecedent is not expressed, but is rather (presumably) anaphoric on some previous sentence in the discourse; 'Had you cut off their tails, they'd have toppled over' is a counterfactual that lacks the 'if/then' structure. Looking to surface syntax as a guide could be misleading.

A semantic account of counterfactuals seems far more fruitful than a syntactic account. A statement need not include words such as 'if' and 'would' (or their correlates in other natural languages) to be a counterfactual, but rather needs to express the right kind of proposition. That is, a counterfactual is a kind of statement that expresses what would occur, given some state of affairs. Counterfactuals are so-called because the state of affairs in question does not actually obtain; evaluating a counterfactual is a process of determining what would happen if the antecedent were true (though it is not). If the antecedent of some counterfactual is *true*, then the utterance of the counterfactual is infelicitous; this does not, however, mean that the utterance is not a counterfactual in the relevant sense.

But what would happen if the antecedent conditions were impossible?

3. Impossibility

The other part of our definition of counterpossibles that needs unpacking is the notion of impossibility. We say that a counterpossible is a counterfactual when its

antecedent is impossible, but when is something impossible? The nature of impossibility is itself a thorny issue, so here I will merely highlight different kinds of impossibility that might be employed in some counterpossibles.

It is important to note that words we use to express possibility and impossibility, such as 'can' and 'must' and so on, are themselves highly context-dependant with respect to their application (see, for example, Kratzer [1977]). We might truly say that it is not possible for me to run a mile in six minutes, but we are not thereby committed to the claim that there is no possible world where I can run a six-minute mile. Rather, we deem those worlds to be irrelevant in the given context. If we change the context, perhaps by bringing up the possibility of having devoted some time to training and practice, we would no longer be inclined to say that I couldn't run a six-minute mile. We also sometimes restrict the set of worlds in question when discussing belief ('John must be in Boston by now'), physical laws ('The bowling ball can't float in mid-air'), and normativity ('You cannot torture children for fun.') But these are not the modal notions of interest here; the notion of impossibility in play is grounded by the ways things might turn out, not what we believe, or what the best physical theory of this world is, or what we should do.

The notion of possibility relevant to a study of counterpossibles is *unrestricted* possibility. When I say that something is impossible unrestrictedly, I do not mean merely that it is impractical, or would violate some law (but see below); rather, I mean that it is *absolutely* impossible. I mean that there is no possible world where it is true; I mean that the scenario described is not a genuine possibility. Once we have decided that something is metaphysically impossible, we cannot then change the context in such a way that it would then seem possible after all. One could attempt to get a better understanding of what is

meant by impossibility, in the relevant sense, by examining some different ways in which the antecedent of a counterfactual could be impossible.

A conditional whose antecedent is not logically possible could be called a *counterlogical*. I assume the truth of classical logic, so the antecedents of counterlogicals would express a formal contradiction. This could be some statement of the form P and not-P, or some statement that entails something of the form P and not-P. This formulation could be modified, however, to apply to any logical theory as long as it admits of valid and invalid rules of inference.

A conditional whose antecedent is impossible with respect to a correct philosophical theory is a *countermetaphysical*. This occurs when the antecedent asserts the truth of a false philosophical theory whose truth value is a matter of necessity. 'If properties were universals, then redness would be repeatable' expresses a countermetaphysical.

A conditional whose antecedent is analytically false is a *counteranalytical*. This occurs when the antecedent expresses something that is semantically or conceptually false. 'If some bachelors were married, then they'd have wedding rings' is a counteranalytical.

A conditional whose antecedent is not mathematically possible is a *countermathematical*. This occurs when the antecedent is false according to arithmetic, set theory, geometry, and so on. 'If 9 were prime, then it would not be divisible by 3' expresses a countermathematical.

There are also other notions of impossibility in the literature, such as the proposal that laws of nature are genuinely necessary, as in Shoemaker [1998], and that what has occurred in the past is genuinely necessary, as in Prior [1957]. These proposals could be understood as kinds of metaphysical necessity in the sense of countermetaphysicals, or as some *sui generis* kind of unrestricted necessity.

Perhaps some of these categories collapse into others. If the logicist or neo-logicist project is successful, for example countermathematicals will reduce to counterlogicals. If false theories in metaphysics and ethics somehow entail contradictions, then they will reduce to counterlogicals. If these false theories are conceptually or linguistically defective in such a way that the phenomena they analyze fail to match our initial concepts of them – perhaps it is constitutive of properties that they be repeatable, or constitutive of rightness that the organ harvest cannot be right – then countermetaphysicals will reduce to counteranalyticals.

This discussion, of course, is quite inconclusive. Determining which kinds of necessities reduce to other kinds of necessities is a project beyond the scope of this dissertation. Nonetheless, it may be helpful, as one is attempting to evaluate analyses of counterpossibles, to keep in mind the various kinds of counterpossibles, and how they might relate to each other. If a non-trivial account of these conditionals is possible, it must be able to account for these various kinds of counterpossibles. A theory should be judged on how well it matches our intuitive judgments about the truth of these sentences, as well make sense of how these sentences are used in math, logic, and philosophy.

CHAPTER 1

COUNTERPOSSIBLES ARE NOT VACUOUS

1.1 Introduction

Counterfactuals with necessarily false antecedents, or *counterpossibles*, pose problems for standard treatments of counterfactual conditionals. We seem to have robust intuitions, at least in many cases, about the truth of counterpossibles. Consider, for example:

(1) If mereological nihilism were true, then there would be no tables.

(2) If mereological nihilism were true, then there would be tables.

(3) If I were a horse, then I would have hooves.

(4) If I were you, I wouldn't eat that.

(5) If wishes were horses, beggars would ride.

(6) If Hume had squared the circle in secret, then giraffes would have wings.

(7) If some bachelor were married, then it would be false that some bachelor were married.[1]

These sentences seem natural enough, either in philosophical or ordinary contexts. Indeed, (5) is in at least somewhat common usage as a proverb. We have fairly firm intuitions that (1) and (3) are true, that (4) is can be asserted in some contexts, and that (2), (6), and (7) are false. But the standard account of counterfactuals, from Lewis [1973] and Stalnaker [1968],

[1] Perhaps some of these examples are not counterpossibles. For a conditional to be counterpossible, the antecedent must be impossible, but determining which antecedents are possible and which are impossible requires a fully-developed account of possibility, as well as a settling certain logical, mathematical, and metaphysical questions. Clearly, (1) and (2) are not counterpossibles if mereological nihilism *is* true; (3) and (4) need not counterpossibles if Kripkean essentialism is false; (4) is not a counterpossible if the proper analysis of 'If I were you' is by relevant properties and roles instead of identity, and so on. I do not intend to take a stand on such issues presently. If these examples are not counterpossibles, they may be replaced with other examples whose antecedents are impossible.

does not deliver these results. For reasons discussed below, Lewis and Stalnaker hold that counterpossibles are vacuous. Such a view, however, is wrong for several reasons. First, it does not respect our linguistic intuitions, and any adequate theory of counterfactuals must respect our firm intuitions about their truth. Second, such a view does not respect philosophical practice, insofar as these sentences are used to draw out the consequences of various philosophical, logical, or mathematical theories. Third, a more careful consideration of the use of 'might' and 'would' counterfactuals shows that the reasons Lewis and Stalnaker give for treating counterpossibles vacuously are in tension with their treatment of the logic of 'might' and 'would.' Indeed, only an account of counterpossibles according to which they are non-vacuous has the tools to explain the behavior of counterpossibles with respect various plausible logical principles.

1.2 Lewis, Stalnaker, and Counterfactuals

At least in English, counterfactuals are generally expressed as conditionals in the subjunctive mood. Yet this is obviously not a definition of what it is to be a counterfactual; as discussed in the introduction, the most obvious problem is that not all counterfactuals are expressed as subjunctive conditionals, even in English. Consider, for example, the bumper sticker that reads 'No farms, no food:' even if it does not look like a conditional, it is clearly meant to be evaluated as a claim about what would happen if there were no farms. It is not helpful to think of counterfactuals in purely syntactic terms. Rather, counterfactuals are propositions that express a certain kind of modal condition between the antecedent and consequent; they make the claim that, at least in some relevant space of possibility, the consequent of the conditional is true where the antecedent is. Because of this modal connection, counterfactuals are often written as $A \mathrel{\Box\!\!\rightarrow} C$, where A and C are propositions

(intuitively, the antecedent and consequent, respectively) and the connective $\Box\!\!\rightarrow$ stands for the counterfactual conditional.

Counterfactuals generally carry the presupposition that the antecedent is false. If I were to utter, 'If I had missed the bus, then I would be late to class,' one would generally assume that I had, in fact, caught the bus. If the antecedent of a counterfactual is true – if, for example, I utter the above sentence even though I *did* miss the bus – then what I said was infelicitous. What I had said is not necessarily false, however, and it is not clearly no longer a counterfactual.

There is, of course, more to say about these issues, but they need not be settled for our present purposes. One could take nearly any view of what a counterfactual is and still face the problem of how to evaluate counterpossibles. Whether a counterfactual is best construed syntactically or semantically or whether its antecedent *must* be false for it to be a genuine counterfactual is a question that should be settled in giving a complete account of counterfactuals. However these questions are answered, one can still formulate counterfactuals with necessarily false antecedents.

According to Lewis and Stalnaker, a counterfactual is true (roughly) iff its consequent is true in the most similar, or "nearest" possible world or worlds where the antecedent is true. According to Lewis [1973], a counterfactual is true iff no world where the antecedent is true and the consequent is false is "closer" to the base world than any world where the antecedent and consequent are both true. A world is "closeby" to the degree that the world is, in some contextually relevant sense, similar to the base world. Lewis gives some of the logical properties of similarity in his [1973], and discusses a particular metric of similarity in his [1979]. The truth conditions are stated in the way that they are in order to allow for the falsity of both the *limit assumption,* according to which there

is *at least* one possible world where the antecedent is true that is closest to the actual world, as well as the *uniqueness assumption*, according to which there is *at most* one possible world where the antecedent is true that is closest to the actual world. Since Stalnaker [1968] accepts both of these assumptions, his statement of the truth conditions of counterfactuals is simpler: a counterfactual is true iff the consequent is true in the *selected* world where the antecedent is true. For Stalnaker, a selection function determines the world of evaluation, and if the consequent is true in that world of evaluation, then the counterfactual is true.

I will assume that the limit assumption is true, but that the uniqueness assumption is false. Not much hangs on this choice, and nothing discussed herein depends on these assumptions; one could easily translate the analyses to be given to include or exclude either of these assumptions. I make this choice largely because the uniqueness assumption seems so implausible,[2] and because stating the truth conditions of counterfactuals is far easier with the limit assumption. Thus, when I discuss the 'Lewis-Stalnaker' view, I really mean a view that neither of them had: a counterfactual is true iff the consequent is true in all the closest possible worlds where the antecedent is true.

What if the antecedent of the conditional is impossible? When there are no possible worlds where the antecedent is true, there are obviously no "nearby" or "selected" possible worlds where the antecedent is true at which we might attempt to determine the truth of the consequent. For Lewis and Stalnaker, all counterpossibles are vacuously true. In light of the previous examples, this seems quite unintuitive. Sentences such as (1)-(7) seem to be meaningful, and capable of being true or false; indeed, some of them *are* false! Our project, then, is to motivate, describe, and defend an account of counterpossibles according to which their truth conditions are not vacuous.

[2] But see Stalnaker [1980]

1.3 Counterpossibles, Language, and Philosophical Practice

The examples above are meant to provide intuitive evidence for the view that counterpossibles cannot all be trivially true. Many of these kinds of sentences are used in natural language, and would not strike a speaker as being vacuous. These sentences clearly have content, both in the sense that they seem to be genuinely meaningful and in the sense that they could be either true or false. Analyzing these sentences as vacuous is a mistake, even if there is a pragmatic story at hand to explain why they *seem* non-vacuous. Ignoring our intuitions about truth conditions in a wide range of cases undercuts our original project of providing plausible truth conditions for conditionals in general.

Furthermore, counterpossibles are regularly used in philosophy, math, and logic: we judge the adequacy of necessarily true or necessarily false theories in metaphysics, ethics, and epistemology by what follows from their truth. Sentences such as (1) could be used in part of an argument against a particular philosophical view; we might also have reason to accept sentences such as 'If Platonism were true, then we would have no way of knowing about numbers' or 'If Utilitarianism were true, then the organ harvest would be morally obligatory' or 'If intuitionism were true, then the law of double negation elimination would fail.' We are invited to accept sentences such as these as (non-vacuously) true, and take their truth as having some force with respect to how we understand and evaluate certain theories. That is, counterpossibles are used to express the commitments of theories, even in cases where these theories are acknowledged to be necessarily false. A necessarily false theory is not committed to *everything*, but it might be committed to certain unacceptable consequences. Insofar as these consequences are deemed unwelcome, we are inclined to reject the theory in question. Even if arguments about the commitments of various necessarily false theories are

not conclusive, counterpossibles whose antecedents suppose their truth cannot be vacuous without rendering meaningless broad swaths of philosophical discussion. Since the meaningfulness of counterpossibles is so important to philosophical and mathematical practice, it would be a mistake to offer an analysis according to which counterpossibles are vacuous.[3]

The view that counterpossibles are all trivially true leads to some rather strange results when we form counterpossibles about the semantics of counterpossibles. Consider the following example:

(8) If some counterpossibles were false, then Lewis would be right about counterpossibles.

For Lewis, (8) is a counterpossible, and therefore (trivially) true. The proper analysis of conditionals in a language, after all, is a necessary truth, even if the facts about natural language are contingent. But (8) is obviously false – Lewis's theory of counterpossibles, after all, would be wrong if there were false counterpossibles! The antecedent of (8) gives the very condition that would falsify Lewis' account, so it would be bizarre to think that (8) is true.[4]

---

[3] These considerations might also be related to the question of how to understand arguments by *reductio ad absurdum*. In such arguments, we reason from assumptions that are (at least in conjunction with other premises) contradictory. Nonetheless, we are able to reason from these inconsistent propositions in a fairly robust fashion; certain *reductio* deductions are licit, but others are not. To be sure, however, at least formal arguments that make use of *reductio ad absurdum* function in a very syntactic fashion: we can make an assumption, use known rules of inference to show that the assumption (and possibly other premises) lead to a formal contradiction, and then accept the negated assumption. It is not clear that counterpossibles in general behave in such a fashion. There is further discussion of this point in chapter 2.

[4] There is also another kind of argument we could use against Lewis and Stalnaker that is more clearly *ad hominem*. Lewis and Stalnaker *clearly* understand counterpossibles in a non-vacuous fashion because they make use of them! Consider, for example, Lewis in his [1986:25]: "If, *per impossibile,* the method of dominance had succeeded in ranking some false theories above others, it could still have been challenged by those who care little for truth." The details of this particular argument are not important, but it does seem that Lewis thinks he can reason about what would happen if some impossible condition were to hold.

Furthermore, we cannot lay the blame for this problem simply at the feet of Lewis and Stalnaker's approach to counterfactuals, and use this as an argument against possible worlds approach to counterfactuals in general. Other approaches to counterfactuals deliver the same result. Non-worlds-based accounts, such as [Goodman 1947], hold that a counterfactual is true whenever the antecedent, along with other background propositions, entails the consequent. Whether or not this approach is ultimately preferable to the possible worlds account as a treatment of counterfactuals, it too is unable to deliver our intuitive judgments about counterpossibles. If the antecedent is impossible, then (presumably) it entails some contradiction, a statement of the form *P and not-P*.[5] And if the antecedent and other statements entailed by it include a formal contradiction, then, by the principle of explosion, any consequent will come out true. For example, if our antecedent is 'I am a horse,' then it, as well as certain other metaphysical theses, such as that 'I' is a rigid designator, and that in every world where I exist, I am a human and not a horse, entail that both I am a horse and I am not a horse. And from this, we can deduce any consequence we would like. Thus, for logical entailment-based approaches, counterpossibles will be trivially true, since any consequent is entailed by a contradiction.[6]

Of course, this does not mean that *any* theory of counterfactuals will make counterpossibles trivially true; nonetheless, since the other approaches to counterfactuals face the same difficulties with respect to impossible antecedents, we cannot merely take the problem of counterpossibles as an argument against Lewis and Stalnaker's views in particular. Indeed, the fact that both worlds-based approaches and entailment-based

[5] This may conflate the distinction between logically and metaphysically impossible antecedents. In either case, this argument certainly holds for counterpossibles with logically impossible antecedents.
[6] The same argument can be lodged against more complex versions of the entailment account, such as that of Kvart [1986].

approaches are incapable of giving non-vacuous truth conditions to counterpossibles suggests that the problem is more general than it might have seemed.

1.4 'Might' Counterpossibles

There are now several reasons to reject a vacuous analysis of counterpossibles on the table. But the problem is even worse than this: given their treatment of 'might' counterfactuals, Lewis in particular is in an especially poor position to claim that counterpossibles are vacuous.

Lewis and Stalnaker differ with respect to their treatment of 'might' counterpossibles (represented as $A \Diamond\!\!\rightarrow C$, and read as 'If A were the case, then C might be the case.'). For Lewis, the 'might' counterfactual is defined as the dual of the 'would' counterfactual: $A \Diamond\!\!\rightarrow C$ is equivalent to $\neg(A \Box\!\!\rightarrow \neg C)$. Intuitively, a 'might' counterfactual is true when the consequent is true in *some* of the nearest worlds where the antecedent is true; if there are no possible worlds where the antecedent is true, the conditional is false. Since all 'would' counterpossibles are true for Lewis, all 'might' counterpossibles are false.

According to Stalnaker [1980] , 'might' counterfactuals are to be analyzed as 'would' counterfactuals embedded in an epistemic modal. $A \Diamond\!\!\rightarrow C$ is roughly equivalent to 'For all I know, $A \Box\!\!\rightarrow C$.' Since $A \Box\!\!\rightarrow C$ is always true whenever A is impossible, it *cannot* be ruled out by anything that I know. Nothing that any agent knows can be incompatible with something that must be true, so nothing any agent knows could ever rule out the truth of a 'would' counterpossible. Since 'might' counterpossibles are epistemic modals applied to

necessarily true propositions, 'might' counterpossibles will be vacuously true on Stalnaker's account.[7]

Stalnaker's analysis of 'might' counterfactuals introduces a somewhat surprising asymmetry between 'might' and 'would' counterfactuals. A 'would' counterfactual is analyzed by determining which possible world is closest to the actual world (in his terms, 'selected'), so a 'would' counterfactual could be true or false regardless of what is known by the speaker or evaluator. A 'might' counterfactual, on the other hand, uses an epistemic modal, so the truth conditions of 'might' counterfactuals always rely on what is known by one or more participants in a conversation. While uses of 'might' often suggest a kind of epistemic modality, there are many cases where they do not: consider, for example, 'I am a philosopher, but I might have been a lawyer instead.' Insofar as we think that 'might' and 'would' counterfactuals are closely related, we should be hesitant to adopt an approach that treats them differently.

1.4.1 'Might' Counterpossibles and Entailment

Lewis [1973] , Williamson [2007], and others worry that our judgments about the truth values of counterpossibles, such as those mentioned above, do not stand up to scrutiny. When we consider counterpossible situations in the proper way, perhaps, we realize that "anything goes" when we reason about impossible antecedents. Consider (6), for example. We are inclined to judge (6) to be false, and take this judgment as evidence against Lewis's account. But what would it take for Hume to have squared the circle? *A priori* truths about Euclidean geometry would have to be vastly different! And if we waive certain

[7] To be sure, Stalnaker does have some tools at his disposal to explain away some of the discomfort associated with the view that that we believe all necessary truths, as in his [1987]. The second chapter discusses an approach to the truth conditions of counterpossibles that is inspired by some of these views.

fundamental truths about Euclidean geometry, who knows what might follow? Perhaps (6) is true after all (though not in any interesting way), since in these worlds, anything at all is true! Once we consider what things would be like if mathematical or logical truths differ, it seems hard to sustain our initial reading of the counterpossible over the vacuous reading. Indeed, the vacuous approach would then get the truth conditions right after all: when we consider what the world would be like if something impossible were to happen, anything at all would be true!

This response is unsatisfying for several reasons. First, it is not at all clear that these considerations compel us to abandon our intuitions about the truth conditions of counterpossibles; instead, we should hold that these considerations compel us to accept that 'anything goes' in *some* contexts of evaluation. All that the defender of the trivial account of counterpossibles has shown is that there is at least one reading of (6) according to which we are hesitant to assert that it is false. But it is no surprise that counterpossibles have more than one admissible reading. Consider an example from [Jackson 1977]:

(9) If I had jumped out the window, I would have injured myself.

We take (9) to be true in most contexts. The nearest world where I jump out the window is a world where I land on the concrete and injure myself. But another reader might balk at this conclusion: I am a reasonable person, and I would not jump out the window if I thought I might injure myself. I would only jump out a window if I had (say) placed a net beneath it beforehand. Now, (9) seems to be false, since the nearest world where I jump out the window is a world where I place a net beneath the window and land safely. We need not let these considerations drive us into a deep skepticism about the status of counterfactuals; rather, we merely realize that how we evaluate counterfactuals depends on context. On the first pass, we sort worlds in a manner that places importance on the height of the window

and the ground, and thus in the nearby worlds, I injure myself. On the second pass, my cautious nature has become salient, and thus we sort worlds in a manner that places greater importance on my fear of injury, and find that in these nearby worlds, I am uninjured because I have taken precautions.[8] Neither reading is *the* correct analysis of (9); our judgment of the truth of (9) varies depending on what kinds of considerations have been brought to our attention.

Something very similar occurs in (6). At first pass, (6) seems false, since Hume's squaring of the circle in secret has no obvious effect on the anatomy of giraffes. We sort worlds in a manner that places importance on a match of facts about the world. On the second pass, however, we re-sort worlds, and now place importance on the deductive closure of geometry. Once we are forced to admit that something impossible has happened, we are simply not sure what to think. Perhaps now, the nearest antecedent world is the "explosion" world where every proposition is true. Neither way of sorting is *the* correct way to sort worlds, and neither of these is *the* correct analysis of the counterpossible absent of any contextually salient considerations; different contexts merely establish different similarity metrics, and so the non-vacuous reading of (6) remains licit in the context in which it was presented. A counterpossible is a kind of counterfactual, and thus our judgments of the truth values of counterpossibles are similarly malleable and sensitive to context. There are, no doubt, contexts in which we might say that 'anything goes' in a counterpossible situation, but this does not mean that anything goes in counterpossible situations *tout court,* or even in some 'standard' context. We need not conclude that counterfactuals are vacuous because of this shiftiness, and we need not conclude that counterpossibles are vacuous because of this

[8] Indeed, on this reading of (9), the salience of my fear of injury is great enough to allow "backtracking" in the analysis of the counterfactual. We change some facts about the past (*viz.,* my placing a net beneath the window) in order to get the proper reading of the counterfactual. What's important for this example is merely the ease with which we re-order the similarity of worlds.

shiftiness. Since Lewis' analysis of counterpossibles never allows counterpossibles to have non-vacuous truth-values, we have reason to doubt Lewis' analysis.

Consideration of 'might' counterpossibles, however, gives us a second reason to reject this argument. Lewis and Williamson claim that careful consideration of counterpossibles suggests that anything follows from an impossibility. Recall, however, that at least for Lewis, all 'might' counterpossibles are *false*. If we think that all 'would' counterpossibles are true because we are inclined to shrug our shoulders in the face of an impossibility, how can we account for the *falsity* of 'might' counterpossibles? There may be contexts where it seems plausible to say that anything at all would be true if Hume had squared the circle, but it seems harder to imagine any context where *nothing* at all might be true if Hume had squared the circle. If the strangeness of 'would' counterfactuals inclines us to say that anything at all would be true in these cases, how could it also incline us to say *nothing* at all *might* be true in these cases?

Defenders of the vacuous approach sometimes claim that the result that all counterpossibles are trivial is not as bad as it seems. The counterpossibles that we want to assert (such as (1)) will be true and assertible, while the counterpossibles we don't want to assert (such as (2)) may be true, but some pragmatic strategy could explain why they are not assertible (see Lewis [1973]). Of course, such a pragmatic strategy would need much more detail to be plausible, and it seems difficult to say what kind of Gricean mechanisms would deliver the desired result. Even leaving the details of this strategy aside, it does not account for the *falsity* of 'might' counterpossibles. 'Might' counterpossibles that we want to assert (such as, 'If mereological nihilism were true, then we might never find out') will be false. To be sure, we could attempt to develop some additional pragmatic account for why sentences such as these are false but nevertheless assertible, but this would be still more difficult than

accounting for why some sentences are true but not assertible. The vacuous approach as advanced by Lewis and others requires not only a pragmatic theory according to which 'would' counterpossibles that are intuitively false are actually true but not assertible, but also an error theory according to which 'might' counterpossibles that are intuitively true are actually false but assertible. These considerations, of course, are not decisive, and it is possible that some elegant pragmatic strategy could explain the patterns of assertibility in a satisfying fashion. Nonetheless, such a view would still be an error theory, and would still require significant pragmatic machinery in order to account for intuitive judgments. Other things being equal, it would be far better to get the semantics of the conditionals right, rather than have the heavy lifting done by poorly-understood pragmatic processes.

A further argument offered in defense of the vacuous approach is also in tension with the falsity of 'might' counterpossibles. Lewis [1973] and Wierenga [1998] argue that if A logically entails C, the counterfactual $A \mathrel{\Box\!\!\rightarrow} C$ is true. Call this principle Entailment:

(Entailment) If $A \rightarrow C$, then $A \mathrel{\Box\!\!\rightarrow} C$

Given the "nearest possible worlds" semantics for counterfactuals, it seems that classical entailment guarantees counterfactuality: if A entails C, then all A-worlds are C-worlds, and so the nearest A-worlds are C-worlds. And since, in classical logic, a contradiction entails anything, $A \mathrel{\Box\!\!\rightarrow} C$ will be true for any C as long as A is a contradiction.[9] Thus, 'would' counterpossibles are all vacuously true, and the vacuous approach to counterpossibles is vindicated.

[9] We need not appeal to the possible worlds semantics of counterfactuals to defend Entailment, though it is as a good way to illustrate the principle. Entailment seems highly plausible, and is also guaranteed by other approaches to counterfactuals. It is clearly valid on entailment approaches to counterfactuals, such as those offered by Chisholm [1946] and Goodman [1947], which are the leading contender to possible worlds accounts of counterfactuals.

It is important to note that not all impossible antecedents clearly entail a logical contradiction. For example, counterpossibles with metaphysically impossible antecedents might be immune to the argument from Entailment sketched above. It may be that a theory of properties according to which they are immanent universals is necessarily false, but that such a theory does not entail any kind of formal contradiction. Indeed, it would be incredible if every false theory in metaphysics, epistemology, or ethics were such that, under analysis, one could derive a contradiction from a statement of the theory: Frege's set theory is perhaps the only serious philosophical theory that can be shown to result in logical inconsistency. It is much more plausible that the argument from Entailment is directed toward the proper subset of counterpossibles whose antecedents do entail a contradiction.[10]

Even setting this caveat aside, the falsity of 'might' counterpossibles is difficult for the vacuous theorist to account for. After all, it seems just as plausible to say that if A logically entails C, the 'might' counterfactual $A \Diamond\!\!\rightarrow C$ is true. Call this principle 'Might' Entailment:

('Might' Entailment) If $A \rightarrow C$, then $A \Diamond\!\!\rightarrow C$

If all A-worlds are C-worlds, then some of the nearest A-worlds are C-worlds. Since a contradiction entails everything, $A \Diamond\!\!\rightarrow C$ should be *true* for any C, as long as A is a contradiction. Indeed, as long as we think that the 'might' counterfactual is weaker than the 'would' counterfactual, we should think that 'Might' Entailment is valid because Entailment is valid. However, even though A will entail C, Lewis holds that $A \Diamond\!\!\rightarrow C$ will be *false* in every instance where A is a contradiction. Thus, it seems difficult for Lewis and Wierenga to appeal to Entailment to defend the vacuous approach to 'would' counterpossibles, when the

[10] For further discussion of this point, see Zagzebski [1990].

equally plausible 'Might' Entailment principle is incompatible with the vacuous approach to 'might' counterpossibles under consideration.

Such a *tu quoque* argument, however, exposes an awkward tension with respect to the status of Entailment. The Entailment principle was presented as quite plausible, yet the Lewisian is unable to accept the equally-plausible principle of 'Might' Entailment. For all this, it is still unclear what to make of the intuitions that made Entailment seem plausible in the first case. If there is tension within Lewis's view over the status of the 'Might' Entailment, there is also tension within the non-vacuous theorist's view over what to make of the plausibility of Entailment and 'Might' Entailment. Of course, one could accept the failure of these principles with some degree of equanimity: perhaps consideration of our intuitions of certain examples is sufficient to reject Entailment as well as 'Might' Entailment. Yet one could also attempt to reformulate weaker, more acceptable versions of the principles:

(Entailment*) If $A \rightarrow C$ and not: $A \rightarrow \neg C$, then $A \Box\!\!\rightarrow C$

('Might' Entailment*) If $A \rightarrow C$ and not: $A \rightarrow \neg C$, then $A \Diamond\!\!\rightarrow C$

Entailment* (and 'Might' Entailment*) allows one to reject the claim that, if some proposition entails everything, then if that proposition were true, everything would (might) be true. Entailment* and 'Might' Entailment* is silent about counterfactuals if the antecedent entails everything; if some proposition entails something and its negation, then all bets are off with respect to which counterfactuals that make use of that proposition as an antecedent are true. The non-vacuous counterpossible theorist can endorse these weaker principles and blunt the force of this version of the 'anything goes' argument.

However we ultimately settle issues related to the Entailment principle, the falsity of 'might' counterpossibles stands in tension with many of the defenses offered for the vacuity

of counterpossibles. We may be inclined to think that 'would' counterpossibles are all true because 'anything goes' in counterpossible situations, or because logical entailment entails counterfactuality; we might think that the result that all 'would' counterpossibles are true is not as unintuitive as it seems, so long as a pragmatic account could tell us why many 'would' counterpossibles are not assertible. But if 'might' counterpossibles are false, then we are at a loss to explain why it is also true that 'nothing goes' in counterpossible situations, or how logical entailment fails to entail 'might' counterfactuality, or how a pragmatic strategy could deliver that result that a wide range of sentences are false but assertible.

Given these considerations, then, we might think that Stalnaker's treatment of counterpossibles has a clear advantage over Lewis's treatment. If 'might' counterpossibles are all true, then the arguments in favor of vacuity given above remain plausible. Given some impossibility, we could still hold that 'anything goes' (since all 'would' counterpossibles are true) and 'anything might go' (since all 'might' counterpossibles are true as well). Since all 'might' counterpossibles are false, we need not offer some error theory about the assertibility of any false counterpossibles. And, finally, Stalnaker is able to endorse both Entailment and 'Might' Entailment. These advantages, of course, come at the cost of denying the duality of 'might' and 'would' counterfactuals, it is nonetheless clear that the replies given to the arguments that counterpossibles are vacuous target Lewis's semantics, and not Stalnaker's. Though I have given independent reasons above for favoring a non-vacuous account of counterpossibles, it seems that a vacuous account along Stalanker's lines is at least more plausible than a vacuous account along the lines of Lewis's.

1.4.2 'Might' and 'Would'

The definition of the 'might' counterfactual in terms of the 'would' counterfactual as described by Lewis delivers very unintuitive results when applied to counterpossibles. Consider, for example, the following counterpossibles:

(10) If mereoloigcal nihilism were true, then there would (still) be tables.

(11) If mereoloigcal nihilism were true, then there would not be any tables.

(12) If mereoloigcal nihilism were true, then there might (still) be tables.

(13) If mereoloigcal nihilism were true, then there might not be any tables.

According to Lewis (and Stalnaker), (10) and (11) are both true; according to Lewis (but not Stlanker) (12) and (13) are both false. These results are unacceptable for several reasons. First, it seems very clear that our intuitive judgments hold that (11) and (13) are true, and (10) and (12) are false. Lewis and Stalnaker cannot deliver this result. Second, it seems odd that sentences of the form (10) and (11) could both be true; this is because even if mereological nihilism were true, it can't be the case that a pair of contradictory things would be true. Third, it seems odd that sentences of the form (12) and (13) could both be false; this is because even if mereological nihilism were true, it can't be that there is *nothing* true to say about what might be the case. Finally, it seems odd that sentences of the form (10) could be true while (12) is false, or (11) true and (13) false. This is because it seems odd to say that something *would* happen under certain counterfactual conditions, but also hold that it's not the case that it *might* happen under those conditions. We will examine each of these claims below.

1.4.3 Would Law of Non-Contradiction (WLNC)

One plausible principle for counterfactuals is the *would law of non-contradiction* (WLNC). According to this law, there cannot be a pair of 'would' counterfactuals with the same antecedent, but contradictory consequents.

(WLNC) Not: $A \mathrel{\Box\!\!\rightarrow} C$ and $A \mathrel{\Box\!\!\rightarrow} \neg C$

Intuitively, given some antecedent condition, it cannot be the case that some pair of contradictory things will be the case. Possible worlds, after all, are consistent. It cannot be the case that both C and not-C are true in *any* world, and therefore C and not-C cannot be true in *all* nearby A-worlds. WLNC, then, is valid on Lewsis's account of counterfactuals with possible antecedents.[11]

Counterfactuals with impossible antecedents, however, will introduce widespread failures of WLNC for Lewis and Stalnaker. Since *all* 'would' counterfactuals with impossible antecedents are true, both members of some pair of counterpossibles of the form $A \mathrel{\Box\!\!\rightarrow} C$ and $A \mathrel{\Box\!\!\rightarrow} \neg C$ will be true when A is impossible. Thus, WLNC will fail for every such pair of counterpossibles.

1.4.4 Might Law of Excluded Middle (MLEM)

Another plausible principle for a logic of counterfactuals is the *might law of excluded middle* (MLEM). According to this law, for any pair of 'might' counterfactuals with the same antecedent and contradictory consequents, at least one counterfactual of the pair is true.

(MLEM) $A \mathrel{\Diamond\!\!\rightarrow} C$ or $A \mathrel{\Diamond\!\!\rightarrow} \neg C$ is true

[11] I assume throughout that the set of nearby possible worlds is held fixed in a pair of counterfactuals. There are apparent failures of WLNC in pairs of counterfactuals such as, 'If Caesar were in charge in Korea, he would have used catapults' and 'If Caesar were in charge in Korea, he would have used nuclear weapons (and not catapults).' But, presumably, we switch contexts between these two sentences, and thus the set of worlds that are deemed to be the nearest worlds changes. This kind of case does not represent a genuine failure of WLNC.

Intuitively, some antecedent condition might give rise to the truth of some proposition, and it might give rise to its negation, but it can't be the case that it might lead to *neither*. Given some antecedent condition, we should expect that some proposition might follow, or the negation of that proposition might follow, but it would be odd if *neither* of a pair of mutually exhaustive possibilities might be obtain. Possible worlds, after all, are complete: MLEM will be valid for Lewis whenever the antecedent is possible, since either C or not-C will be true in any world, and therefore C or not-C will be true in any of the nearby A-worlds.

According to Lewis's semantics, MLEM is not valid for counterfactuals with impossible antecedents. Since *all* 'might' counterfactuals with impossible antecedents are false, both members of some pair of counterpossibles of the form $A \Diamond\!\!\rightarrow C$ and $A \Diamond\!\!\rightarrow \neg C$ will be false whenever A is impossible. Thus, MLEM will fail for every such pair of counterpossibles.

1.4.5 Would Implies Might Principle (WIMP)

Another highly plausible principle concerns the relation between 'might' and 'would' counterfactuals. Specifically, a 'would' counterfactual implies its corresponding 'might' counterfactual.

(WIMP) If $A \Box\!\!\rightarrow C$, then $A \Diamond\!\!\rightarrow C$[12]

It follows from the fact that if some antecedent conditions obtained, then some consequent would be true that if those antecedent conditions obtained, then that consequent *might* be true. That is, if something *would* happen given some counterfactual circumstance, it follows that it *might* happen given those circumstances. It seems obvious that if it were true that kangaroos would topple over if they lacked tails, we could infer that kangaroos *might* topple

[12] WIMP is also discussed in Vander Laan [2007].

over if they lacked tails. This principle is also valid for Lewis, at least when the antecedent is possible: if *all* nearby A-worlds are C-worlds, then *some* nearby A-worlds are C-worlds.

According to Lewis's approach, WIMP is never true when the antecedent is impossible. All 'would' counterpossibles are true, and all 'might' counterpossibles are false. This means that any counterpossible of the form $A \Box\!\!\rightarrow C$ will be true when A is impossible, regardless of the content of C, since 'would' counterpossibles are all true. However, the corresponding 'might' counterfactual of the form $A \Diamond\!\!\rightarrow C$ will never be true, since might counterpossibles are all false. Thus, on this view, WIMP will never hold when the antecedent of the pair of conditionals is impossible.[13]

1.4.6 True Mights and False Woulds

According to Lewis's analysis, all 'would' counterpossibles are vacuously true and all 'might' counterpossibles are vacuously false, but this choice is somewhat arbitrary. We could also hold that all counterpossibles are true, as Stalnaker and Wierenga do[14], or that all counterpossibles are false. Doing so, however, would not get us out of the problems discussed above. If all counterpossibles were true, then we would still have to contend with

---

[13] One might think that this discussion is less compelling when the modal operators are understood as quantifiers over worlds. We can, for example, restate WIMP in terms of quantifiers and possible worlds by saying that if all of the nearest A-worlds are C-worlds, then some of the nearest A-worlds are C-worlds. Yet when there are no A-worlds at all, as will be the case in counterpossibles, this principle simply fails. After all, it can be true that all A-worlds are C-worlds if there are no A-worlds, but it cannot be true that some A-worlds are C-worlds if there are no A-worlds; the universal quantifier does not have existential import, but the existential quantifier does. But it seems like this response simply avoids the intuition behind WIMP in favor of a particular formalism of modal logic: WIMP is plausible quite apart from any understanding of necessity as a universal quantifier or the question of the existential import of the universal quantifier. WIMP is plausible because, in natural language, 'would' is stronger than 'might.' It would be absurd for a 'might' claim to be true and a 'would' claim to be false. If this is correct, then the modal operators in this kind of context are not best captured as classical quantifiers over worlds. Perhaps they are not quantifiers at all, or perhaps the 'would' is a universal quantifier with existential import.

[14] Wierenga [1998] entertains a might operator that would make WIMP and MLEM valid. According to him, $A \Diamond\!\!\rightarrow C =_{df} \neg (A \Box\!\!\rightarrow \neg C)$ or $A \Box\!\!\rightarrow C$.

violations of WLNC. If all counterpossibles were false, we would still have to contend with violations of MLEM, Entailment, and 'Might' Entailment. In both cases, we would also have to reject the duality of 'might' and 'would' counterfactuals.

We could, however, preserve the duality of the two counterfactuals while answering most of the logical issues raised above by making 'would' counterpossibles false, and 'might' counterpossibles true.[15] If this is the case, worries about WLNC, MLEM, and WIMP vanish. If 'would' counterpossibles are all *false,* then there will be no violations of WLNC in the case of impossible antecedents. For any impossible antecedent A and consequent C, neither A $\Box\!\!\rightarrow$ C nor A $\Box\!\!\rightarrow \neg$C will be true. Similarly, if 'might' counterpossibles are all *true*, then there will be no violations of MLEM in the case of impossible antecedents. For any impossible antecedent A and any consequent C, both A $\Diamond\!\!\rightarrow$ C and A $\Diamond\!\!\rightarrow \neg$C will be true. Furthermore, there will not be violations of WIMP when the antecedent is impossible. Since every 'would' counterpossible is false, we need not worry about cases where some 'would' counterfactual is true, but its corresponding 'might' counterfactual is false. And, of course, we can make these changes without threatening the duality of the 'might' and 'would' counterfactual operators: regardless of whether or not the antecedent is possible, the operators will be such that A $\Box\!\!\rightarrow$ C =df $\neg$(A $\Diamond\!\!\rightarrow \neg$C).[16]

---

[15] Lewis briefly considers such operators in his [1973], though he does not consider them as a pair. He rejects them as being intuitively inferior to his preferred 'might' and 'would' counterfactual operators, but as we have seen, it is unclear what to make of his judgments about the plausibility of these operators.

[16] It should be noted that making all would-couterpossibles false would commit us to massive violations of what we might call the *would law of excluded middle* (WLEM). But, *pace* Stalnaker, the WLEM is not plausible. It does not seem true to say either 'If I were to flip this coin, it would land heads' or 'If I were to flip this coin, it would not land heads.'[16] And making all 'might' counterpossibles false would commit us to massive violations of the *might law of non-contradiction* (MLNC). But MLNC clearly is not plausible. It seems true to say both, 'If I were to flip this coin, it might land heads' and 'If I were to flip this coin, it might not land heads.'

Note that, unlike the other vacuous approaches on offer, the 'true mights and false woulds' approach is not capable of endorsing Entailment. The defender of such an approach would need to endorse the weaker notion of Entailment that the non-vacuous theorist is committed to.

Making all 'would' counterpossibles false and all 'might' counterpossibles true allows us to treat the most of the logical principles discussed above as valid. If we are to accept a vacuous approach for counterpossibles, this is the approach we should accept.

1.5 Counterpossibles and Negation

There are, however, further worries that face the principle discussed above. For example, it seems true to say, 'If France were a monarchy and France were not a monarchy, then France would be a monarchy' and also true to say, 'If France were a monarchy and France were not a monarchy, then France would not be a monarchy.' It seems false to say, 'If France were a monarchy and France were not a monarchy, then there would be fewer kangaroos.' For ease of exposition, I will replace a proposition such as *that France is a monarchy* below with P:

(14) $P \,\&\, \neg P \mathrel{\Box\!\!\rightarrow} P$

(15) $P \,\&\, \neg P \mathrel{\Box\!\!\rightarrow} \neg P$

At least arguably, (14) and (15) are both true: if P and not-P were both true, then P would be true, and not-P would be true. If there were a truth value glut with respect to some proposition P, we should expect that both P and not-P to obtain. Taken together, however, these counterpossibles violate WLNC, since (14) and (15) are a pair of counterpossibles with the same antecedent, but contradictory consequents.

The fault is not merely with WLNC, since it is not difficult to make trouble for the other putative principles as well. MLEM, for example, fails when we consider pairs of counterfactuals whose antecedents express that there is some truth value gap:

(16) $\neg(P \vee \neg P) \Diamond\!\!\rightarrow P$

(17) $\neg(P \vee \neg P) \Diamond\!\!\rightarrow \neg P$

Intuitively, both of these sentences are false: if neither P nor not-P were the case, then it is false that P might be true, and false that not-P might be true. If this is correct, however, then we cannot endorse MLEM.

The truth of WIMP is also in question. According to WIMP, (14) and (15) imply the following sentences, respectively:

(18) $P \,\&\, \neg P \Diamond\!\!\rightarrow P$

(19) $P \,\&\, \neg P \Diamond\!\!\rightarrow \neg P$

At least according to Lewis's definition of the 'might' counterfactual, however, (18) and (19) are equivalent to:

(18') $\neg\,(P \,\&\, \neg P \,\Box\!\!\rightarrow \neg P)$

(19') $\neg\,(P \,\&\, \neg P \,\Box\!\!\rightarrow P)$

Yet (15) and (18') cannot be true together, and (14) and (19') cannot be true together.

One potential response to this worry would be to rephrase our various logical principles in terms of truth and non-truth, and hold that such reformulated principles could accommodate the kinds of cases discussed above.[17] We can understand what is happening in these cases by introducing a new kind of negation, built from classical negation, represented by the symbol ~. I will continue to use the symbol ¬ for classical negation, and read ¬P as

[17] This kind of proposal is also discussed in a somewhat different context by Goodman [2004].

'P is false.' By contrast, ~P can be read as 'P is not true'; the 'not' here is understood as the 'not' of classical negation.[18] This ~-negation, then, is a generalization of classical negation, since whenever propositions are "only" true or false, ~-negation behaves just like ¬-negation. In the non-classical cases, however, ~-negation is a kind of glut- and gap-closing negation, since even in counterfactual cases where, for some proposition P, P and ¬P are both true (or both false), P and ~P will remain mutually exclusive and exhaustive. Consider, then, a replacement principle such as:

(WLNC*) Not: $A \mathrel{\Box\!\!\rightarrow} C$ and $A \mathrel{\Box\!\!\rightarrow} {\sim}C$

Note that (14) and (15) make no trouble for WLNC*. This is because if P and ¬P were true, then obviously P would be true, but we would not be inclined to think that ~P were true. After all, ~P is the claim that P is not true, and since P *is* true, ~P would not obtain in the scenario described in these counterpossibles.

Use of this kind of negation can also be of use in amending the other principles discussed above:

(MLEM*) $A \mathrel{\Diamond\!\!\rightarrow} C$ or $A \mathrel{\Diamond\!\!\rightarrow} {\sim}C$ is true

If MLEM* is true of 'might' counterfactuals, then (16) and (17) are not problematic. The falsity of these sentences is compatible with the truth of MLEM*; if $\neg(P \vee \neg P)$ were true, then it would not be the case that P might be true, but it does not follow that ~P might not be true. Indeed, it seems that ~P is true if neither P nor ¬P were true, since ~P is equivalent to the claim that P fails to be true. If there were a truth value gap with respect to P, then it would follow that P is not true.

[18] This kind of negation is similar to Meyer's *Boolean negation* and Sainsbury's *option negation*. See Meyer [1974] and Sainsbury [2007]. Unlike these kinds of negation, however, ~-negation is not related to choosing or options, and it is constructed from classical negation, rather than being a kind of *sui generis* operator.

What can we say about the failure of WIMP? This problem can also be solved by making use of ~-negation. We had previously defined the 'might' counterfactual from the 'would' counterfactual and ¬-negation. Instead, we could define the 'might' counterfactual in the following way:

(Might*) $A \Diamond\!\!\rightarrow C$ is equivalent to $\neg(A \Box\!\!\rightarrow \sim C)$.[19]

If this is how 'might' counterfactuals are to be understood, sentences such as (18) and (19) are no longer problematic, since they are not equivalent to (18') and (19'). Rather, they are equivalent to the following 'would' counterfactuals:

(18") $\neg (P \,\&\, \neg P \Box\!\!\rightarrow \sim P)$

(19") $\neg (P \,\&\, \neg P \Box\!\!\rightarrow \sim\neg P)$

These sentences state that it is false that, if P and not-P were true, then P would fail to be true, and it would be false that, if P and not-P were true, then it would be untrue that P is false. These sentences are compatible with (14) and (15), so long as we are careful about not conflating the two notions of negation.

1.5.1 Many Negations

Such a proposal faces problems, however, since it is just as easy to formulate counterpossibles that make claims about what would follow from propositions that are true and not true, rather than true and false. WLNC*, then, will face problems when we consider what would be true if some proposition P were true and not true. Consider, for example:

(22) $P \,\&\, \sim P \Box\!\!\rightarrow P$

(23) $P \,\&\, \sim P \Box\!\!\rightarrow \sim P$

[19] See Vander Laan [2007] for further discussion of this treatment of the relation of 'might' and 'would' counterfactuals.

Just as the pair of (14) and (15) are a counterexample to WLNC, the pair of (22) and (23) are a counterexample to WLNC*. After all, if P were true and not true, then P would be true, and P would be not true. The challenge, then, becomes how to handle counterpossibles whose antecedents express the failure of ~-negation. The same challenge can also be raised against MLEM* and WIMP that make use of Might*.

Perhaps we could respond to this worry by formulating a new principle that generalizes WLNC*. We could consider a third kind of negation, built from ~-negation: –P is true whenever P is not true, where the 'not' is read as the 'not' of ~-negation. That is, –P is true whenever ~P is "only" true, or where neither P nor ~P is true; it is false in cases where P and ~P are both true, or where P is "only" true. This negation closes truth value gluts (and gaps) such as those opened by sentences (16) and (17). This allows us to state a different principle about the behavior of counterfactuals:

(WLNC**) Not: $A \mathrel{\Box\!\!\rightarrow} C$ and $A \mathrel{\Box\!\!\rightarrow} -C$

Sentences (22) and (23) make no trouble for WLNC**. Even if P and ~P were true, we would not be inclined to think that –P were true. Similar changes could be made to shore up MLEM* and Might*.[20] It is easy, however, to see here how we could create trouble for WLNC** by formulating counterpossibles whose antecedents assert that both P and –P are both true. For any kind of negation we could imagine, there will be counterpossibles about some contradiction stated in terms of that negation.

We could create a hierarchy of laws that deal with the logical behavior of counterpossibles, each new kind of negation defined from a previous negation operator.

[20] Of course, Might* is not a logical principle, but rather a proposed definition for the 'might' counterfactual in terms of the 'would' counterfactual. If there is no stable definition for the 'might' counterfactual – if any proposed definition of the 'might' counterfactual cannot make sense of *all* uses of the counterfactual – then we might worry about how we are even to think of the 'might' counterfactual.

Nonetheless, we should not mistake one of these laws for a statement of the behavior of all counterfactuals. Each step of the hierarchy only explains the behavior of counterfactuals at the level "below" it. As soon as some principle is formulated, there is a recipe to create a pair of conditionals that serves as a counterexample. Thus, rather than finding a principle for the behavior of counterfactuals with respect to contradictions, we have found a family of such principles. We can now start to see the beginnings of a response to these logical problems that is at least open to the defender of a non-vacuous account of counterpossibles: even if there are exceptions to principles that at first seemed plausible, there may be nearby replacements for these principles that can be made sense of on non-classical grounds.

One lesson of the above considerations is that the failure of the vacuous theory of counterpossibles to make sense of laws such as WLNC, MLEM, and WIMP does not constitute a straightforward argument against the vacuous theory. After all, Lewis or Stalnaker might not be able to endorse some or all of these principles, but neither can the non-vacuous theories! Indeed, even reformulating these principles only puts off the problem, since there will be problematic pairs of counterlogicals that make use of the same kind of negation that is used to reformulate the principle. Nonetheless, there is some reason to think that the non-vacuous theorist is at least somewhat better off than the vacuous theorist of counterpossibles, since she is able to at least endorse this family of logical principles. That is, Lewis's semantics for counterpossibles hold that WLNC is violated for every relevant pair of counterpossibles, while successful non-vacuous semantics for counterpossibles can endorse WLNC in most cases, and offer a revised principle (such as WLNC*) in the cases where it fails.

1.6 Conclusion

Vacuous approaches to counterpossibles face various problems. They do not make sense of our linguistic behavior or philosophical practice, and their treatment of 'might' counterpossibles is in tension with the very arguments that are used to support the view. Indeed, no available approach is capable of endorsing all of the plausible principles of the behavior of counterfactuals discussed above. Of the vacuous approaches, we can do best by making all 'might' counterpossibles true and all 'would' counterpossibles false; nonetheless, such an approach is unmotivated in light of arguments from philosophical and linguistic practice. A non-vacuous approach to counterpossibles cannot endorse the above putative principles, either, but it can endorse a family of logical principles that capture the intuitions that make the original versions seem plausible. Doing so requires that we endorse a kind of 'gap-or-glut filling' negation that makes sense of what is ruled out in certain impossible situations.

I am not suggesting that classical logic is not valid at this world with respect to negation, or that we cannot consider impossibilities within some classical framework. Indeed, the theory of counterpossibles that I ultimately endorse *is* classical, in the sense that we can use classical logic to determine what is true at a world and to determine the logical laws of impossible worlds. However, what makes a world impossible is that its propositions are not closed under classical entailment; these worlds will contain contradictions. Reasoning about what would happen if a contradiction were true involves ignoring certain classical rules of logic if we are to avoid absurdity. To expect counterpossibles to respect principles like Entailment or WLNC is to misunderstand what the antecedents of these conditionals express.

CHAPTER 2

COUNTERPOSSIBLES WITHOUT IMPOSSIBLE WORLDS

2.1 Introduction

The previous chapter contained a collection of arguments that suggest that counterpossibles cannot be vacuous. If we are to make sense of linguistic intuitions as well as philosophical and mathematical practice, we should admit that counterpossibles are meaningful, and can be either true or false. Furthermore, the view that all 'might' counterpossibles are false stands in tension with some of the motivation for the view that counterpossibles are vacuous, and the vacuous view of counterpossibles is in tension with several plausible logical principles for the behavior of 'might' and 'would' counterfactuals. A non-vacuous view of counterpossibles is better able to make sense of our linguistic intuitions and philosophical practice, and is capable of accounting for the failure of these logical principles and offering suitable alternatives.

The analysis of counterpossibles that will ultimately be endorsed is one that expands the Lewis-Stalnaker semantics of counterfactuals in such a way as to allow *impossible* worlds into the similarity metric; such an approach is also discussed by Nolan [1997], Yagisawa [1987], Vander Laan [2004], Priest [2006], Mares [2004], and others. These impossible worlds are worlds that contain true contradictions, but need not be worlds where, *pace* classical logic, every proposition is true. Since there are now worlds that contain contradictions, evaluating counterpossibles poses no special difficulty on this approach. A counterfactual is true, then, iff the consequent is true in all of the nearest *worlds* – possible or otherwise – in which the antecedent is true.

Yet before such a view can be discussed in more detail, it should be noted that there is some space between the view that counterpossibles are not vacuous and the view that impossible worlds are required to give truth conditions to counterpossibles. This chapter explores a family of options that attempt to give plausible truth conditions to counterpossibles without any kind of commitment to the existence of impossible worlds.

Why might such a view be attractive? First, countenancing the existence of *impossible* worlds, alongside the possible worlds, might be ontologically profligate. Believing in possible worlds is, at least in some circles, already a large pill to swallow; adding *impossible* worlds to the mix might be too much to bear. Not only are we adding many *more* mysterious entities to our ontology, but these new entities are chaotic and unruly in a way that the possible worlds are not. Second, it is not entirely clear in what ways these worlds would be similar to our world, or similar to each other. How can we adjudicate the similarity of worlds with contradictions? Which kinds of contradictions are more acceptable than others? Can some world with a contradiction be more similar to the actual world than some world without a contradiction?

Of course, these considerations are not decisive, and they will be addressed in later chapters. Nonetheless, an alternate account that would provide non-trivial truth conditions for counterpossibles, while avoiding the complications associated with the countenancing impossible worlds would certainly be worth pursuing.

2.2 The consistent revisions approach

The first proposal relies on an analogy between the treatment of inconsistent belief and counterpossibes. Insofar as we understand (and can reason from) the inconsistent

antecedents of counterpossibles, we do so by construing them in a somewhat special fashion.

One method of linking the analysis of counterfactuals to a treatment of belief is to couch the antecedent and consequent of the counterfactual in belief reports. Rather than attempting to determine what would follow from the truth of *that I am a horse*, we could determine what would follow from *that I believe that I am a horse*. To get the proper truth conditions for the counterpossible, we would couch the consequent under a belief operator as well: if I were to believe that I am a horse, then I would believe that I have hooves. But there are several serious problems with this approach. First, it is unclear *whose* beliefs are relevant for determining the truth of the counterpossible: the utterer? the evaluator? an ideal epistemic agent of some sort?

Even if we could settle this question, we would be holding the truth of these counterpossibles hostage to broadly psychological considerations of what kinds of things agents would believe given certain other, contradictory beliefs. Who knows what I (or even some ideal agent) would believe if I were to actually believe that I am a horse? Maybe I'd believe that I were a special hoof-less horse, or maybe my set of beliefs would be too irrational and bizarre to make any real predictions about. And the problem is far worse if we consider metaphysical theses: after all, determining what follows from necessarily false philosophical theories is an important and substantive project, and one that we can be wrong about. I could easily have false beliefs about what would follow from mereological nihilism or Platonism that should not affect the truth or falsity of counterpossibles.

Rather than construing the antecedent and consequent of counterpossibles as operating under belief operators, we could rather construe them as being comprehensible in the fashion suggested by the compartmentalization approach to contradictory belief. Lewis

[1982], Varzi [1997], Stalnaker [1987], and others suggest a strategy for giving content to inconsistent beliefs within a possible worlds framework. Inconsistent beliefs pose a problem for possible worlds theories of belief content, since an agent's belief state is often taken to be the set of possible worlds that are not ruled out by how the agent takes the world to be.[21] If the agent has contradictory beliefs, however, *all* possible worlds are ruled out, since no possible world can represent the contradiction. On Lewis and Varzi's approach, contradictions are not accepted *tout court*, but are rather compartmentalized into consistent sets of beliefs. An agent's beliefs exist in these segregated compartments, such that she believes different propositions in these separate compartments. No compartment contains a belief of the form P-and-not-P, though one compartment may contain P, while another compartment contains not-P. A proposition is believed by an agent when it is contained in at least one compartment,[22] though perhaps we might prefer to relativize what is believed to a particular compartment or a particular situation; this matter need not be settled for present purposes. Even when some agent has contradictory beliefs, then, we can give content to his or her beliefs without resorting to impossible worlds; collections of possible worlds alone can suffice.

The inconsistent antecedents of counterpossibles can be evaluated within a classical framework by creating consistent revisions of the antecedent analogous to the consistent revisions of belief states. We would thereby divide the counterpossible into two counterfactuals, one for each revision of the antecedent.[23] A counterpossible under consideration would then be true when one of the resulting counterfactuals is true, just as a

---

[21] See Hintikka [1962]

[22] It should be noted that, on this approach, beliefs aren't closed under conjunction. An agent may believe P (since P exists in one compartment) and not-P (since not-P exists in another compartment), but not P-and-not-P, since the contradiction isn't present in any compartment.

[23] If the consequent is also impossible (e.g., 'If Graham Priest is right about logic, then there would be true contradictions'), we would divide the consequent as well.

proposition is believed by an agent if it exists in one of the belief compartments. Call such an account of counterpossibles the *consistent revision account*.

The consistent revision account of counterpossibles is plausible for the same reasons that the consistent revisions account of belief is plausible. Insofar as we are capable of grasping contradictions, we do so by grasping some part of the contradiction. We believe in contradictions by believing in something of the form P and something of the form ¬P, and are reduced to cognitive dissonance when we consider P & ¬P at once. Similarly, we are able to evaluate a counterpossible by determining whether some logical *part* of its antecedent would result in the truth of the consequent, but are reduced to cognitive dissonance when we attempt to evaluate what would result from the antecedent of the counterpossible taken as a whole.

Before the consistent revisions account can get off the ground, an important question must be addressed: how do we generate the revisions of the antecedent? The obvious strategy is to construe the antecedent of the counterfactual as some conjunction of the form *P-and-not-P*, and then make one revision of the counterpossible that takes *P* to be the antecedent, and another revision that takes *not-P* to be the antecedent. But this will not do. First, many counterpossibles do not seem to have antecedents that are conjunctive in form; consider most of the examples that have motivated this project, such as 'If Platonism were true, then we would have no way of knowing about numbers.' If there is something of the form *P-and-not-P* "in" the antecedents of counterpossibles such as these, it would be hard to find. This is not to say that it would be impossible to convert those counterpossibles into conjunctions, but it does not seem plausible that there would be some principled way of doing so. Second, many counterpossibles that do have explicitly conjunctive antecedents do not seem to be amenable to this treatment; consider:

(1) If I were to discover that the number of stars were even and odd, then I'd be surprised.

The counterpossible should not divide into 'If I were to discover that the number of stars were even, I'd be surprised' and 'If I were to discover that the number of stars were odd, I'd be surprised,' since neither of these counterfactuals are true. I would not be surprised if the number of stars were even, and I would not be surprised if the number of stars were odd; I'm surprised only when the number is even *and* odd. The contradictory nature of the antecedent is required to get the intuitive reading of the counterpossible. How can the consistent revision account deliver the intuitively correct result?

2.3 The Best Revision Account

We might still think that the right approach to counterpossibles is one in which a counterpossible is true if some related counterfactual is true. Perhaps the relevant counterfactual is not one in which the antecedent is some conjunct of the counterpossible's antecedent, but rather some other related proposition. Such a related proposition would be some related proposition from which we can recover the meaning of the contradictory antecedent. We can find a "best revision" for some impossible antecedent such that it is consistent (and thus amenable to the standard treatment of counterfactuals) while retaining as much of the content of the original antecedent as possible. If this best revision counterfactual is true, then the original counterpossible is true; if it's false, then the counterpossible is false.

Take R(x) to be a function from propositions to propositions that results in the best revision for the antecedent in question.[24] If A is the antecedent of a counterpossible to be evaluated, how do we determine the resulting proposition R(A)? Generally speaking, we need to consider the context of utterance (or the context of evaluation) as well as the antecedent itself to determine which proposition best captures the content of the antecedent. Indeed, the same considerations used to evaluate the truth or falsity of ordinary counterfactuals are in play here. When evaluating counterfactuals, we determine which antecedent-world is most similar to the actual world in the relevant context, and then determine whether the consequent holds in this world. Similarly, we use the context of the counterpossible to determine what content can be gleaned from its antecedent. Replacing the original antecedent with this consistent part of it will result in a counterfactual that can be evaluated in the ordinary way.

However the consistent revisions of counterpossibles are derived, it should be noted that this account gives a unified treatment to counterfactuals and counterpossibles. Intuitively, when evaluating subjunctive conditionals, we grasp the antecedent by taking the consistent revision of the antecedent that retains most of its content. If the antecedent is already possible, then there need be no revision of the antecedent. The best revision of the

[24] Note that, for this approach, we need some way of individuating propositions that is more fine-grained than functions from possible worlds to truth-values. If propositions are merely these kinds of functions, then all inconsistent propositions will be identical to each other; they will all be the function that returns *false* for every possible world. This will mean that, however the "best revision" function works, it will assign the same consistent revision to any inconsistent antecedent, since every inconsistent antecedent is the same proposition. More fine-grained approaches to propositions are required. Accounts of structured propositions such as that of [King 2007], or the interpreted logical forms discussed in [Ludlow 2000] could suffice. The kind of structured propositions discussed in this work in chapter 4 would suffice for this purpose as well, if our task is to individuate propositions such as *that 2 + 2 = 4* from *that all bachelors are male*. Note, however, that the kinds of structured propositions discussed in chapter 4 are not as fine-grained as some other accounts, and thus cannot differentiate a proposition such as *that Superman flies* from *that Clark Kent flies*, since Superman and Clark Kent will be picked out by the same set of individuals across worlds. Intuitively, these propositions ought to be differentiated, since one could believe one without believing the other, but the kinds of impossible worlds described here are not the proper tool for that job.

antecedent in such a case is the antecedent itself; if A is consistent, then R(A) = A. In this fashion, we evaluate all subjunctive conditionals using the same method; the "best revisions" approach need not be a separate analysis invoked only when the antecedent is impossible. Furthermore, this method guarantees that our original judgments about counterfactuals are unchanged; as long as the antecedent of the conditional is possible, the best revision of the antecedent is the antecedent itself, so there would be no change to our analysis of counterfactuals.

Of course, this is still just the outline of a theory of counterpossibles; more must be said about how this function operates before the view can be evaluated.

2.3.1 The Best Revision and Similarity

One means of determining the best revision of the antecedent is to posit similarity relations that hold between propositions; these similarity relations are analogous to the similarity relations that hold between worlds. On such an approach, we determine which consistent proposition is, in the contextually appropriate sense, most similar to the counterpossible's antecedent. Relying on the similarity of propositions to find the best revision should not be seen as *ad hoc;* indeed, we already have a notion of comparative similarity with respect to propositions that we employ in discussions about (imperfect) translations or synonymy. A sentence can be *very nearly* the same as some other sentence, in virtue of meaning *very nearly* the same thing. We can cash out such talk in terms of the comparative similarity of the propositions that these sentences express.

Though we might have some intuitions about the comparative similarity of propositions, such intuitions by themselves are not adequate to develop a general account. I have no such general account of similarity to offer, but we should not expect any such

account to match all of our intuitions about which propositions seem more similar to others, any more than intuitive views of the similarity of worlds will be of use in determining which possible worlds are closest to the actual world.

The proposal, then, is to take the similarity of propositions as unanalyzed, and use that notion to determine how the function R(x) works. That is, given some context, R(A) will be the possibly-true proposition that is most similar to the antecedent of the counterfactual. We then evaluate the counterfactual R(A) $\Box\!\!\rightarrow$C for truth; A $\Box\!\!\rightarrow$ C is true iff R(A) $\Box\!\!\rightarrow$ C is true. If there is a tie for the proposition most similar to the antecedent, then we could either say that the counterpossible is true when *all* of these nearby propositions are such that the resulting counterfactuals are true, or we could say that the counterpossible is true when at least *some* of the nearby propositions result in true counterfactuals.

This view, however, is still disappointingly vague unless we can determine what makes some proposition more or less similar to another. Perhaps one could hope to attach a theory of the closeness of propositions to an extant theory of the closeness of worlds. One proposition is similar to another if, by some measure, the possible worlds where the first proposition is true are nearby, in logical space, to the possible worlds where the second proposition is true. Of course, it would be very difficult to determine the degree to which one collection of possible worlds is nearby to some other collection of possible worlds, but we need not even attempt to address this. For our present purposes, we need to locate some possibly-true proposition that is nearby to some necessarily-false proposition. Since there are no possible worlds where the necessarily-false proposition is true, such an approach could not even get off the ground. Some alternate approach is required.

2.3.2 Qualitative Similarity and Counterparts

One approach for determining when some proposition is similar to another is the qualitative similarity approach: a proposition is similar to another proposition to the extent that the objects referred to by one proposition are qualitatively similar in relevant respects to the objects referred to by the second proposition.[25] Call the objects referred to in the revised proposition the *counterparts* of the objects referred to in the antecedent of the counterpossible. The counterpart of an object in this sense is one that is qualitatively similar to it in relevant respects; these counterparts must share as many of these qualitative features as possible, yet be such that the proposition that results from replacing the subject of the antecedent with its counterpart is possibly true. In this sense, my counterpart can "play my role" in some other world, even if in some worlds, my counterpart has properties that I could not have. According to the current proposal, then, we determine which possible proposition is most similar to the impossible antecedent by determining which possible proposition ascribes the properties and relations in question to the object or objects most qualitatively similar (in relevant respects) to those expressed by the counterpossible's antecedent. Thus, when we evaluate counterpossibles, we do not consider what would happen if some object possessed a property that it could not possess, but rather consider what would happen if some similar object bore the properties ascribed to it by the counterpossible's antecedent.

Of course, counterpart theory is nothing new with respect to understanding the modal properties of objects.[26] Yet its application here is quite independent of any theory of the representation of *de re* modal properties. If one is a counterpart theorist already, then one could accept this proposal as a kind of natural extension of the theory: this theory is an

[25] Assume that a proposition is a structured entity comprised of objects, properties, and relations.
[26] See, for example, Lewis [1968] and Lewis [1973].

attempt to exploit the kinds of relations of qualitative similarity that are used to explain how objects can be represented as existing at other world to explain how objects can be represented as having properties that they could not have.[27] Of course, one need not be a counterpart theorist with respect to transworld identity to accept such an analysis of counterpossibles: one could hold any view about how objects exist across worlds, but still think that counterpossibles present a special case that require thinking about the objects in question in a somewhat different way.

According to the counterpart theory of counterpossibles, when we grasp the antecedent of a counterfactual such as

(2) If I were a horse, then I would have hooves.

we do not somehow imagine something that is essentially not a horse (*viz.,* me) being a horse; rather, we imagine something that is a counterpart of me as a horse. On such an approach, then, we consider my qualitative properties and relations to determine the appropriate counterparts of myself, and determine whether the counterfactual that results in replacing me with such a counterpart is true. A counterpart of me that could be a horse would be an object that has as many of the qualitative features that I have as possible, yet could be a horse. Of course, specifying just what these features are is difficult, and much of the discussion would hinge on what kinds of properties one thinks is essential to being a horse or being a human. However this is to be spelled out, it should not be that my counterpart has the same modal properties that *I* have: that is, we can assume for the sake of the example that I am necessarily human, and my counterpart is not.

[27] Assuming, of course, that there is no admissible counterpart relation according to which the object in question has counterparts with the property ascribed by the antecedent. If counterpart relations are tolerant enough to rule out all essential properties, then these kinds of conditionals will not be counterpossibles.

On this approach, to determine the truth of (2), we determine the truth of (2'):

(2') If my counterpart were a horse, then it would have hooves.

In the nearest world where my counterpart, or something that plays my role, is a horse, that thing has hooves; after all, if we assume that the counterpart in question is a horse at some world, it would have the kinds of features we ordinarily associate with horses at that world. Therefore, the revised counterfactual is true, and thus the counterpossible is true. Of course, if such an approach were to be developed and defended, more would have to be said about how we locate these counterparts and what their modal properties would be like.

Yet, however these questions are answered, this approach does not seem applicable to cases of counterpossibles with antecedents that are logically or mathematically impossible. It is unclear how to understand counterpossibles such as

(1) If I were to discover that the number of stars were even and odd, I'd be surprised.

What are we supposed to find counterparts *of?* Presumably, not myself: there is no counterpart of myself whatsoever that could discover that the number of stars were even and odd, since 'discover' is factive: nothing could discover the impossible. It seems we would have to find some counterpart for the number of stars that could be both even and odd. What would this counterpart be? Clearly not some other number: no number could possibly be both even and odd. Yet if the counterpart is not a number, then how could it be even or odd at all? More generally, it seems difficult to imagine how this approach is supposed to find the relevant kinds of qualitatively similar entities that could have collections of properties that are logically or mathematically impossible.

It seems that the qualitative similarity approach is not capable of determining the relevantly similar, revised antecedent for a wide range of counterpossibles. Of course, one

could endorse some other "best revision" approach which attempts to determine the relevantly similar proposition by some other means, but it is unclear what other strategy could be plausible. Nonetheless, a more general argument against "best revision" approaches could forestall an attempt to come up with a more plausible method of determining how to revise the antecedent of counterpossibles.

2.4 Some objections: Surrogate Propositions and Inconsistency

At this point, one might object: perhaps this approach has merely avoided the difficult problem by changing the subject, since we are no longer dealing with the truth conditions of the counterpossibles, but rather with the truth conditions of counterfactuals that are somehow similar to the counterpossibles with which we began. Instead of evaluating conditionals with antecedents such as *I am a horse* or *I have squared the circle* we evaluate conditionals with antecedents such as *Something like me is a horse* or, *I come to believe that I have squared the circle*, and so on. The conditionals that we analyze for truth are, by design, distinct from the original counterpossibles. By analyzing these surrogate conditionals, one might complain, I have merely ducked the hard question of how we can deal with impossible propositions in a Lewisian framework. But this objection misunderstands the enterprise. This approach is merely an attempt to give non-vacuous truth conditions to counterpossibles, and we have done so by identifying them with the truth conditions of counterfactuals. The surrogates provided are not arbitrary conditionals unrelated to the original counterpossibles; rather, they are selected to capture the cognitive content of the impossibilities expressed by the impossible antecedents. A counterpossible has the truth conditions it does in virtue of the nearby counterfactual. This does not mean that the counterpossible is identical with the content of the counterfactual, or that we can never

properly express counterpossible propositions. The claim is merely that counterpossibles are true whenever their associated counterfactuals are true.

But there is a more serious worry in the neighborhood that does not have as easy an answer. Perhaps some counterpossibles have the truth conditions that they do in virtue of the contradiction expressed in the antecedent, and any attempt to avoid this contradiction by way of replacing the antecedent with some possible proposition is bound to give us the wrong results. That is, this entire family of approaches attempt to give the truth conditions of the counterpossible in terms of the some counterfactual with a possible antecedent. These approaches, then, are unable to give intuitively correct truth conditions to a counterpossible whose truth depends on the impossibility of its antecedent.

At least one kind of case where the impossibility of the antecedent is salient to the truth of the conditional *does* have an adequate response. Recall the 'anything goes' argument discussed in chapter 1: on at least some ways of understanding the contradictory antecedents of counterogicals, making the contradictory nature of the antecedent salient makes the result that all counterpossibles are trivially true far more plausible. When we start to consider what would happen if some contradiction were true, the argument goes, it becomes unclear just what else would be true; indeed, given some impossible antecedent, maybe anything at all is true. Yet this kind of special case is impossible according to the best revision approach, since the antecedent will always be logically possible, and explosion will never result. Furthermore, this trivial reading of the counterpossible is exactly what we require to make sense of proofs by *reductio*. In such a proof, we can show that some assumption leads to a contradiction, and thus the assumption must be false and its negation true. It is the contradiction itself that gives us this result, and not some consistent revision of it that allows us to draw this conclusion. Any form of the best revision approach to counterpossibles

replaces the contradictory antecedent with something consistent, but in the case of an "anything goes" context or a proof by *reductio*, it is precisely this inconsistency that is required to derive the desired consequent.

The response is that presenting a proof by *reductio* must set a special kind of context that is similar to the context set by the 'anything goes' argument. In this special case, the consistent revision function will return an impossible proposition after all, such that everything follows from it. Thus, *reductios* will be treated as a special context where R(A) is a contradiction after all, but the contradiction will behave in a classical fashion, and explosion will (rightly) result.

There is a more serious objection to this family of proposals, however. Consider a counterpossible such as the following:

> (3) If the number of stars were both even and odd, then some contradiction would be true.

First, note that the consequent expresses an impossibility; if the analysis is to work, we must also find some nearby contingent proposition for the consequent to express. The solution, then, is to apply the revision function sketched above to the consequent in order to give (2) the desired truth value. A counterpossible of the form $A \mathrel{\Box\!\!\rightarrow} C$, then, would be true whenever $R(A) \mathrel{\Box\!\!\rightarrow} R(C)$ is true.

It is not clear that moving toward some revision of the consequent will solve the problem, however. In (3), it seems that any attempt to make the antecedent logically possible will be inadequate to account for its content in such a way as to guarantee the truth of the conditional. After all, intuitively, there would be a contradiction in the world of evaluation *because* the antecedent is contradictory. That is, the link between the antecedent and consequent of (3) that makes it seem (non-vacuously) true depends on the antecedent being

impossible, and an analysis that treats the antecedent (and consequent) as possible would miss this connection. This points toward a more general problem: however it is that we decide what the best revision is, whether it be an approach that considers the beliefs of the speaker, the counterparts of various objects, or some unanalyzed notion of a relevantly similar proposition, it will be possible to come up with a counterpossible which cannot be successfully analyzed by that approach. This is because the 'best revision' approach will always leave some gap between the meaning of the antecedent and the meaning of the revised antecedent, and this gap in meaning can be exploited in such a way as to produce a counterexample.

In the case of (3), if we apply the consistent revision strategy to both the antecedent and consequent, we will have no guarantee that the resulting counterfactual will have the correct truth value. It seems implausible that a 'best revision' for the antecedent of a conditional such as (3) would have much to do with our beliefs, or anything about the counterparts of mathematical or logical objects. After all, it is difficult to say what I would believe if I were to come to believe that the number of stars were even and odd, and as we have seen, it is difficult to make sense of what the relevant counterpart relations in this case. The best hope is to understand the best revision of the antecedent of this conditional in terms of some unanalyzed "nearby" proposition.

One complaint, of course, is that there simply is not enough information about propositional similarity to go on. What might this proposition "nearby" to the antecedent be? My complaint is not that similarity in this notion of similarity is unanalyzed; everyone is entitled to his or her primitives. Rather, the complaint is that it is difficult to imagine what the most similar, logically possible proposition to some given impossibility could even look like. It is difficult to make the case that a sentence such as (3) is analyzed correctly or not if

we have no intuitive grasp of what the corresponding counterfactual would be. Thus, the theory is completely untestable. Any attempt to elucidate the similarity relation for propositions will open it up to easily generated counterexamples.

2.5 Counterfactuals and Logic

There is one further family of non-trivial analyses of counterpossibles to be considered in this chapter. We might think that the counterfactual conditional is some kind of entailment operator after all, and that Goodman-Chisholm analysis, according to which we evaluate a counterfactual by determining whether the antecedent (as well as relevant background information) entails the consequent, is nearly correct after all. One could then argue that a counterpossible is true whenever its consequent can be formally derived from the antecedent in the proper way. Since avoiding triviality is important, however, the notion of entailment in such a theory cannot be classical, lest any consequent be true when the antecedent is inconsistent. The entailment relation must in question must, at least, reject the principle of explosion, according to which everything follows from a contradiction. According to such an analysis, $A \mathrel{\Box\!\!\rightarrow} C$ is true whenever $A \rightarrow_L C$, where $\rightarrow_L$ represents the entailment operator of some language L.

Which language could we fill in for L? We need not commit ourselves to a particular logic here, but it seems plausible that a relevance logic[28] would be a strong candidate. Such logics avoid explosion, and thus would not result in triviality; they also enforce a requirement that the premises of an argument be "relevant" to its conclusion; the information contained in the premises must be used in deriving the conclusion. There are various ways to enforce this requirement, either syntactically (by marking premises in such a way that one must make

[28] See, for example, Anderson and Belnap [1975] and Restall [2000].

use of them in a derivation) or semantically (by using a so-called *ternary* accessibility relation,[29] which would hold that worlds can only be accessible to one another in such a way that the information contained in the premises is present in the worlds where the conclusion is true). What is important for our purposes is that there are plausible logical systems that could be used in such a way as to give counterpossibles non-trivial truth conditions without the use of impossible worlds.[30]

In order to accept such a view, one need not endorse some non-classical logic *per se*. This view does not claim that, for example, the principle of explosion is invalid. One could endorse this view as a classical logician and a non-pluralist, claiming that classical entailment is the only game in town. Perhaps non-classical logics describe some *other* relation between propositions, perhaps they are merely syntactic games. Yet so long as they describe some kind of intelligible system, we can use that system to determine when counterpossibles are true.

A more serious problem is that, even if we could decide on a plausible logical system, it would never be adequate to describe the truth conditions of all counterpossibles. For any choice of L, there will be some weaker logic, L', such that we could use counterpossibles to discuss the behavior of L'. Yet using the entailment relation of L to describe what follows from L' would clearly be mistaken. If, for example, we used relevant implication to determine what would follow from intuitionistic logic, we would find that all the theorems of *relevance* logic would be true. Clearly, any choice of L for counterpossible entailment will not allow one to determine what would follow if other languages were true.

---

[29] Due to Routley and Meyer [1983].

[30] One might object that logical approaches do require the use of impossible worlds, since most of the semantics for these logics makes use of worlds that are inconsistent or incomplete. But the possible worlds of logical systems are set-theoretic entities, not worlds in the sense that (for example) Lewis discusses in [1986].

Perhaps there is a more modest kind of logical analysis of counterpossibles. Rather than make use of some specific non-classical logic, one might hold that a counterpossible is true if its consequent can be derived in classical logic, in some "reasonably direct" way. Of course, what it is for a proof to be "reasonably direct" is quite vague, but it surely must avoid use of principles such as explosion, lest the theory of counterpossibles be vacuous. Yet we cannot merely rule out this rule of inference on its own, since explosion can be derived from disjunction introduction and disjunctive syllogism, both of which seem to be unimpeachable, even for paradigmatically "reasonably direct" proofs; a reasonably direct derivation could easily include one or both of these principles and not lead to triviality. Furthermore, it seems, sometimes the principle of explosion is an important part of reasoning, as in certain proofs by *reductio* and in certain "anything goes" contexts. Thus, it seems quite hard to know just what a "reasonably direct" proof would and would not rule out. Clearly, there are no particular rules of inference to be denied, nor some number of steps of a formal derivation that one is entitled to use. The notion of a reasonably direct proof simply seems too vague to be of much use.

Of course, these arguments are not decisive; the defender of some logical account of counterpossibles could attempt to come up with some theory of what a "reasonably direct derivation" is in such a way that makes it at least as precise as counterpossibles, and at least in theory testable. Yet as it stands, there seems to be no way to avoid the principle of explosion without adopting a non-classical logic, and adopting any particular non-classical logic will leave the theory incapable of dealing with all counterpossibles.

2.6 Conclusion

An analysis of counterpossibles that aims to provide correct, non-trivial truth conditions without countenancing the existence of impossible worlds is worth exploring. Since such a theory does not rely on impossible worlds, it would not need not answer uncomfortable questions about ontology or similarity among logically impossible worlds. Rather, such an account offers an ontologically cheaper picture of how we reason about counterpossibles that gives us intuitively plausible truth conditions by relying on less problematic, nearby propositions in place of impossible antecedents.

These approaches, unfortunately, are ultimately inadequate to the task. If the analysis of counterpossibles proceeds in such a way that ignores the inconsistent nature of their antecedents, then it will be possible to find counterexamples to the theory that trade on this gap between the impossible antecedent and the possible revision of it. After all, many counterpossibles are true *because* the antecedent is impossible, and if that counterpossible is to be analyzed in terms of some counterfactual with a possible antecedent, there is simply no reason to assume that this counterfactual will have the appropriate truth value. If, on the other hand, the analysis relies on some particular choice of a non-classical entailment relation, one must be faced with the choice of *which* entailment relation to make use of, which would close off the analysis to a large family of counterlogicals.

CHAPTER 3

IMPOSSIBLE WORLDS AND SIMILARITY

3.1 Introduction

The aim of this chapter is to provide an account of what makes a world – possible or impossible – more or less similar to the actual world, so that truth conditions for counterpossibles can be given in such a way as to extend the Lewis/Stalnaker semantics of counterfactuals. According to the Lewis/Stalnaker account, a counterfactual is true iff its consequent is true in the "closest" or "most similar" possible worlds where the antecedent is true.[31] That is, a counterfactual is true when the consequent of that counterfactual is true in those possible worlds at which the "minimal changes" necessary are made to smoothly accommodate the truth of the antecedent. A sentence such as, 'If dinosaurs were to still exist, then there would be fewer people' is true, as there are fewer people in a world much like this one (in relevant respects), yet where dinosaurs still exist. It would seem plausible that such a "nearby" world is one where we change facts about how the dinosaurs actually went extinct, as well as various facts about history and the environment that would have been affected by the presence of dinosaurs; plausibly, we would change facts about evolutionary development or predation that would affect the human population in this world. Other facts that are not relevantly related to the presence of dinosaurs, such as the color of snow or the shape of our planetary orbit, are not changed at such a possible world.

[31] The Lewis/Stalnaker account of counterfactuals is developed in Lewis [1973] and Stalnaker [1968] and discussed in more detail in chapter 1. There are, of courses, differences between the way that Lewis and Stalnaker analyze counterfactuals, such as the analysis of 'might' counterfactuals and whether, given some counterfactual, there is a unique "closest" possible world or set of worlds. These differences will not affect how the account of counterfactuals is developed here; I assume, for the sake of exposition, that there may be "ties" with respect to which worlds where the antecedent of the counterfactual are true are closest, but that there will always be at least one closest antecedent world.

Much of the disagreement about the analysis of counterfactuals, then, lies in determining which facts vary and which facts are held fixed in the "most similar," "closest," or "nearby" possible worlds.[32] What kind of changes must be made in order to properly alter what is true at a world so as to locate the nearest possible world where the antecedent of the counterfactual is true?

As Lewis argues, it is possible to come up with a kind of "standard" similarity metric that would allow one to determine how a world is minimally altered in this fashion. With some minor refinements, Lewis's analysis is capable of explaining the truth conditions of counterfactuals and meeting some of the challenges that have been leveled against it. Furthermore, with these refinements in hand, it also becomes clear how to extend Lewis's treatment in such a way as to give truth conditions to counterpossibles.

3.2 Lewis and Counterfactuals

It is clear that offhand intuitions about how similar some possible world is to the actual world will not count for much. Lewis said this as early as his [1973:95], although much of the early discussion of his account nevertheless relied on such judgments about similarity. Of course, noting that our intuitions about the similarity of worlds are a poor guide to transworld similarity does not bring us much closer to developing an account of the phenomenon. The challenge of explicating this notion remains; indeed, it is exacerbated if we think that our intuitions about similarity will play little if any role in determining the comparative nearness of worlds.[33]

---

[32] I will use locutions such as "similarity," "nearness," and "closeness" synonymously.

[33] The kind of similarity that holds between worlds should not be confused with a notion of "objective similarity" that holds between objects. When we say that two objects are similar in such a way as to underwrite claims about counterpart theory, for example, we are not appealing to the same notion of similarity that underwrites trans-world identity. There are, however, some common

Consider the case of Fine's Bomb, discussed in his [1975]. Imagine that, at some time *t*, the President's finger is poised over the button that would launch nuclear weapons at Russia. Thankfully, she thinks better of it, and decides not to press the button. Intuitively, the following counterfactual is true:

(1) If the President had pressed the button at *t*, there would have been nuclear war.

But how can we get this result with something like the view of counterfactuals sketched above? If we think that the nearest possible world where the President presses the button is like this world in relevant respects, then we should think that it is a world that is, at least as much as possible, very much like the actual world. Yet, consider a few possible worlds. In w_1, the President presses the button, the mechanism works as designed, and there is nuclear war. In w_2, the President presses the button, but the firing mechanism fizzles in such a way that no missiles are launched, and there is no nuclear war. If we are to rely on intuitions about similarity, it seems clear that w_2 is closer to the actual world than w_1: after all, a world with nuclear war is a world that develops in vastly different ways after *t*, and would look very different from the actual world. A world without a nuclear war such as w_2 deviates from the actual world only in matters of comparatively little importance. The button may have the President's fingerprints, and she may have memories of pressing the button, but such differences seem inconsequential compared to the nuclear holocaust present in w_1.

As Lewis explains in his [1979], only certain kinds of matches of facts count toward the closeness of worlds; not all such matches are created equal. Lewis presents a list of desiderata in determining which kinds of matches are important:[34]

---

threads: for example, the requirement that there be a "perfect match" of matters of fact resembles the requirement that the appropriately natural properties be duplicated.

[34] For the purposes of this discussion, Lewis assumes that determinism is true. If the actual world is indeterministic, then the picture will be at least somewhat more complicated; one must understand the notion of 'miracle' that Lewis deploys in a somewhat different fashion, so as to count events that

1) It is of first importance to avoid big, widespread diverse violations of law.

2) It is of second importance to maximize the spatiotemporal region throughout which perfect match of particular fact prevails

3) It is of third importance to avoid even small, localized, simple violations of law

4) It is of little or no importance to secure approximate similarity of particular fact

A world that contains a widespread and diverse violation of nomological law (or "large miracles") will never be as closeby as a world that does not contain such a violation. Second, a world will be closeby to the extent that it has a large region of *perfect* match of matters of fact, unless such a match comes at the expense of a large miracle; it is not important that a world like w_2 has approximately the same kind of history as the actual world, but it should rather match the actual world perfectly. Third, a closeby world should minimize the occurrence of "small miracles," or more localized violations of nomological law.[35] In some contexts, it is at least somewhat important to achieve an "approximate" match of matters of fact, but not in other contexts; we can ignore this complication for present purposes.

In light of this, Lewis's response to Fine is now clear: we can imagine two ways in which the nuclear-holocaust-free world could seem nearby to the actual world, but on either way of understanding w_2, a world such as w_1 will end up as closer to the actual world according to Lewis's analysis. Both worlds will, of course, require a "small miracle" in order to make it true that the President presses the button; there must, somehow, be an exception to the laws of physics, perhaps localized in some region of the President's brain, which

---

are nomically possible but are coincidental to the point of seeming to be conspiratorial. See his discussion of 'quasi-miracles' in the postscript to [1979].

[35] When a possible world is said to contain some violation of law, this should be understood as a violation of *our* laws. If some miracle occurs that causes some possible world to deviate from our word, then our laws of physics were violated; presumably, the possible world that contains this miracle either has different laws, so as to account for this change, or *no* laws that apply to the phenomenon in discussion.

causes her to press the button that launches the nuclear weapons. But worlds like w_2 will either *not* perfectly match the future of the actual world, or they will also require a "large miracle" so as to achieve this match.

Imagine a world w_{2A} at which there is no nuclear war, but the myriad effects of the President's button-pressing still spread out into the future. Maybe that moment is recorded in her memoirs; maybe her decisions later in life are profoundly affected by this dark moment in her past. There will certainly be a great many mundane effects as well: the President's fingerprints will be on the button, the sound waves will emanate from the click of the button, the heat will dissipate from the wire, the air will be disturbed, and so. Various changes in the world will propagate from the button-pressing. The development of this world after t will *not* perfectly match what happens in the actual world: while the past of the actual world and w_{2A} will be perfectly alike, the future of the two worlds will *not* be perfectly alike, as the traces of the President's actions spread through the world after t. Since this match of fact in the future is not perfect, it will not count toward its similarity to the actual world. World w_1, then, is closer to the actual world than w_{2A}. Both possible worlds match the history of the actual world just prior to t perfectly, and neither matches the history of the actual world after t. However, w_1 requires only one small miracle (which causes the President to press the button), while w_{2A} requires two small miracles (one causes the President to press the button, and the second causes the firing mechanism to fail.)

We could also imagine a world w_{2B} at which the various traces of the President's button-pressing – the memories, fingerprints, sound waves, and so on – are all wiped clean from the world. In worlds such as these, some diverse and widespread violation of law would be required to eliminate all of the myriad traces of the President's button-pressing; the world would have to conspire to remove the memories, the sound waves, the heat from the

wire, the disturbances in the air, and so on. Note that w_{2B} *does* match the actual world perfectly after *t*, but it does so at the cost of a large miracle. Thus, neither of these worlds will be as nearby as w_1, where there *is* nuclear war, but no large miracle.

It is important to note that Lewis does not offer this approach as some kind of analysis of our intuitive notion of trans-world closeness. As Fine has shown, our intuitions about the similarity of worlds will land us in trouble if they are used to evaluate the truth of counterfactuals. A more technical conception of transworld closeness, divorced to at least some degree from our intuitions, can be judged adequate to the extent that it is capable of being deployed in such a way as to make sense of our intuitions about the truth values of counterfactuals. If Lewis's desiderata for the closeness of worlds does not seem intuitive, it is merely because a theory of the closeness of worlds must ultimately answer to our intuitions about counterfactuals such as (1).

3.3 Complications

Nonetheless, the account given by Lewis in his [1979] cannot be the last word on the matter. The first kind worry for Lewis's account is that, in at least some cases, there are "deviant" possible worlds that contain a perfect match in matters of fact without requiring some large miracle; in these cases, Lewis's desiderata would give us the wrong truth values.

Consider a process that results in a "deviant" match of matters of fact, such that the kind of analysis suggested in Lewis's [1979] will get the wrong result. One such case is presented by Elga in his [2001]. Imagine that, at the actual world, Greta cracks an egg on a pan at 8:00; by 8:05, a fried egg sits on the pan. Consider the following counterfactual:

(2) If Greta had not cracked an egg at 8:00, there would be no fried egg at 8:05.

This counterfactual is presumably true: the nearest worlds where Greta does not crack an egg in the pan are worlds where there is no subsequent fried egg. At such worlds, a small miracle occurs, perhaps isolated in Greta's brain, which results in her deciding not to crack an egg in the pan, and thus, no fried egg results; call a world such as this w_3. There are, to be sure, possible worlds where Greta does not crack the egg, but a fried egg miraculously appears in the pan at 8:05, but Lewis would claim that these worlds would either not match the actual world's history after 8:00 perfectly (if, for example, there are differences in Greta's memory or in the heat distribution in the kitchen), or would require a large miracle (so as to ensure that all traces of Greta's decision *not* to crack the egg are wiped from the world).

Consider, however, a world such as w_4, where a fried egg appears on the pan at 8:05; the history of w_4 and the actual world match perfectly after 8:05, complete with all of the apparent traces of Greta's decision to crack the egg. Yet the fried egg that appears at 8:05 is not the result of Greta having cracked the egg, or some large miracle that conspires to make the world seem like Greta had cracked the egg, but is rather the result of a more lawful, anti-entropic process: prior to 8:05, the fried egg "unrotted" on the pan, forming in some improbable-seeming way out of organic material around the pan. Prior to 8:00, the history of a world like w_4 will look quite different from the history of the actual world, yet at 8:05, the worlds perfectly converge *without* the kind of large miracle that Lewis would object to.

Worlds such as w_3 and w_4 both match the actual world in an analogous way, since w_3 perfectly matches the actual world from before 8:00 and w_4 perfectly matches the actual world after 8:05. It is clear why w_3 requires only a small miracle: we require only some small change in some region of Greta's brain in order to make Greta decide not to crack the egg, and the rest of the world evolves in some predictable way from there. But why think that w_4 also requires only a small miracle, when the world contains some incredibly improbable

event that results in a conspiracy of false memories and other seeming traces of Greta's decision?

Elga asks us to imagine the process that occurs in the actual world from 8:00 to 8:05, in which Greta cracks an egg, the egg heats in the pan, and a fried egg results. Consider now this process run backwards in time, from 8:05 to 8:00, in which a fried egg cools in a pan, and eventually flies back into its shell; such a process is possible according to the dynamic laws, but is fabulously improbable. Such a process is also fragile, in the sense that a small change to some of the particles in the system will result in a wildly different result, in which the egg will not return to its shell, but rather remain in the pan and continue to cool and rot (albeit in reverse). Such a small change would constitute a small miracle; thus, a small miracle change to the 8:05 to 8:00 process results in a very different world in which there is a fried egg, but no egg-cracking. If we take the actual world, then, and make a small, localized change to the particles in the pan at around 8:05, and then run the dynamic laws *backwards*, the result will be a possible world w_4 which differs from the actual world by a small miracle, and in which there is a fried egg at 8:05, but no egg-cracking at 8:00. Since w_3 and w_4 are tied according to Lewis's metric, the defender of such an approach would not be able to claim that a counterfactual such as (2) is true.[36] A modification of Lewis's account is required to meet Elga's challenge.

3.4 Causation and Explanation

There are, to be sure, other proposals that seek to modify Lewis's account of the similarity of worlds in such a way as to handle the kinds of objections that Elga and others

[36] Elga's example is a particularly persuasive case where reconvergence does not seem to require a 'large miracle.' Bennett discusses other such cases in [1984]. For other putative counterexamples to the similarity metric described in Lewis [1979], see Wasserman [2006], Kment [2006a], and Hawthorne [2005].

advance against Lewis. The strategy is to locate some feature of the kind of match of matters fact that is deviant; the kinds of match of matters of fact that Elga and others rely on to formulate their counterexamples could then be ruled out in our similarity metric of possible worlds.

According to Schaffer [2004], the only facts that count toward the comparative nearness of possible worlds are those facts that are causally independent of the antecedent. In the case of Elga's egg, we should not consider w_4 to be as nearby as w_3, because while w_4 perfectly matches the history of the actual world after the appearance of the fried egg, such a match is *not* causally independent of the antecedent: whether or not there is a fried egg (in the actual world) depends on the result of a causal chain that is anchored in the Greta's egg-cracking. Since, in the actual world, the fried egg is caused by Greta's cracking of the egg, any kind of match related to the existence of the egg (which includes much of the match that exists between the actual world and w_4 after 8:05) simply does not count toward trans-world similarity. Thus, w_3 will be closer to the actual world after all.

We might not be content, however, to accept Schaffer's account of counterfactuals. First, as Schaffer notes, it does not allow us to reduce causation to counterfactuals in some clearly non-circular fashion. To be sure, the counterfactual account of causation faces some trouble,[37] but it seems premature to rule out such an account by building causation into our notion of counterfactual analysis. This is no refutation of Schaffer's analysis, but there is at least some reason to look for some analysis of counterfactuals that does not rely on causation, so as not to close off hope of any such analysis.

Another reason not to accept Schaffer's analysis is that it does not seem general enough for our purposes. Ultimately, our approach should be capable of making sense of

[37] See, for example, Lewis [2000], Schaffer [2000], and Hall [2003].

conditionals whose antecedents express nomological, metaphysical, or logical impossibilities. But Schaffer's account is concerned only with avoiding a certain kind of deviant match of matters of fact that arises in cases such as Elga's; it is not of much help when evaluating counterfactuals whose antecedents express nomological impossibilities, since it gives no explicit guidance with respect to how to measure the comparative similarity of worlds with different laws. And, of course, it is not clear how such an approach can be extended to give an analysis of when counterpossibles are true.

Perhaps causation is the wrong notion to hang our analysis on. According to Kment [2006], the only facts that count toward the comparative nearness of possible worlds are those that have the same explanation in both worlds. Thus, in the case of Elga's egg, world w_4 will not be closer to the actual world than w_3, since the presence of the fried egg has different explanations in w_4 and the actual world: in the actual world, there is a fried egg because of Greta's actions, and in w_4, there is a fried egg because of some anti-entropic process.[38] Thus, the presence of the fried egg, and indeed the rest of the history that follows from the existence of that fried egg, simply does not contribute towards the closeness of a world like w_4 to the actual world; worlds such as w_3 will match the history of the actual world perfectly prior to 8:00, but worlds such as w_4 will not match the history of the actual world after 8:05, since the events after that time have a different explanation than they do in the actual world. The two worlds, then, are not tied for similarity to the actual world.

Kment's approach places a lot of weight on the notion of explanation, so more must be said about the notion of explanation that is capable of doing the work required. If the kind of explanation is epistemic, then Kment's analysis of counterfactuals is hostage to how

[38] One could argue that we cannot explain the existence of an egg by some probabilistic mechanical process, especially if the probability of that phenomenon is close to zero. Perhaps it would be better to say that the fried egg has *no* explanation in w_4. In either case, though, it is clear that it does not have the *same* explanation that it does in the actual world.

well we understand (or can understand) the world. Yet it seems that what would happen under counterfactual circumstances does not depend on our knowledge or our cognitive abilities. There is a fact of the matter about what would happen if I were to press a button or crack an egg, and what would happen does not depend on whether we know how the button is wired or what our present beliefs about of the laws of statistical mechanics are. The notion of explanation, then, should be ontic; explanations should be understood in such a way that they are somehow part of the structure of the world. Kment suggests as much, proposing that we should understand explanation as Salmon does in his [1984]. Yet Salmon's account of explanation relies both on causation, since for him explanation is an attempt at fitting the phenomenon to be explained into a *causal* structure, as well as counterfactual analysis, since this causal structure is understood in terms of the ability of objects to transfer "marks." The possession of some ability seems to be a counterfactual notion: something possesses the ability to transfer a mark if it *would* transfer a mark under certain (generally) counterfactual conditions.

Insofar as the proposed account of explanation makes use of causation, it will inherit the charges against Schaffer. More importantly, we cannot adopt Salmon's account of explanation in the analysis of counterfactuals without courting blatant circularity as long as explanation is understood in counterfactual terms. Finally, it would seem that such an account would leave us no closer to understanding the truth conditions of counterfactuals with nomically or metaphysically impossible antecedents. Perhaps a proposal along the lines offered by Kment and Schaffer could be made to work: one could attempt to reformulate such an account of the similarity of worlds if one is able to elucidate a notion of explanation (or causation) adequate for the project. The moral to be drawn for the purposes of the present analysis, however, is that, even if laws and explanation can play a role in the

similarity metric for counterfactuals, they cannot be used in the way that Schaffer and Kment suggest.

3.5 Special Science Laws

Dunn [2010] suggests a different kind of way of amending Lewis's similarity metric in response to Elga. According to Dunn, the similarity metric should be changed by adding a requirement that it is important to avoid violation of special science laws. This requirement is inserted into Lewis's desiderata after the third requirement, that it is important to avoid "small miracles." Dunn's revised similarity metric, in full, is as follows:

1) It is of first importance to avoid big, widespread diverse violations of fundamental, nomological law.
2) It is of second importance to maximize the spatiotemporal region throughout which perfect match of particular fact prevails
3) It is of third importance to avoid even small, localized, simple violations of fundamental, nomological law

3.5) It is of fourth importance to avoid violation of special science laws.

4) It is of little or no importance to secure approximate similarity of particular fact.

Note that, in the first and third requirements, we have made clear that Lewis seeks to avoid violations of *fundamental* law, so as to differentiate these requirements from the kind of match of special science laws that Dunn recommends.

This approach to the comparative similarity of possible worlds offers a straightforward response to Elga's challenge: a world such as w_3 will be closer to the actual world than a world such as w_4, since w_3 does not violate special science laws in the way that w_4 does. The anti-entropic process that occurs at w_4 violates the second law of

thermodynamics, and even though the laws of thermodynamics are not fundamental or without exception, such a violation of law results in w_4 being farther away from the actual world than w_3. Thus, even though $w3$ and $w4$ are tied with respect to violations of fundamental law and tied with respect to the size of the region of perfect match of matters of fact, w_3 matches the laws of thermodynamics in a way that w_4 does not.

Perhaps there is also another, related way of responding to Elga's objection. Assume that we can determine the laws of a world in a broadly Humean fashion, according to which a world is associated with an ideal system that can be used to derive what is true at the world. As Lewis describes the project:

> "Take all deductive systems whose theorems are true. Some are simpler, better systematized than others. Some are stronger, more informative than others. These virtues compete: An uninformative system can be very simple, an unsystematized compendium of miscellaneous information can be very informative. The best system is the one that strikes as good a balance as truth will allow between simplicity and strength. How good a balance that is will depend on how kind nature is. A regularity is a law iff it is a theorem of the best system." (1994: 231)

On such an approach, we distinguish between laws and accidental generalizations by determining whether the generalization in question is a theorem of the system that achieves the best balance of simplicity and strength. The strength of a system is a measure of its informativeness, or the number of possibilities it rules out. Determining the simplicity of the system is a somewhat more delicate matter: we certainly do not want an account of simplicity that relies on the psychology of a collection of individuals, or on the syntax of some language used by these individuals, lest the account of laws be too subjective. Yet we

need not admit defeat so easily: there must be some notion of simplicity that supports the judgment that, for example, "a linear function is simpler than a quartic function" (*ibid*).

The best system account is often favored by those with Humean scruples: we can use such an approach, it is argued, to describe the laws of a world without appeal to any "spooky" unanalyzed modality, either by way of counterfactuals are nomic necessities. The fabric of the world is merely a collection of objects and their properties, and the laws of the world are merely collections of generalizations about those objects. An account of laws along these lines could make for good ideology. If these kinds of generalizations are to play a role in a theory of counterfactuals, however, then a non-modal account of them is *essential* to our theory: after all, if the analysis of counterfactuals contains an important role for laws, then laws cannot be understood in counterfactual (or otherwise problematically modal) terms on pain of circularity.

According to such an approach, the laws are the generalizations of this "best system." But what about the other theorems that comprise the system? Other kinds of facts about the state of the word might earn a place in the best system, even if we would not want to call these kinds of propositions *laws*. These facts are useful with respect to determining what the world is like, even if they are not generalizations in the best system of the world. Consider, for example, the past hypothesis, according to which the universe began in a state of low entropy.[39] The past hypothesis, along with the laws of statistical mechanics, can explain why the entropy of large systems increases over time.[40] Such a fact is necessary to

---

[39] See Albert [2000] for a discussion of the past hypothesis.

[40] The reduction of such laws to statistical mechanics by determining the comparative volume of paths through phase space, along the lines of Boltzmann is not by itself sufficient to explain the development of macroscopic systems. These paths through phase space are themselves governed by the dynamic laws of motion, and are themselves time-reversible. It is true that the vast majority of paths through phase space that begin in a region of low entropy wander into regions of high entropy (and ultimately equilibrium), and that once a path is in a region of high entropy, it is highly likely to

explain why systems evolve in the way that they do, and can be stated in a way that is quite simple. It is likely, then, that the past hypothesis earns a place in the best system, given its simplicity and its ability to explain the behavior of macroscopic systems.[41]

The idea, then, is to add the class of statements like the past hypothesis into the similarity metric. Such claims are not laws, but rather theorems about the nature of the world. Even though such theorems are not laws, they are still special kinds of claims about the nature of the world, and thus worlds in which these theorems are false are worlds that are importantly different from the actual world. Our revised similarity metric, then, would be as follows:

1) It is of first importance to avoid big, widespread diverse violations of nomological law.
2) It is of second importance to maximize the spatiotemporal region throughout which perfect match of particular fact prevails
3) It is of third importance to avoid even small, localized, simple violations of nomological law

---

remain there. Noting this, however, does not give us anything like an arrow of time, since these paths run both forwards and backwards through time. A low entropy state will probably be in a state of higher entropy in the future, but it will have also probably have been in a state of higher entropy in the past, if we can only judge how a system is likely to evolve by the comparative volume of these paths.

[41] One might argue that the past hypothesis would *not* be part of the best system of the world, since even if it plays an important role in the reduction of the laws of thermodynamics, it is redundant if the initial conditions of the world are also a part of the best system. Yet I would not want to be committed to the claim that the initial conditions *are* part of the best system. First, they may score poorly enough on the simplicity requirement that gains made in informativeness will not be sufficient. Second, it may not be the initial conditions that are so highly predictive, but rather the conditions of the universe at *any* time, if the fundamental laws are time-reversible. Nonetheless, even if we concede that claims such as the past hypothesis would not be part of a "best system," we could still hold that such claims are important with respect to the evaluation of counterfactuals; that is, there could be a category of sufficiently informative, sufficiently simple sentences that are "useful hypotheses" even if they are not contained in the ideal system, and these useful hypotheses could be used to evaluate the nearness of worlds.

3.5) It is of fourth importance to avoid violations of the theorems of the best system that are not laws.

4) It is of little or no importance to secure approximate similarity of particular fact.

This would let us see why w_4 is more distant than w_3: at w_4, the past hypothesis is not true. A world such as w_4 did not begin in a state of low entropy, but rather a state of high entropy (at least in the "infected region"). Thus, even though this world has the same laws as the actual world, w_4 contains other conditions that are quite different, and would therefore be judged as farther away from the actual world than w_3.

Ultimately, either the addition of a requirement to match special science laws or a requirement to match the theorems of the best system that are not generalizations could provide the requisite response to Elga. Dunn's proposal has the advantage of being able to avoid difficult questions about which propositions at a world are part of its "best system," while my proposal has the advantage of avoiding difficult questions about how a Humean ought to accommodate the special sciences into a "best system" of the world. Both approaches, it should be noted, threaten Lewis's project of locating the "arrow of time" in counterfactuals: in the case of Dunn's approach, this is because the arrow of time could be embedded in the laws of the special sciences, and in my approach, this is because the arrow of time is stipulated by the past hypothesis. Nonetheless, the strategy of adding requirements to Lewis's standard similarity metric in light of putative counterexamples suggests that we could also add requirements to Lewis's metric in such a way as to accommodate counterpossibles. Furthermore, thinking of laws in Humean terms gives us some insight into how to understand the different kinds of logical or metaphysical laws of impossible worlds.

3.6 Counterpossibles

The discussion so far has been concerned with the truth conditions of counterfactuals with metaphysically possible antecedents; it has merely been an assumption that a plausible account of such counterfactuals could be generalized so as to make sense of the truth conditions of counterpossibles. Modifying the account to respond to Elga's objection is necessary in the course of giving a complete account of counterfactuals, but we must also make good on the promise of providing a similarity metric that makes sense of counterfactuals with necessarily false antecedents. We have seen that a proposal in the spirit of Lewis's similarity metric is capable of meeting an array of challenges leveled against it. Nonetheless, our ultimate goal is not to defend Lewis's proposal for counterfactuals *per se*, but to offer an analysis that can accommodate counterfactuals with metaphysically impossible antecedents.

There have been few worked-out accounts of how transworld similarity might work in light of the existence of impossible worlds. Nolan [1997], for example, seems to alternately assume that Lewis's account of 'similarity-in-relevant-respects' carries over more or less straightforwardly in the case of impossible worlds at some points, and relies on intuitive judgments about the similarity of worlds at other points. Brogaard and Salerno [2008] hold that a world is nearby to the extent that it holds fixed "background facts" and preserves "relevant" *a priori* consequences, which are relativized to a speaker and a context. Yet even if this notion of a relevant *a priori* consequence can be made sense of, the resulting account of counterpossibles is too speaker-relative to make sense of counterlogicals, or to be of much use in discussions of metaphysical dependence (see chapter 5). The proper way to see how we could expand Lewis's account is to look at an example and see what kinds of considerations may be in play.

Consider, for example, a counterfactual whose antecedent is metaphysically impossible. Consider an example from chapter 1:

(3) If mereological nihilism were true, then there would be no tables.

Presumably, this counterpossible is true. On a 'nearest worlds' approach of counterfactuals, this would mean that the nearest worlds where mereological nihilism is true are impossible worlds where tables do not exist.

A few assumptions for the sake of this example: first, if we are to take (3) as a counterpossible, we should assume that mereological nihilism is false in all possible worlds. This is a substantive claim, of course: not only does composition occur, but it occurs in all possible worlds that contain conditions favorable to how composition actually occurs. Furthermore, we should assume that the antecedent of (3) does not entail a contradiction. It is obviously not a formal contradiction, but we should also assume that no formal contradiction could be derived from a statement of the theory, so as to separate (3) from the class of counterlogicals.

Why should we think that the nearest worlds where mereological nihilism is true are worlds without tables? There are, after all, impossible worlds where mereological nihilism is true, but tables exist. One kind of world is w_5: such a world is very much like our own, and everything that is true in the actual world is true in w_5 as well. But w_5 is an impossible world where we add the proposition *that mereological nihilism is true* to the set of truths of the actual world. Such a world is, at least on one measure, very similar to the actual world: it differs from the actual world by the truth of only one proposition!

Clearly, this kind of similarity cannot be a guide to the nearness of worlds. That is, the kind of match of facts that a world such as w_5 bears to the actual world does not make it appropriately nearby because such a match comes at the expense of a violation of logical law.

Recall that a world such as w_5 contains all the truths of the actual world, with the addition of the proposition *that mereological nihilism is true*. Since mereological nihilism is both true and false, w_5 contains a logical contradiction; any match of matters of fact it bears to the actual world are not important, given this contradiction.

Of course, worlds such as w_5 are not the only worlds at which mereological nihilism is true, but where tables exist. Consider a world such as w_6, which contains all the truths of the actual world, with the exception of the proposition *that mereological nihilism is false*, and instead contains the proposition *that mereological nihilism is true*. This world, then, lacks the logical contradiction that gets us in trouble in w_5, but it will have other logical inconsistencies instead: consider, for example, the proposition *that if there are composite objects, then mereological nihilism is false*. This proposition is, by hypothesis, true in w_6, as is the proposition *that there are composite objects*. Yet the consequent of this proposition is false in this world; thus, world w_6 contains a clear failure of *modus ponens*. Once again, the impressive match of matters of fact that this world bears to the actual comes at the expense of a violation of logical law.

It should now be clear how to approach the generalization of Lewis's similarity metric: just as Lewis earlier counseled that we avoid large violations of natural law, we can also hold that it is important to avoid violations of logical law. A world cannot be relevantly nearby if it achieves a match in matters of fact and matters of nomological law at the expense of violating logical laws.

There are other ways in which mereological nihilism might be true, and that tables would exist. Perhaps, in some worlds, tables are simples; call a world such as this w_6. In such worlds, there need not be any logical contradictions; it is not true, for example., that in w_6, the proposition *that tables have parts* is true and *that tables do not have parts* is true as well. Rather, these are worlds where tables exist in virtue of violating some metaphysical, rather

than logical, law. For example, w_6 may be a world where tables are extended simples,[42] or worlds where tables are atoms and without extension. Nonetheless, even if a world such as w_6 does not contain a formal contradiction, it must be less similar to the actual world than a world where mereological nihilism is true and tables do not exist. That is, the match of matters of fact that occurs in light of there being tables in both the actual world and w_6 is not important, because a world such as w_6 contains a larger "metaphysical miracle" than the world where mereological nihilism is true and tables do not exist.

We have already seen that it is necessary that our similarity metric hold that matches of matters of fact should not come at the expense of some logical contradiction; it is now clear that such matches also should not come at the expense of larger changes in *metaphysical* law, either. We are, then, in a position to state the revised similarity metric in full:

1) It is of first importance to minimize the degree of violation of logical laws.
2) It is of second importance to minimize the degree of violation of metaphysical laws.
3) It is of third importance to avoid big, widespread diverse violations of nomological law.
4) It is of fourth importance to maximize the spatiotemporal region throughout which perfect match of particular fact prevails
5) It is of fifth importance to avoid even small, localized, simple violations of nomological law
6) It is of sixth importance to avoid violations of the theorems of the best system that are not laws.
7) It is of little or no importance to secure approximate similarity of particular fact.

[42] See, for example, McDaniel [2007].

There is one last issue, however, that needs to be discussed: What is it for a violation of metaphysical or logical law to be large or small? A violation of nomological law can be great or small to the degree that such a violation is "large, widespread, and diverse;" this formulation (at least arguably) makes sense because these violations of law occur in regions of spacetime. But violations of logical or metaphysical law do not obviously have any relation to regions of spacetime. So how could we judge their size?

3.7 Similarity and Laws

What counts as a greater or lesser violation of a logical or metaphysical law? The idea is that a violation of law is great or small to the degree that the world in question has logical or metaphysical laws that are very different from the laws of the actual world. It cannot be that the expressions used to describe these laws resemble each other to a greater or lesser degree. How these laws are represented, in terms of natural language or some language of thought, is far too subjective for present purposes. But what other way of comparing the similarity of laws is there? This is a version of a problem that the defender of a Humean account of laws already faces, in the guise of presenting a suitably objective notion of simplicity. As long as factors such as simplicity enter into the determination of what the laws are, there must be some reasonably objective method of determining what makes one candidate law more simple than another; such a determination of simplicity cannot rely on anything speaker-dependent and remain objective enough for these purposes. The simplicity of some candidate law cannot depend on how it is formulated, since simplicity cannot depend on which linguistic representation we use or which (abundant) properties are used to formulate the theory. In defense of a conception of simplicity that is not tied too closely to our interests, Lewis suggests that we formulate the laws in terms of perfectly natural

properties; doing so makes the simplicity problem at least tractable (see, for example Lewis [1983] and Loewer [1996]). If such a suggestion can help determine how simple a law is, it can also be used to determined how similar two candidate laws are: once we have a kind of ideal language to formulate the laws, we can compare how these laws are expressed in that language to determine how similar they are. Laws could be similar to the extent that they refer to the same kinds of objects or the same properties or relations in this privileged language. On one version of the story, some objects and properties are *fundamental*, and it is the sharing of these fundamental objects, properties, and relations that make for objective similarity.[43]

Determining the logical laws of a world, then, proceeds in a broadly Humean fashion: the logical laws are the generalizations that represent how the propositions of a world are related to one another in the best systematization of what is true at that world. We can determine what the logical laws of a world are in a way analogous to how we might determine the physical laws of a world. If an impossible world is a set of structured propositions (as in chapter 4), we can evaluate candidates for a systematization of a world by how well such a system accounts for the truths of that world. A good systematization of a world will consist of a deductive system that derives the truth of propositions at a world from its other propositions in such a way that strikes a balance of informativeness and simplicity. The tradeoffs will look very familiar: a strong but complex system could simply list all the propositions true at a world; a simple but weak system could only list one (non-maximal) truth.

[43] I advance this theory of the similarity of laws *very* tentatively. Such a theory might be problematic in light of chapter 5, where I discuss fundamentality in terms of counterpossibles.

A world that behaves classically with respect to conjunction, for example, is a world such that whenever a proposition of the form $\varphi \wedge \psi$ is true, so are the propositions φ and ψ; similarly, whenever propositions of the form φ and ψ are true, so is $\varphi \wedge \psi$. Rules of conjunction introduction and elimination, then, would be capable of systematizing the facts of this world with respect to conjunction; we can account for the truth of the propositions in this world (at least in part) by understanding how conjunction behaves. The (classical) logical rules with respect to conjunction would earn their keep in the system that makes sense of what is true at these worlds. A rival law of conjunction for this world could be ruled out on the grounds that it is too complicated (such as, from $\varphi \wedge \psi \wedge (\psi \rightarrow \varphi)$, infer that $\varphi \wedge \psi$) or not as informative (such as, from $\varphi \wedge \psi$, infer that $\varphi \rightarrow \psi$).

One should expect ties when determining the logical laws of a world: there may be no way of determining which connectives to take as primitive and which to take as derived, and there may be no way of determining whether it is better to derive the various atomic truths of a world from a maximal truth, or *vice versa*. Perhaps there are many equally good formulations of the laws of a world, or perhaps there is one proper formulation but no way of determining which one is correct; we need not make this choice for present purposes. What *is* important, however, is that we can determine some plausible candidates for the logical laws of a world, and compare those candidates to the laws of other worlds. The logical laws of a world are given by the best systems of those worlds that account for all the logical facts; we then determine the degree of violation of logical laws in some world by determining the degree to which the logical laws of that world differ from the logical laws of our world.

The same approach is applied to determine the metaphysical laws of a world: the metaphysical laws of a world are also determined by the "best system" that properly

accounts for the facts of a world that relate to composition, the nature of properties, time, and so on. Mereological universalism is true at a world when that theory best explains the facts of how objects and their parts are related at a world: in particular, mereological universalism is true at a world where any collection of objects compose a further object, as the theory of mereological universalism is an apt generalization of how composition occurs. Similarly, a world that contains *no* composite objects is a world where mereological nihilism is true, as that law would be an apt generalization of the mereological facts of the world in question. A world that contains exactly one object may be a world where the two theories are tried with respect to determining the metaphysical laws of the world; perhaps both laws are true of this world, or exactly one law is true, but we cannot know which. And once we can meaningfully determine what the metaphysical laws of a world are in such a way that account for all the metaphysical facts, we can determine which worlds have metaphysical laws that are more or less similar to the laws of the actual world. The worlds whose metaphysical laws are more similar to our metaphysical laws, as expressed in some suitably canonical system, are worlds that contain smaller violations of metaphysical laws.

Even among worlds that have different metaphysical laws, there is room for these violations of law to be greater or smaller. In the case of a counterpossible such as (3), we should expect worlds where mereological nihilism is true to be worlds with different kinds of objects than the actual worlds. But we should not expect these worlds to needlessly change their metaphysical laws, as would be required in worlds such as w_6, where metaphysical laws about the nature of extension must be altered as well as laws related to composition.

3.8 Conclusion

It is, of course, impossible to discuss every interesting counterpossible, and show why the intuitively correct world is more similar to the actual world than any other impossible world one might consider. Nonetheless, the above discussion should provide sufficient guidance with respect to how the considerations of nearness discussed above would bear on further examples. A world – possible or otherwise – is close to the actual world to the extent that, first, in minimizes violations of logical laws; second, to the extent that it minimizes violations of metaphysical laws; third, to the extent that it avoids large violations of nomological laws; fourth, to the extent that maximizes the region of perfect match of matters of fact; fifth, to the extent that it avoids small violations of nomological laws, and finally, to the extent that it avoids violations of non-lawful theorems of the best system (or violations of special science laws). Counterpossibles are evaluated just as counterfactuals are: we determine what the laws of the world are in a broadly Humean fashion, and judge how different those laws are from the laws of the actual world.

The approach to counterfactuals discussed above is unified, in the sense that there is a single standard similarity metric for all counterfactuals. The similarity metric is also a generalization of Lewis's similarity metric, in the sense that it does not differ from his with the respect to the truth of any counterfactual whose antecedent is possible, at least assuming we amend his similarity metric in response to Elga's challenge as discussed above. The metric is derived both from considerations about the truth values of counterpossibles as well as Humean approaches to law extended so as to make sense of logical and metaphysical laws.

CHAPTER 4

WHAT ARE IMPOSSIBLE WORLDS?

It is well-known that the framework of possible worlds suffers from certain difficulties if it is to be used as an analysis of certain philosophical phenomena. Possible worlds are coarse-grained, in the sense that the space of possible worlds models only what is logically or metaphysically possible. Thus, analyses that make use of possible worlds face problems if they are to differentiate content that is logically equivalent. The propositions *that nine is prime* and *that I am a married bachelor* are both impossible, but they are distinct. I could, for example, believe one without believing the other; different things might be the case if one of these propositions were true, but not the other. In this chapter, I address some of these difficulties by supplementing the possible worlds framework with impossible worlds, which represent ways the world could *not* be. I develop an account of impossible worlds of a certain kind, and show how they allow us to give better analyses of philosophical concepts than we would be able to do with possible worlds alone.

In section 1, I will briefly discuss some of the limitations of possible worlds and potential applications of impossible worlds. In section 2, I will develop an account of impossible worlds that is both ontologically innocent and robust enough to do the work required of these worlds. In section 3, I respond to a version of an argument against the existence of impossible worlds advanced by Lewis.

4.1 Uses of impossible worlds

Why believe in impossible worlds? One kind of argument aims to show that we are *already* committed to the existence of impossible worlds. In his [1973], Lewis argues that

possible worlds are simply ways things could have been. We all believe that there are ways things could have been, and possible worlds are merely ways things could have been, so we all believe that there possible worlds. Lewis backed off this argument, in least in part because we also all believe there are ways things *couldn't* have been, as pointed out by Naylor in his [1986]. Since Lewis did not want to use impossible worlds to represent the ways things could *not* have been, he rejected his original argument for possible worlds. But, we could also easily turn Naylor's modus tollens into a modus ponens. Since there are obviously ways things could not have been, there are obviously impossible worlds.[44]

We might also think that we are committed to impossible worlds because their existence is entailed by other theoretical commitments. As King argues in [2007], depending on what we take possible worlds to be, we should accept the existence of impossible worlds as well. If possible worlds are complete and consistent sets of sentences or propositions or states of affairs, there is no principled reason why we shouldn't also believe in incomplete or inconsistent sets of sentences, propositions, or states of affairs.

The argument I want to defend here, however, does not rely on how we are to understand "ways," nor does it rely on our commitment to sets and propositions. Rather, I hold, with Lewis, that possible worlds earn their ontological keep by the use they can be put to in most fields of philosophy, playing an important role in explicating concepts in logic, metaphysics, epistemology, and ethics. Yet the work done by possible worlds could be *better* done with the addition of *impossible* worlds. Consider, for example, the kind of work Lewis puts possible worlds to in his [1986]. Lewis holds that possible worlds are extremely useful in explaining modality, closeness, content, and properties. If we consider each of these

[44] This argument is made in more detail by Vander Laan in his [1997]
.

applications in turn, we can see that our understanding of each of these families of concepts can be improved with the addition of impossible worlds.[45]

4.1.1 Modality

The most obvious application of possible worlds is to explain modality. By quantifying over worlds, we can develop a much better understanding of what it is for something to be possible, necessary, or actual.[46] We can understand the sense in which the laws of physics are necessary (and thus locutions such as 'It is not possible to travel faster than the speed of light') as restricted quantifications over possible worlds. That is, 'Nothing can travel faster than light' is true because in all *nomically accessible* worlds (i.e., worlds which have the same laws of physics that our world has), nothing travels faster than the speed of light. Of course, there are possible worlds that have different laws of physics, and at these worlds, some things travel faster than light; nonetheless, these are not the worlds we were quantifying over. If we say, 'The coin could land tails,' we usually mean that there are nomically accessible possible worlds where the coin *does* land tails. Similar stories can be told for other flavors of necessity: some worlds share our history, some worlds share facts about what is practical, some worlds correspond to our moral or legal obligations, and so on.

---

[45] This kind of argument is similar to that of Yagisawa [1986], though I reject his account of extended modal realism.

[46] It is worth noting that Lewis does not think that acceptance of possible worlds is required for the semantics of modal and non-standard logic: we can make sense of the "possible worlds" invoked by the semantics of logic with set theory alone. Nonetheless, a more robust notion of possible worlds could be useful to make better intuitive sense about what is being claimed about accessibility (and thus, modal claims) by the supporters of different logical systems in a given discussion [1986: 18-20]. If we need "real" possible worlds to make sense of what kind accessibility relation to accept in an account of modal terms, then it is obvious that we need "real" impossible worlds to make sense of arguments about which non-classical logic to accept in an analysis of belief revision, for example. Most semantics of intuitionist or relevance logic, for example, make use of non-standard worlds that contain truth value gaps or gluts.

Epistemic possibility, however, presents a challenge. How are we to make sense of sentences such as 'Goldbach's conjecture could turn out to be false,' or 'relevance logic could be correct?' Assuming that Goldbach's conjecture is true and that classical logic is the proper analysis of the entailment relation, these are mere *epistemic* possibilities: for all we know, there could be some very large even number that is not the sum of two primes; it could turn out that the conclusion of some formal argument has to be relevant to its premises. There is (presumably) no possible world where Goldbach's conjecture is false, and no possible world where relevance logic is correct. Nonetheless, these claims certainly seem to be existential quantification over worlds, in the same pattern as 'The coin could land tails.' Impossible worlds would provide an attractive and straightforward way of understanding claims about epistemic possibility. A logically impossible proposition could still be epistemically possible if we can quantify over impossible worlds. The claim that Goldbach's conjecture could be false, then, should be understood as 'At some epistemically accessible world, Goldbach's conjecture is false,' even if there is no such possible world. The epistemically possible worlds are just the worlds we *think* are possible.[47]

4.1.2 Closeness

Possible worlds are also required to make sense of *closeness*. An account of closeness is required to make sense of counterfactuals: when we say something like, 'If I had missed the bus, then I would be late for class,' I plausibly mean something like 'In the closest

[47] To be sure, impossible worlds aren't the only way of understanding epistemic possibility, though they probably are the simplest. Other accounts face various problems. Perhaps the leading contender for understanding epistemic possibility is two-dimensional semantics (see, for example [Chalmers 2004], according to which we can make sense of some epistemic possibilities by analyzing the primary intensions of the propositions expressed. This strategy is helpful in understanding some epistemic possibilities; two dimensions semantics is helpful in understanding how we could conceive of water as not being H_2O. It is less clear, however, how two dimensional semantics would help us understand how Goldbach's conjecture or how some non-classical logic could be true.

situation where I miss the bus, I am late for class.' The best way to make sense of such a 'closest situation' is to think of possible worlds ordered by some closeness relation, as in [Lewis 1973]. We can then determine whether a counterfactual is true (roughly) by determining whether the consequent is true in the closest possible world where the antecedent is true. But, it seems that there can be counterfactuals with impossible antecedents. Consider 'If nine were prime, then it would not be divisible by three.' Impossible worlds allow us to retain the structure of Lewis's account of counterfactuals: we can say that this sentence is true because, in the nearest (impossible!) world where nine is prime, it is not divisible by three.[48]

Similar considerations apply to the comparative closeness of theories. False physical theories can also be close to the truth to varying degrees, and we can model this in terms of the comparative closeness of possible worlds where these theories are true. This seems quite plausible, but we might also wonder how we are to analyze our intuitions about certain necessarily false theories being closer to the truth than other necessarily false theories. Impossible worlds allow us to retain the analysis offered above: impossible worlds where some necessarily false theories are true are closer to the actual world than other impossible worlds where different necessarily false theories are true.

4.1.3 Content

Possible worlds are also useful in discussing the content of propositional attitudes such as belief. We can represent what some agent believes as the set of possible worlds that she thinks might be actual. We can say that some possible worlds are *doxastically accessible* to some agent: these are the worlds that the agent think might be the actual world. An agent

[48] This proposal is discussed in [Nolan 1997], [Vander Laan 2007], and, of course, chapter 3 of this work.

thinks some propositions might be true iff that proposition is true in some of her doxastically accessible worlds; she believes some proposition iff that proposition is true in every doxastically accessible world.[49] We then represent how an agent makes inferences by updating the set of doxastically accessible worlds with the acquisition of new information: as an agent learns something about the world, she narrows the set of worlds she takes to be actual.[50]

This picture, however, makes several unreasonable assumptions about the rational behavior of agents. Agents are represented as being logically omniscient and perfect reasoners. They believe all logical truths, since logical truths are true at every possible world, and therefore true at every doxastically accessible world. The agents described in this model believe no contradictions, since contradictions are not true at any possible world, and therefore not true at any doxastically accessible world. Finally, these agents believe the logical entailments of all their beliefs, since if P entails Q, and P is true at every doxastically accessible world, then Q is true at every doxastically accessible world as well. Assuming we want these agents to be more realistic in these respects, we should look for ways to relax these assumptions. Realistic agents must be permitted to believe the impossible, fail to believe the necessary, and fail to believe some things that follow from their other beliefs.[51]

Impossible worlds provide the most straightforward remedy for this problem. There are impossible worlds at which contradictions are true, impossible worlds at which not all

---

[49] See, for example, [Hintikka 1962]. The model presented is a simplification of Lewis's scheme, which makes better sense of belief *de se*.
[50] See, for example, [Stalnaker 1987].
[51] Some might cede these points, and hold merely that the formal model described here is merely an epistemic ideal; the model describes perfect reasoners, and thus describes a kind of normative goal for reasoning. It is certainly useful to keep the Hintikka model as a kind of ideal for rationality, but the extended model with impossible worlds will allow us to model more than just ideal rationality. If we can describe the behavior of realistic agents, we can use it to make predictions about how people will behave and update beliefs, for example.

tautologies are true, and impossible worlds at which some propositions are true without their (classical) entailments. It would, therefore, be easy to represent agents who are not logically perfect if we allow the doxastically accessible worlds to include impossible worlds. If an agent believes in some contradiction, then all of her doxastically accessible worlds are impossible worlds that contain some contradiction. If she fails to believe in some logical truth, then some of her doxastically accessible worlds are impossible worlds that lack this logical truth. If an agent fails to believe some proposition Q that is entailed by some proposition P that she believes, then P is true in all of her doxastically accessible worlds, though some of her doxastically accessible worlds are impossible worlds where P is true but Q is not.

4.1.4 Propositions

Possible worlds are also useful in providing an analysis of propositions. Possible worlds also allow us to analyze propositions as functions from possible worlds to truth values (or, equivalently, as sets of possible worlds). The proposition *that grass is green* is the set of all worlds according to which grass is green.

There are, however, intuitively distinct propositions that pick out the same set of possible worlds: *that all bachelors are unmarried* and *that 2 + 2 = 4* pick out the same set of worlds, *viz.* the set of all possible worlds. Impossible worlds could differentiate these necessarily equivalent propositions or properties: there are, after all, impossible worlds where bachelors are married, and impossible worlds where 2 + 2 does not equal 4. At some impossible worlds, both of these are true; at other impossible worlds, one of these propositions is true, but not the other. Thus, the set of all worlds (possible and impossible)

where bachelors are unmarried is not the same set as the set of all worlds where 2 + 2 = 4.[52] One could then either hold that a proposition is a *non-empty* set of worlds, or if there were good theoretical reason to do so, use the empty set to represent the content of certain non-truth evaluable expressions.

To be sure, the defender of a possible worlds account of properties and propositions has tools at her disposal to represent these entities in a more fine-grained fashion. For example, one could use possible worlds to describe structured propositions, which allow us to individuate necessarily coextensive properties and propositions. Though there are many accounts of this kind of structured entity,[53] we can illustrate this strategy with the system Lewis sketches in [1986:56-59]. The project is to augment the possible worlds account with some kind of structure that mirrors the structure of the language used to express the proposition. On an unstructured account of propositions, for some proposition P, the propositions P and ~~P pick out the same set of possible worlds (*viz.,* the set of all worlds where P is true). A structured account of propositions could represent negation as a relation N that holds between unstructured propositions and the sets of worlds where that proposition does not hold. The structured proposition ~P could then be represented as <N, P>, while the structured proposition ~~P could be represented as <N, <N, P>>. On the structured account of propositions, P and ~~P are distinct propositions, since they pick out different 'tuples.[54]

---

[52] Impossible worlds can also be used to distinguish necessarily coextensive properties. We might think that triangularity and trilaterality are distinct properties, even though every three-angled figure is also a three-sided figure across all possible worlds. Impossible worlds allow us to represent figures that are triangular but not trilateral, or vice versa; thus, the two properties pick out different figures in different impossible worlds.

[53] See, for example, King [2007] and others.

[54] To be sure, structured propositions will not solve every problem of hyperintensionality. One could believe one, but not the other, of two necessarily equivalent and structurally isomorphic structured

If structured propositions give us a fine-grained notion of properties and propositions, why do we need impossible worlds? If we can come up with some entity that enables us to make the distinctions we need to make between properties and propositions, why not simply make use of them and not take on the additional ontological burden of impossible worlds? Because we can do more with worlds than we can with structured propositions alone. It is, of course, useful to give some perspicuous account of propositions in the course of metaphysics. But we cannot simply point to some class of set-theoretic structures that identify and differentiate content in the right way and think that our work is done. Our account of propositions must differentiate distinct propositions, but it must also do more: it must also pick out the objects that are able to play the role that is required of them by modality, counterfactuals, the content of propositional attitudes, and so on. Structured propositions alone cannot do this work. As seen above, *worlds* are required to give an account of modality, content, and closeness. It is *worlds* that stand in the similarity relations discussed in chapter 3; it is *worlds* that can be used to model how agents (ideal or otherwise) reason; it is *worlds* that represent ways what a believer takes to be true. Structured propositions can give us fine-grained entities to individuate certain properties and propositions, but they cannot be put to the same use that worlds are.

---

propositions. Whether or not this is a problem for the defender of the structured proposition account depends on whether there is some plausible alternative treatment for such cases. For example, if the equivalent and isomorphic propositions in question are *that London is pretty* and *that Londres est jolie*, then it is plausible that the reason that Pierre would assent to one but not the other are metalinguistic in nature. Certainly, if Pierre realizes that the two sentences are equivalent (or even that 'London' and 'Londres' are names for the same city), then he would retract his acceptance of one of these propositions. To be sure, this account is threatened if *all* hyperintensional content can be treated metalinguistically, but this does not seem to be the case. Pierre's confusion about 'London' and 'Londres' is not akin to how some logicians understand non-classical entailment relations, and it would border on absurdity to hold that nobody grasps the meanings of their terms, since everyone falls short of ideal rationality.

An adequate account of propositions is one on which we can put them to the kinds of uses argued above. Propositions are not idle pieces of metaphysics, but rather objects that have modal properties, stand in entailment relations, counterfactually entail other propositions, are the proper targets of certain propositional attitudes, and so on. If an impossible proposition is merely some ordered set, we do not have any handle on how to understand epistemic possibility, counterpossibles, or belief in the impossible. To say that if A were true, B would be true, for example, is to say (roughly) that in the nearest world where A is true, B is true. It is unclear what to make of these instructions if A and B are merely sets: there is no theory on the table of the comparative nearness of sets. Of course, one could attempt to develop such a theory, but the considerations that are generally in play with respect to the nearness of worlds (*e.g.*, matching regions of matters of fact or preserving laws) are more naturally applied to worlds than they are to sets. If we have impossible worlds, however, these questions of how to understand the work done by these propositions become tractable, since worlds provide the kind of framework that allows us to meaningfully talk about modality, entailment, counterfactuals, belief content, and the like.

4.1.5 Impossible worlds and triviality

There is at least one serious worry about these uses of impossible worlds. The reason that possible worlds play such a central role in describing modality, closeness, belief content, properties, and propositions is that possible worlds allow us to discuss all and only *possible* scenarios. If impossible worlds are to play a role in our treatment of these concepts, we risk trivializing our analyses. For example, we can use impossible worlds to distinguish the proposition *that nine is prime* from *that some bachelor is married*, since there is some impossible world where *that nine is prime* is true, but not *that some bachelor is married*; thus, the

sets of worlds are not the same. However, it seems that we will also be forced to distinguish the propositions *that some bachelor is married* and *that some bachelor is wedded* since, at least depending how the worlds are specified, there could be impossible worlds where one is true, but not the other. Intuitively, however, these *are* the same proposition; the content of what is said should not be affected merely by how we express it. Similarly, our models of belief should not license outlandish inferences, and any realistic agent should be expected to make certain simple inferences, even if she fails to make more complex ones. Formal epistemology would be trivialized if the structure provided no norm for how agents *should* reason, and provided no predictions for how agents likely *will* reason. This seems true even if there are cases in which it is vague whether two propositions are the same, or vague whether some inference is outlandish or reasonable. If the space of impossible worlds lacks the inferential structure of the space of possible worlds, then it is unclear what sense we can make of how agents update their beliefs. The worry, then, is that impossible worlds lack the kind of logical structure that makes possible worlds an attractive tool for explaining many philosophical concepts.

The solution is to restrict which impossible worlds are used when giving our accounts of modality, properties, propositions, beliefs, and so on. We determine which impossible worlds are relevant to our analysis by considering the worlds that are close to the actual world in some relevant sense. For example, a proposition can be analyzed as the set of some restricted class of worlds where it is true; context determines whether we analyze the proposition as the set of all possible worlds where it is true, the set of possible worlds and some impossible worlds, or the set of all worlds, possible and impossible. My response to the worry of triviality is that our analyses are only trivial when we allow *too many* impossible worlds into the analysis. Rather than allow all impossible worlds to factor into

our analyses of modality, propositions, content, and the like, context selects the impossible worlds that are relevant to the task at hand.[55]

4.2 The ontology of impossible worlds

4.2.1 Ersatz Impossible Worlds

It is plausible to think of impossible worlds as ersatz constructions of some sort. On pain of contradiction, they cannot be concrete.[56] Perhaps they are mere fictions of some sort,[57] but even if they are, it would be important to better characterize this fiction if it is to do any explanatory work.

If impossible worlds are constructions, what are they constructed from? Brogaard and Salerno, in their [2008], argue that they are constructions out of sentences of some world-making language. This kind of impossible world is intuitively plausible, since we use language to describe impossible scenarios in a way that cannot be done with concrete worlds or (perhaps) pictures.[58]

A linguistic ersatzist account of impossible worlds inherits many of the difficulties faced by the linguistic ersatzist program in general. Most importantly, it is not clear that a world-making language would have the resources required to describe all the impossible worlds. The linguistic ersatzist faces embarrassing questions about how the appropriate world-making language represents content related to alien individuals or alien properties. Furthermore, it is unclear how one could understand the meanings of the terms of the

---

[55] Context could also select a relation among the structured propositions that would individuate them for the purposes at hand. Perhaps the structured propositions would be divided into equivalence classes, but it is also likely that which propositions bear the same content will be vague, even holding a context fixed.
[56] See [Lewis 1986, fn 3]. I return to this argument below.
[57] See, for example [Rosen 1990].
[58] See, for example, [Sorensen 2002]

world-making languages without the kind of semantic tools that already rely on possible worlds: the referring terms and predicates of the world-making language cannot express a function from possible worlds or individuals at possible worlds, as they traditionally do. The ability of this language to express impossibilities would be constrained by its ability to name individuals and properties that do not exist in the actual world, as well as the grammar of the world-making language. An account of ersatz impossible worlds that does not rely on language would be more fruitful.

4.2.2 Worlds as simple set theoretic constructions

A more promising account of impossible worlds is one in which impossible worlds are set-theoretic constructions of possible worlds, where possible worlds are not themselves set-theoretic constructions. Creating impossible worlds in this fashion has many advantages: it is as ontologically innocent as possible worlds, as they are merely constructions of possible worlds. Furthermore, if the possible worlds remain basic in a way that the impossible worlds are not, then we can continue to use the possible worlds to describe metaphysical possibility and necessity without threat of circularity. We can consider a few ways of constructing these impossible worlds to determine whether they will be able to play the role required of them in section 1.

Rescher and Brandom, in their [1979], describe what they call *schematic* worlds, at which there are truth value gaps, and *inconsistent* worlds, at which there are truth value gluts. These worlds are created from the possible worlds. Consider two possible worlds, w_1 and w_2. Some proposition P is true at w_1, and ~P is true at w_2. We can now describe a schematic world by a process they call *schematization*, or *world conjunction*: $w_1 \cap w_2$ describes a world such that, for any proposition ϕ, ϕ obtains in the world iff ϕ obtains in w_1 and w_2.

ϕ is true at $w_1 \cap w_2$ iff ϕ is true at w_1 and ϕ is true at w_2.

Note that, in this schematic world, P does not obtain, since P is not true at w_2 and ~P does not obtain, since ~P is not true at w_1. We can describe an inconsistent world by a process they call *superposition,* or *world disjunction*: $w_1 \cup w_2$ describes a world such that, for any proposition ϕ, ϕ obtains in the world iff ϕ obtains in w_1 or w_2.

ϕ is true at $w_1 \cup w_2$ iff ϕ is true at w_1 or ϕ is true at w_2

Note that, at this inconsistent world, P is true, since P is true in w_1, and ~P is true, since ~P is true in w_2.[59]

Not that, even though P and ~P can both be true in an inconsistent world, P ∧ ~P will not be true at any world, since that conjunction is not true at any possible world; similarly, even though neither P nor ~P can be true in a schematic world, P ∨ ~P will always be true, since that disjunction is true in all possible worlds.

Can Rescher-Brandom worlds do the work required of impossible worlds? These inconsistent and schematic worlds describe ways things couldn't be, since they give us worlds where P and ~P are both true, or where neither P nor ~P is true. As we have seen, however, they do not give us worlds where contradictions are true. This means impossible worlds do not describe some of the ways the world couldn't be: namely, they don't describe worlds where P ∧ ~P obtains, or worlds where P ∨ ~P fails to obtain. Rescher-Brandom worlds, then, would conflate all of these impossibilities; if propositions are represented by the set of all worlds where they obtain, we would not be able to differentiate intuitively distinct contradictions, since they all pick out the set of no worlds. Similarly, Rescher-

[59] Note that schematization and superposition can also be applied to the non-standard worlds, so a world can be both schematic *and* inconsistent, if we take the superposition of two schematic worlds, or the schematization of two inconsistent worlds.

Brandom worlds would be unable to represent the beliefs of logically imperfect agents who believe contradictions or fail to believe tautologies.

We could attempt to answer this problem by modifying the approach in such a way that the semantics given above hold only for atomic formulae, and changing our semantics for the evaluation of complex formulae[60]. Roughly following the approach of [Restall 1997], we create non-standard worlds by superposing or schematizing the atomic formulae of possible worlds, which gives us the truth of atomic formulae, but *not* the complex formulae.[61] Instead, we can define how the connectives work at worlds the following way:

$\sim\phi$ is true at w iff ϕ is false at w[62]

$\phi \wedge \psi$ is true at w iff ϕ is true at w and ψ is true at w.

$\phi \vee \psi$ is true at w iff ϕ is true at w or ψ is true at w.

Thus, we can have inconsistent worlds where $P \wedge \sim P$ is true: in our original example, the superposed world $w_1 \cup w_2$ describes such a case. Since P is true at w_1 and $\sim P$ is true at w_2, $P \wedge \sim P$ is true at $w_1 \cup w_2$. Analogously, at schematic worlds, $P \vee \sim P$ will not be true: at the schematized world $w_1 \cap w_2$, P is true at w_1 but not w_2, and $\sim P$ is true at w_2, but not w_1. Thus, on such a system, we are able to describe non-standard worlds where logical falsehoods are true and worlds where tautologies are not true.

---

[60] Note that we shift here to discussion of atomic and complex formulae, rather than propositions.
[61] This is also similar to the approach of [Priest 1979]. Restall and Priest only consider worlds that are inconsistent. The impossible worlds described here, however, can be inconsistent or schematic. See also [Berto 2009], whose impossible worlds are also built from sets of possible worlds.

[62] It might seem more natural to define negation as $w \vDash \sim\phi$ iff $w \nVdash \phi$, but this rules out the possibility of P and $\sim P$ being true at some world together. If we want P and $\sim P$ to be true at some inconsistent world, then we cannot tie the truth of $\sim P$ to the lack of truth of P. Rather, the model must ascribe falsity to some proposition independent of whether it has also ascribed truth to it. Another strategy to allow P and $\sim P$ to be true at a world would be to make $\sim P$ true when P is not true at some world w^*. See, for example, [Dunn 1993].

This proposal still does not provide the kind of worlds required to represent *all* logical contradictions. Our logic is weak, but it will certainly have some theorems: for example, the law of double negation ($\sim\sim\phi \leftrightarrow \phi$) is valid in this framework. In all worlds as described above, $\sim\sim P$ is true whenever P is true. Thus, these impossible worlds do not describe some ways things could not be.

We cannot respond to this problem simply by weakening the logic we have chosen, either. We could stipulate, for example, that the law of double negation is not a theorem, and include in our model worlds where some proposition P can be true without $\sim\sim P$ being true, and vice versa. But this does not address the central worry. As long as we provide some system of worlds that describes some entailment relation, the theorems associated with that entailment relation will be true in all worlds described by the system.

If propositions are sets of worlds as described by this kind of system, then, they simply will not be fine-grained enough. Such a system will conflate the meanings of any two sentences that are true in the same set of worlds, given some system. If it seemed wrong that all propositions that are logically true in classical logic are indistinguishable, then it should seem wrong that all propositions that are logically true in some weaker system are indistinguishable. Thus, if impossible worlds are to let us differentiate properties and propositions that are necessarily coextensive, the structures provided so far are simply not up to the task.

If impossible worlds are to capture the phenomenon of epistemic possibility, these worlds simply do not go far enough. As we have seen, the law of double negation is valid in this system. A formula such as ($\sim\sim P \supset P$) will be true in all worlds, and thus we cannot make any sense of the notion that the failure of the law of double negation is epistemically possible in some scenario.

Similarly, these impossible worlds will not solve the problem of logical omniscience. To be sure, since agents can now be doxastically related to Restall worlds, they can be represented as believing some logical falsehoods such as $P \wedge \sim P$. They cannot, however, be represented as believing other logical falsehoods, such as $\sim(\sim\sim P \supset P)$. The agents represented by this model will be logically omniscient with respect to the theorems of whatever logic is described by the worlds posited by the model: they will believe all of its logical truths, they will fail to believe anything that is logically impossible in the system, and they will accept all the relevant entailments of their beliefs.

A further problem with this kind of proposal is that it is not applicable to many of the potential applications that motivated our project. The kinds of impossible worlds discussed above are all ways of representing various kinds of formal impossibilities: Rescher-Brandom worlds give us structures where propositions such as P and ~P can be both true or both false, and Restall worlds give us structures where propositions such as $P \wedge \sim P$ are true and $P \vee \sim P$ are false. These worlds might be sufficient if our project is to provide a structure to represent certain logical falsehoods, but our ambitions are greater than this. We are not concerned merely with formal contradictions, but impossibilities in general.

Consider propositions such as *that nine is prime* or *that some bachelor is married*. Presumably, it is impossible that nine is prime, or that some bachelor is married, but we still want to be able to represent these propositions in a non-trivial way, if they are to serve as potential targets for belief ('Bob believes that nine is prime'), antecedents of counterpossibles ('If some bachelor were married, his wedding would be well-attended') and so on. But these propositions do not seem to be conjunctions of contraries. At least on the surface, they attribute a property to some object. It would be strange, then, to attempt to

use schematic or inconsistent worlds to describe some scenario where nine is prime: just what would be conjoined or disjoined to make nine prime?

To be sure, the proposition *that some bachelor is married* could be complex in a way that is not evident from its syntax when expressed in English. Perhaps 'bachelor' is to be represented as 'man' and 'not married,' and thus the proposition is analyzed as *that some not married man is married*, which *does* seem to involve a conjunction of contraries. Under analysis, all impossibilities reduce to contradictions; once we have such a formal contradiction, we can analyze it as the superposition of the possible worlds expressed by the contradiction's conjuncts. At some point, however, it becomes unclear how to carry out this strategy. It is simply implausible that all impossibilities can be analyzed as contradictions. How can we tell a similar story for *that nine is prime?* If there is a formal contradiction there, it is harder to uncover, and it is certainly more difficult to argue for some specific formulation of this contradiction as the "correct" representation of the proposition.

4.2.3 Worlds as complex set-theoretic constructions

An adequate account of impossible worlds, then, builds them not from the possible worlds, but rather from something capable of representing formal contradictions as well as impossibilities in a more general sense. Impossible worlds are not merely set-theoretic constructions out of possible worlds, since the schematization or superposition of possible worlds does not allow us to express all the impossibilities required of a robust account of impossible worlds.

4.2.3.1 From Possible Worlds to Structured Propositions

We have already seen one method of using possible worlds to build structured propositions in section 1.4. We can continue to use this system for purposes of illustration, though not much hangs on our choice of how we build structured propositions. This system is able to differentiate intuitively distinct propositions better than the simple set-theoretic impossible worlds. That is, for some proposition P, P and ~~P pick out the same set of worlds for Restall, since the law of double negation holds at all worlds in that system. Lewis's account of structured propositions, however, is able to distinguish the two: <P> and <N, <N, P>> are, after all, distinct 'tuples.

These structured propositions are also capable of representing the impossibilities that Restall cannot: the structured proposition *that some bachelor is unmarried* can be represented simply as the 'tuple <E, <B, M>>, where E is a relation that holds between some property and the worlds where something instantiates that property; B is the unstructured property of being a bachelor, and M is the unstructured property of being married.

Of course, our work is not done yet. For the reasons discussed above, we need impossible worlds if we are to make better sense of modality, closeness, and belief content. Nonetheless, structured propositions such as these are useful in order to describe and differentiate propositions in a more fine-grained way than would be otherwise available. Furthermore, structured propositions give us the tools necessary to construct the impossible worlds that are capable of giving complete accounts of epistemic possibility, the analysis of counterpossibles, and the contents of belief.

4.2.3.2 From Structured Propositions to Impossible Worlds

Structured propositions are ordered sets composed of possible worlds, collections of entities at possible worlds, and relations between and among them. We can then form sets

of these structured propositions to create worlds, which would be able to describe ways things could (or could not!) be. Call these worlds *structured worlds*. Structured worlds are built from structured propositions, rather than from worlds or superpositions or schematizations of worlds. A structured proposition is true at a structured world iff that proposition is a member of the set that composes the world. A structured world could be complete and consistent, but a structured world that is not complete and consistent is an impossible world. If some impossible world is described by the set {A, B}, where A and B are two unstructured propositions, then all that is true at this world are A and B. It does not follow, for example, that ~A is not true, that A ∧ B is true, or that A ⊃ A is true, since these complex propositions are not members of the set {A, B}.

There's little more to say about the behavior of structured worlds. Previous accounts described a logic of impossible worlds, giving truth conditions for the behavior of logical operators and connectives, but there can be no such logic given for structured worlds in general. This is because these impossible worlds are anarchic by design: any set of structured propositions is an impossible world. Thus, there is no interesting logical structure to the space of these worlds. Any proposition could be true at some world if that proposition is a member of the set that describes the world; any other proposition could fail to be true at that world if that proposition fails to be a member of that set. Thus, principles such as the law of double negation and modus ponens will not be valid in these worlds.

Since there is no logic of structured worlds, they are flexible in a way that their competitors are not. Impossible worlds are capable of describing *any* way that the world could not be, since structured propositions have the resources to express content in a way

that is arbitrarily fine-grained. Impossible worlds will therefore be capable of representing epistemic possibilities that possible worlds by themselves cannot.[63]

Furthermore, worlds such as these can give a formal epistemologist the structure required to represent agents that are not logically omniscient with respect to any logical system. Since anarchic impossible worlds are arbitrary fine-grained, there are worlds that can represent any contradiction; an agent can be doxastically related to worlds where any logical principle fails; an agent can fail to be doxastically related to any world where any given logical truth holds; an agent can believe some proposition without believing some other proposition that is entailed by it in any logical system.

If structured worlds are these kinds of set theoretic constructions, it is clear that we should accept them in our ontology. They are cheap: they are merely sets composed of things we already accept. We should believe in structured propositions because they are merely set-theoretic constructions of possible worlds, the entities in then, and relations between and among these things; we should believe in impossible worlds because they are sets of structured propositions. Furthermore, they are useful: if we can argue from the utility of possible worlds to their existence, then the same kind of case can be made for structured worlds. If we want to understand epistemic possibility, model the beliefs of rationally fallible agents, analyze propositions and properties in a more fine-grained fashion, give truth conditions to counterpossibles, and so on, structured impossible worlds are up to the task.

4.2.3.3 Structured Worlds and Counterpossibles

[63] As mentioned above, it is likely that a separate account will have to be given for certain other cases, such as the beliefs of agents who do not know the meanings of the terms they use to describe their beliefs.

Of course, if structured impossible worlds can accomplish all these tasks, they would be ideal vehicles to represent the impossible worlds needed for an analysis of counterpossibles. Since they are merely sets of structured propositions, sets of impossible words could represent the impossible antecedents of counterpossibles in an appropriately fine-grained fashion. Furthermore, as collections of structured propositions, they are capable of being systematized in the way suggested in chapter 3; we can meaningfully describe the logical laws of an impossible world, not because these laws are imposed by some logical system used to describe truth at impossible worlds, but because we can impose deductive systems on these worlds in such a way as to account for their facts. It is in this sense that these worlds can be said to be lawful, and could therefore stand in similarity relations to other worlds, based on how closely their laws resemble ours and how closely the propositions of these worlds match the propositions true in the actual world.

There is one further complication to discuss: an impossible world could represent some impossibility about modal space. An impossible world could, for example, represent the claim that the structure of the space of possible worlds is best described by S3, even though (we can assume) it is actually described by S5; an impossible world could represent the claim that all worlds accessible to it are worlds where some proposition P is true, even though P is false in some accessible worlds; an impossible world could represent the claim that, if A were true, then C would be true, even if the nearest A-worlds are not C-worlds. But this leads to an immediate worry: earlier, I had claimed that a structured proposition is true at a world iff it is a member of the set that comprises that world, but now it seems that an impossible world can "misrepresent" the modal truths of that world.

One kind of response to this worry is to concede that the modal truths of these worlds are problematic in just the way they claim to be: some counterfactual can be true at

an impossible world even if its consequent is not true in the nearest worlds where its antecedent is true. Any attempt at giving an analysis of counterfactuals (or claims about necessity and possibility) assumes that the world of evaluation of these claims is a possible world. Thus, questions about which counterfactuals would be true if some impossibility were to obtain, or questions about what would be necessary or possible from the standpoint of an impossible world, result in answers that are at odds with how we ordinarily understand our modal vocabulary. The space of these worlds is too anarchic for us to make good sense of.

A more optimistic response is that we *can* make sense of such modal claims in cases where the world of evaluation is an impossible world. Unlike claims about matters of fact at these worlds, however, the truth of these claims might not depend on whether the appropriate structured proposition is contained in the set that comprises the world, but rather by the shape of modal space around that world. Claims about whether a counterfactual is true or not might depend on what the closest worlds where the antecedent is true look like; claims about whether some proposition is necessary or possible might depend on what is true at the set of worlds that are accessible to the world of evaluation. There are, at least with respect to these modal facts, two different notions of truth, and whether the relevant notion of truth is that of truth *qua* set membership or truth *qua* features of the worlds around it depends on our interests.[64] If, for example, we are interested in representing the beliefs of an agent that is less than ideally rational, it is important to determine which structured propositions are true in the set of structured world that the agent takes to be actual; if we are interested in representing what would actually follow from the truth of how this agent takes things to be, it is important to determine which impossible

[64] Assuming that we can determine which worlds are close to an impossible world, of course; one such proposal is offered in chapter 3.

worlds are really close to (or accessible from) the worlds that the agent thinks might be actual.

4.3 Truth and Falsity in Impossible Worlds

4.3.1 Lewis's Puzzle

In footnote 3 of his [1986], Lewis argues that impossible worlds, if they existed, would spread contradictions to the actual world.[65] Consider a world *w* where both P and not-P are true. It seems that both 'In *w*, P' and 'In *w*, not-P' are true. Lewis claims that 'In *w*, not-P' entails 'Not: in *w*, P'. The result will be that both 'In *w*, P' and 'Not: in *w*, P.' are true. Thus, the contradiction is not merely contained in *w*, but spreads to the actual world as long as we are able to use an 'in *w*' operator to discuss what is true at *w*. In other words, we contradict ourselves when we describe the goings on of impossible worlds.[66]

The solution to Lewis' problem, as he suggests and as is discussed in more detail in Stalnaker [2003] and Lycan [1994], is simply to accept a less robust metaphysics of impossible worlds that will block the inference from 'In *w*, not-P' to 'Not: in *w*, P.' For Lewis, the inference is generally valid because possible worlds are concrete entities, so an operator such as 'in *w*' is a restrictive modifier. When we say that something is true at some world *w*, what we say is very much like what we would say if something were true in Massachusetts. If it is true that, in Massachusetts, there are no Republicans, then it is not true that, in Massachusetts, there are any Republicans. But impossible worlds need not be like Massachusetts, and operators that let us refer to the truths of impossible worlds need not function as restrictive modifiers. Indeed, once we grant that impossible worlds are

[65] Of course, putting the point this way makes it seem like Lewis is presenting some non-trivial counterpossible claim.
[66] Of course, not *everyone* would object to the threat of contradictions at the actual world. Dialetheists such as Priest and Beall would not object to this result, nor might Yagisawa.

neither complete nor consistent, we find very little reason to hold that the 'in *w*' operator functions in such a way as to validate the inference in question.

The notion of impossible worlds here, for example, is one according to which a world is a set of structured propositions. It simply does not follow from the claim that not-P is in some set to the claim that it is not the case that P is in that set. Indeed, the inference clearly fails when both P and not-P are in that set together. If a world fails to be consistent, then we should expect failures of the inference from 'in *w*, not-P' to 'not: in *w*, P'; if a world fails to be complete, then we can expect failures of the inference from 'not: in *w*, P' to 'in *w*, not-P.' Thus, the present conception of impossible worlds simply does not face Lewis's threat that contradictions at impossible worlds will spread to the actual world.

4.3.2 Dialethic Contamination

There is, however, another kind of contradiction contamination problem that this account of impossible worlds may face. Note that we have only explicitly discussed what is *true* at an impossible world, and not what is *false*. Presumably, some impossible worlds contain propositions that are both true and false: this is necessary in order to represent the beliefs of a dialetheist, or to represent the antecedents of counterpossibles that discuss dialethic logic. Consider, then, an impossible world w_1, at which some proposition P is both true and false. What can we say about the truth of the claim, 'in w_1, P'? Presumably, this claim is true: after all, P is true at w_1. But, presumably, this claim is also false, since P is false at w_1. Thus, it would seem that the claim 'in w_1, P' is both true and false *at the actual world*, and thus the contradiction in w_1 has spread to the actual world. This puzzle forces us to say more about what it is for something to be false in a world.

Assuming that a proposition can be both true and false in a world, it cannot be the

case that a proposition is false in a world iff it fails to be a member of that world. After all, set membership *is* classical: either a structured proposition is a member of a world or it is not. Thus, if a proposition were false at a world whenever it failed to be a member of that world, no proposition could be both true and false at some world. It would make more sense, given this system, to say that a proposition is false at a world when its negation is true. So, to say that some proposition is false in a world is to say that its negation is a member of that world.

Of course, this still does not give us an analysis of the 'in *w*' operator. Inspired by what has been said above, we might hold that, for some proposition P and world w_1, it is false that 'in w_1, P' whenever ~P is true at w_1. More generally, 'in *w*, ϕ' is false iff 'in *w*, ~ϕ' is true. This allows a proposition to be both true and false at some impossible world, since a proposition and its negation could both be members of that world. Yet this operator lands us in trouble, since the claim 'in w_1, P' is both true and false; this is true because P is a member of w_1, and false because ~P is a member of w_1. Thus, the contradiction within w_1 would spread to the actual world.

The next option, then, is to hold that some proposition is false in a world iff it is not true at that world. That is, 'in *w*, ϕ' is false iff not: (in *w,* ϕ). The negation in the metalanguage is ordinary, classical negation. On this reading of the falsity conditions of the 'in *w*' operator, a dialethia in some impossible world will not create a dialethia in the actual world. Some proposition P can be both true and false in a world w_1, in virtue of that world containing both P and ~P. But that dialethia will not spread to the actual world, because 'in w_1, P' is true (since P is a member of w_1), but not false (since it is not true that it's not the case that, in w_1, P is true.) The contradiction, then, would not spread to the actual world,

since a statement made using that operator will be false whenever it is not true.[67]

Note that it would also have been possible to define the 'in *w*' operator by first describing its falsity conditions (in terms of set membership), and the describing its truth conditions (in terms of the negation of its falsity). These operators, of course, are not equivalent: if P is both true and false in w_1, then 'in w_1, P' would turn out to be only false if the operator were described in this second way. What are we to make of these different ways of defining the 'in *w*' operator?

One could argue that it is better to define the operator in terms of truth, since truth is metaphysically or conceptually prior to falsehood. That is, even though truth and falsity are duals of one another, truth is somehow privileged in the structure of the reality, and thus when we consider what the facts of worlds are, it is appropriate to make use of a logical operator that conceptually privileges the truths of a world over its falsehoods.

If one is not comfortable with this metaphysics, one could also keep both operators. A locution such as 'in *w*' is ambiguous between describing a world (roughly) in terms of its truth and in terms of its falsehoods. Some contexts might favor one reading over the other; in other contexts, it would not matter which operator we use, and so there would be harmless semantic indeterminacy.

Regardless of how the details are filled in, it should be clear that the defender of impossible worlds has nothing to fear from other-worldly contradictions. The problem raised by Lewis is not an issue for this account of impossible worlds, since they would not license the kind of inference from 'in *w*, ~P' to '~in *w*, P' that Lewis discusses. The modified problem, restated in terms of truth and falsity, would not be a problem either, since there is a

[67] This analysis also makes sense of truth value gaps. Consider now a world *w'*, where P is neither true nor false. This is a world of truth value gaps, rather than truth value gluts. What are we to make of a statement such as 'in *w'*, P'? It is not true, since P is not a member of *w'*. Since the 'in *w'* operator's falsity conditions are classical, it will be only false that 'in *w'*, P.'

reasonable, classical reading of these operators that does not result in any contradiction at the actual world.

4.4 Conclusion

In this chapter, I have argued for a particular conception of impossible worlds. We should accept impossible worlds so construed because they are ontologically cheap and theoretically useful. We can use these worlds to describe ways the world could not be. Moreover, we can use these worlds to make better sense of epistemic possibility, the analysis of counterpossibles, the content of belief, and the analysis of properties and propositions, and we can use these worlds to distinguish properties and propositions that would otherwise be conflated.

This proposal is also capable of answering objections related to the "spread" of contradictions from impossible worlds into the actual world. Lewis's argument against impossible worlds will not work because these worlds do not allow one to reason from 'in some world, not-P' to 'not: in some world, P.' Furthermore, the version of the argument that makes use of what is true and false at impossible worlds can be met if one is careful to introduce operators that allow one to describe what is true and false at these worlds.

Thus, structured impossible worlds are both adequate to the task of serving as the impossible worlds needed in our theory of counterpossibles, and are immune to several kinds of objections one might level against impossible worlds. They are robust enough to represent the kinds of content they would need to represent; they are capable of standing in similarity relations; they do not "spread" contradictions to the actual world; and they are easily added to our ontology because they are merely set theoretic constructions of possible

worlds. In short, this kind of impossible world should be accepted because it is serviceable, cheap, and logically unproblematic.

CHAPTER 5

DEPENDENCE AND COUNTERPOSSIBLES

5.1 Introduction

Discussion of grounding or metaphysical dependence is common in philosophical discourse, and one often invokes the notion of grounding or dependence in the course of giving an analysis of some class of objects, properties, or phenomena. Moral properties might be said to be grounded in or depend on physical properties; the existence of some object might depend on the existence of its parts (or *vice versa*); mental properties might be grounded in physical properties; semantic facts might depend on facts about usage; the beliefs of some agent might depend on her credences, and so on. Other cases are somewhat more subtle. Sometimes, a philosopher will argue that we should countenance some entity in our ontology, at least in part because it is 'nothing over and above' things we already believe in. In other cases, a philosopher will consider herself entitled to accept a certain collection of facts because they are true 'in virtue of' some other collection of facts. In all of these cases, there is either an implicit or explicit appeal to the grounding relation.[68] Having a better understanding of the grounding relation would give us a better understanding of what is at stake when we reduce some phenomenon to some other phenomenon, believe that

[68] One might worry, of course, that there is more than one notion that corresponds to our locutions of dependence, and thus hold that talk of "the grounding relation" is misleading. Perhaps there is no single notion that corresponds to such a relation (see, for example, Hofweber [2009]), or perhaps there is a family of such relations (see, for example, Bennett [forthcoming] on building relations). Nonetheless, I will optimistically assume that there is a notion of dependence to which we can give a unified analysis. If attempts to give such an analysis falter, or seem wildly disjunctive, then we could treat that as evidence that there is more than one dependence relation after all.

some objects depend on some other objects, or accept some truths as holding 'in virtue of' other truths.

The grounding relation itself might also play a role in the analysis of other concepts in philosophy. For example, we might hold that grounding plays a role in a developed theory of truthmaking (see Schaffer [2010]), if we hold that truthmakers are the fundamental entities that ground the truth of propositions. Rosen [2010] and Trogdon [2010] describe intrinsicality in terms of grounding. Nor is discussion of grounding new to philosophy; the Euthyphro dialogue, for example, seems to involve a discussion of whether piety is grounded in the attitudes of the gods, or whether the attitudes of the gods are grounded in piety.

Most accounts of grounding or dependence attempt to explicate it in modal terms, as in supervenience accounts, or merely take it as primitive. Yet there are well-known problems with the supervenience account of grounding, and an attempt to take such a crucial notion as primitive might strike some as disappointingly unilluminating. The approach favored in this chapter is a hybrid account that analyzes the notion of dependence in counterfactual terms as well by explanation. A depends on B iff a) if exactly one of either A or B were true, then it would be B that is true, and b) A is explained by B. Developing and defending this view, however, will require a discussion of the counterfactual conditional stated above, as well as the role to be played by explanation.

Some preliminaries: I will use locutions such as 'dependence,' 'metaphysical dependence,' and 'grounding' interchangeably. I also assume that this is the same relation that others refer to with locutions such as 'is true in virtue of' or in discussions about metaphysical or ontological priority. Thus, at least for present purposes, 'A depends on B,'

'A is grounded in B,' and 'A is true in virtue of B,' and 'B is prior to A' all mean the same thing.

I take the grounding relation to hold between propositions. When we say that A depends on B (or A is grounded in B), we should construe A and B as standing for propositions. Some authors with stricter ontological scruples might balk at this. To be sure, nothing here hangs on any particular conception of propositions, as long as they are fine-grained enough to represent the content of necessary truths and falsehoods; we could easily substitute equivalence classes of sentences for propositions. For some purposes, one might prefer a grounding relation that takes the relata to be objects or properties, so as to make sense of claims such as, 'The whole is grounded in its parts' or 'mental properties depend on physical properties,' respectively. Such claims can be construed as propositions about the existence of these objects or instantiation of these properties, such as *that the whole exists* is grounded in *that its parts exist* or *that mental properties M are instantiated* depends on *that physical properties P are instantiated*.[69]

5.2 Some Theories of Dependence

5.2.1 Against Supervenience

[69] Claims about the dependence of one object on another might not always reduce to claims about the existence of one object being dependent on the existence of another. Perhaps something stronger is being claimed: when we say that a house depends on its bricks, we are not merely saying that the existence of the house depends on the existence of its bricks. We might also be claiming that the house depends on the bricks and *their particular arrangement*; we might also be claiming that any change in the house's intrinsic properties must correspond to some change in the properties of the bricks. Of course, none of this is to say that a reduction of object-dependence to property-dependence is impossible. Rather, it seems that determining the exact properties that stand in the dependence relation might depend on what one means by a particular claim of object- or property-dependence.

Claims about dependence were once analyzed as claims about supervenience. One might claim that we can understand metaphysical dependence in terms of patterns of existences, or the instantiations of properties, or the truth of certain propositions across worlds. Different supervenience claims will correspond to different grounding relations, to be sure, but assuming that the relata of the dependence relation are propositions, there is a very straightforward account of dependence by way of supervenience:

> (Dependence as Supervenience) A depends on B iff no two possible worlds differ with respect to the truth of A but agree with respect to the truth of B.

On such an account, metaphysical dependence is the necessary co-variation of truth. This is in line with a project of attempting to secure a theory of reduction without adopting any kind of heavy-weight or spooky notion of ontological dependence; there is no dependence relation apart from what is true about the various entities across worlds.

As pointed out by Schaffer [2009] and others, the supervenience relation has the wrong formal features to capture the notion of dependence in question: supervenience is reflexive and non-asymmetric, while metaphysical dependence is not. Supervenience is reflexive because everything supervenes on itself: there can be no change in something without a change in that very thing; supervenience is non-asymmetric because when A and B co-vary necessarily, A supervenes on B and B supervenes on A. Dependence, on the other hand, is irreflexive and asymmetric, since nothing depends on itself, and if A grounds B, then B cannot ground A.[70]

There might be an easy fix for these formal problems, however. The spirit of a supervenience account, after all, is merely that we explain dependence in terms of what is

[70] It is possible that in some cases, something can ground itself. But we presumably do not want this to always be the case, nor should it be the case that something is the *sole* ground of itself.

true at various possible worlds. If the simple supervenience account fails, then perhaps a small modification is in order:

> (Dependence as Asymmetrical Supervenience) A depends on B iff a) A supervenes on B (as described above), and b) B does not supervene on A.

This account neatly solves the worries of reflexivity and symmetry. On this account, nothing depends on itself, since the second clause of the analysis cannot be satisfied when A=B. Similarly, the relation is now asymmetric, since there must be some difference between the A-worlds and B-worlds for A to depend on B.

Even our revised account of dependence by supervenience, however, is inadequate as an analysis of metaphysical dependence. The grounding relation is hyperintensional in a way that supervenience is not. Even though all necessary truths trivially supervene on my existence, it should not be the case that (say) the truth of 2 + 2 =4 depends on my existence. Furthermore, the null set grounds other pure sets, {Socrates} is grounded in Socrates, and so on. Yet, since each of these pairs exists in the same possible worlds, (Dependence as Supervenience) would hold that they depend on each other, while (Dependence as Asymmetrical Supervenience) would hold that neither one depends on the other.

5.2.2 Against Naturalness

According to Bricker [2006], A metaphysically depends on B if A supervenes on B, and B is perfectly natural. Note that Bricker offers this as merely the basis of an analysis of grounding. First, this is merely a sufficient condition for metaphysical dependence: the analysis requires that the more basic proposition be perfectly natural, yet this will not be the case in many interesting examples of metaphysical dependence. Second, this account requires a notion of naturalness that is appropriate in the discussion of propositions.

Naturalness, after all, is generally thought of as applying to properties and relations. A derived notion of naturalness that can apply to propositions must be deployed if this notion is to be of use in a discussion of propositions depending on other propositions.[71]

These problems are not insurmountable. At least at a first pass, it would be quite easy to generalize Bricker's account so as to make sense of a great many more of the kinds of examples we are interested in: A depends on B iff A supervenes on B and B is more natural than A. A proposition is perfectly natural if it attributes some perfectly natural property to some perfectly natural object; one proposition is then more natural than another if it contains fewer conjunctions or disjunctions of perfectly natural propositions, as in Lewis [1983]. As long as one can make sense of this notion of naturalness, then this analysis of dependence seems tempting. If we are inclined to think that mental facts depends on physical facts, it is because the mental facts supervene on the physical facts, and the physical facts are more natural (in Lewis's sense) than the mental facts.

What can be said about the Socrates/{Socrates} dilemma? Presumably, the defender of the naturalness approach would want to say that the existence of {Socrates} depends on the existence of Socrates, and not the other way around. This is because, in the first place, the existence of {Socrates} supervenes on the existence of Socrates: no two worlds differ with respect to the existence of {Socrates} and Socrates. Furthermore, the proposition that Socrates exists is more natural than the proposition that {Socrates} exists.

But what licenses the naturalness theorist to claim that Socrates is more natural than {Socrates}? It cannot be, of course, that {Socrates} depends on Socrates; saying this would blatantly court circularity. Nor does it seem that the suggestion of counting the number of

[71] Similarly, we would need some derived notion of fundamental objects if we are interested in how objects depends on other objects.

conjuncts or disjuncts of perfectly natural properties is appealing. It is, after all, entirely unclear how to build up the property of being {Socrates} out of the perfectly natural properties. It is even more difficult to see how to do this if the perfectly natural properties are the properties used by fundamental physics.[72]

It would be better to describe naturalness in terms of metaphysical dependence, rather than the other way around. After all, the notion of naturalness in play is itself rather mysterious. Indeed, the notion was mysterious enough when applied to properties and relations; it is only more mysterious when applied to propositions. Once we have an analysis of the notion of grounding, we can use it to determine which properties are more or less natural; the more natural properties are those that *ground* the less natural properties. One may well reply that the notion of metaphysical dependence is at least as mysterious as the notion of naturalness, and so we gain nothing by complaining about the mysteriousness of the notion of naturalness employed by Bricker. This charge, however, is premature: if we are unable to come up with a satisfactory account of dependence in terms that are more tractable than naturalness, then this reply would have force. But there are still more tools that we might use in our analysis, and these tools might yet give us a better understanding of how grounding works than we would have if we analyzed grounding in terms of naturalness.

Perhaps one could argue that naturalness is of use in other domains, and so we should accept naturalness for the sake of its usefulness: a notion of metaphysical naturalness

---

[72] One could claim that Socrates is more fundamental than {Socrates} because, however many conjunctions and disjunctions of natural properties are required to make sense of the proposition *that Socrates exists*, one more – namely set membership – is required to make sense of the proposition *that* {*Socrates*} exists. But this does not seem like the proper way of determining the degree of naturalness of impure sets. First, it is a matter of some dispute whether {Socrates} has many of the kinds of properties that Socrates has, such as mass or spatial location. Even if the set did inherit Socrates's physical properties, this approach to naturalness would have the result that {Socrates} is as fundamental as Socrates-plus-some-electron, or that {Socrates} is more fundamental than the singleton of any "simpler" entity.

is useful with respect to giving an answer to Kripkenstein- or Goldman-related problems of induction, for example, or in an account of duplication and similarity. Yet rather than being an argument for accepting the notion of naturalness as is, this might rather be construed as a more urgent demand to explain naturalness by way of other, better-understood notions. If there is an account of metaphysical dependence, and that account can be used to explain naturalness, then we can freely appeal to the notion of naturalness to solve these puzzles.

5.2.3 Against Explanation

One might also attempt to analyze the dependence relation by way of explanation. According to the simplest version of such a view, A depends on B iff B explains A. The relation of metaphysical dependence, then, is analyzed solely in terms of the relation of explanation. Such an account holds some promise if we are to overcome the difficulties of the supervenience approach. It is at least plausible that explanation is hyperintensional in a way that supervenience is not; for example, the existence of Socrates has a different explanation than the existence of {Socrates}, even if the two entities exist in the same set of possible worlds. Furthermore, intuitions about explanation seem to capture many of the fundamental features of the relation of metaphysical dependence: if the truth of some proposition is *explained* by the truth of some other proposition, then it seems that the *explananans* is more fundamental than the *explanandum*, and perhaps the *explanandum* is true in virtue of the *explanans*.

Of course, a complete defense of such a view would require saying much more about what it is for the truth of one proposition (or collection of propositions) to explain the truth of another. Doing so would be outside the scope of this paper, but a few comments are in order. The notion of explanation in question cannot be pragmatic, such as that of Van

Fraassen [1977], lest the account of metaphysical dependence be similarly pragmatic. Nor can explanation be strictly causal, as in Lewis [1986b] or Scheffler [1957], as our account of explanation must be general enough to make sense of the dependence facts of propositions that lack causal histories, such as logical and mathematical propositions.[73] This is not to say that many instances of explanation are not causal in nature, or that causation can play no role in the analysis. Nonetheless, if our account of explanation requires that causation play some central role, and that logical, mathematical, and metaphysical truths do not stand in causal relations, then it is not clear how these necessary propositions could stand in explanatory relations, and thus not clear how they could stand in relations of metaphysical dependence, either.

There are, however, accounts of explanation that could meet these challenges. One account of explanation that can answer these kinds of challenges is a unificationist approach, such as that of Friedman [1975] and Kitcher [1989]. According to such an account of explanation, one provides an explanation of some phenomenon by showing how it fits into a unified theory of the entire world, where a unified theory is one that contains relatively few primitives or argument types. To be sure, our present account of dependence does not require a unificationist account of explanation; it merely requires that there be some theory of explanation that is appropriately hyperintensional and non-pragmatic.

How could the account of dependence as explanation function in practice? For some kinds of cases, the explanation account of dependence seems to be on the right track: a collection of facts about some microphysical system explains the facts of the corresponding

[73] Indeed, if causation were sufficient for dependence, then there would be many spurious instances of dependence. For example, a dualist who holds that physical events can cause mental events could wrongly be said to hold that those mental events thereby depend on physical events.

macrophysical system, and thus the microphysical system grounds the macrophysical system. Depending on how our theory of explanation is ultimately formulated, we could say that the world as described in terms of microphysics uses a smaller vocabulary of entities and their properties and relations. In the case of Socrates and his singleton, this proposal is committed to the claim that the existence of Socrates explains the existence of {Socrates}, and that the existence of {Socrates} does not explain the existence of Socrates. Intuitions in this case seem somewhat less straightforward, to be sure, but one certainly could argue that the existence of an individual explains the existence of a set that contains that individual, and not *vice versa*. It is more economical to describe the world in terms of individuals and their relations rather than it is to describe it in terms of sets and their features. This is easiest to see if we assume that impure sets are not located in space and time and do not have the physical properties of their concrete elements such as mass: it would not be clear, for example, how to recover physical features of Socrates such as his location and mass and lineage from truths about {Socrates}, while determining the features of {Socrates} requires no tools that are not provided by the principles of set theory. Thus, we could describe the world in terms of objects such as Socrates and derive truths about the sets that contain them, but we cannot describe the world in terms of objects such as {Socrates} and then derive truths about their members. What if impure sets have the physical properties of their members? If we allow that sets such as {Socrates} could have the physical properties of their members and stand in causal relations, then we lose the kind of asymmetry mentioned above. But there is a different kind of asymmetry: Socrates has the property of being a member of {Socrates}, while {Socrates} has the property of containing Socrates as an element. The question of which kind of property makes for a more economical description of the world would then come down to a debate over whether describing the world in set-

theoretic facts (and then deriving the physical facts) is easier (*e.g.,* makes use of fewer primitives) than describing the world in physical facts (and then deriving the set-theoretic facts). I assume a description of the world in set-theoretic terms will face difficulties with respect to various modal and other properties of sets, but if the world could ultimately be described just as easily in set-theoretic terms, then I would accept this as evidence that Socrates might not ground {Socrates} after all.

Other kinds of cases are more problematic for a reduction of dependence to explanation. One paradigmatic kind of explanation is that of explaining the present state of the universe by citing a previous state and the laws: some state of the universe at t_0, in conjunction with the laws of physics, explains a latter state of the universe at t_1. Yet it does not seem right to say that some previous state of the universe (along with the laws) *grounds* a latter state of the universe; there is a causal and explanatory connection between the past and the present, but not a connection of metaphysical grounding. Thus, it seems that even if the notions of explanation and dependence are related, they are not equivalent; explanation is not a sufficient condition for dependence.

Of course, the advocate of a view of dependence as explanation has a number of ways of revising the theory to accommodate this challenge. One could attempt to combine the explanation requirement with a modal requirement, or one could attempt to focus on a particular kind of explanation that is not present in the past-present case. I will discuss both of these options in turn.

If explanation is not sufficient for metaphysical dependence, then perhaps the account should be supplemented with one of the modal notions discussed above. The most straightforward strategy would be to combine the requirements of explanation and supervenience, which would result in something close to Horgan's relation of

superdupervenience:[74] such an approach, however, would still wrongly hold that the past and the laws ground the present state of the universe. At least if determinism is true, the state of the world at some earlier time t_0 and the laws explain *and entail* some subsequent state of the world at time t_1, and thus the state of the world at t_1 also supervenes on the state at t_0 and the laws. This is because if some previous state (at t_0) and the laws entail the present state (at t_1), it then follows that the state of the world at t_1 could not be different unless either the laws or the state at t_0 were also different.[75]

It also will not help to analyze metaphysical dependence in terms of explanation and *asymmetrical* supervenience. Such an approach would still wrongly hold that the present metaphysically depends on the past and the laws, since some state of the world *asymmetrically* supervenes on the laws and a past state of the world. This is because one cannot alter the present without altering the laws or the past, but the opposite is not true: state of the world at t_1 could be derived from more than one set of initial conditions and dynamic laws.[76] Furthermore, recall that other paradigmatic cases of grounding do not meet the asymmetrical supervenience requirement: the existence of {Socrates} supervenes on the existence

---

[74] Horgan discusses this relation is his [1993], and defines it as "ontological supervenience that is robustly explainable in a materialistically explainable way," where to give a robust explanation of supervenience "is to explain it qua ontological, rather than explaining it merely as a feature of the "logic" of the higher-order terms and concepts." Such an account is meant to make sense of physicalist-friendly reductions (or pseudo-reductions) of higher-order properties to lower-order properties. If we are to approach metaphysical dependence in full generality without commitment to physicalism, we need not hold that the explanations in question need to be "materialistically explainable."

[75] What if determinism is false? If the laws are merely stochastic, then our theory of explanation would also have be different so as to make sense some previous state of the world and laws making some latter state highly probable, rather than being guaranteed. The supervenience claim would then return, restated as a claim about probability: there can be no change in the objective chance of the current state obtaining without some change in the previous state of the world or the laws.

[76] This is not to be confused with the claim that the past supervenes on the present state and the laws, which may be true depending on how the physical laws are formulated.

Socrates, and the existence of Socrates supervenes on the existence of {Socrates}, and thus, according to such a proposal, Socrates will not ground {Socrates}.

Perhaps the dependence relation cannot be analyzed as explanation *per se*, but rather as some non-causal kind of explanation. Assume that there is some sense to be made of a specifically non-causal notion of explanation: could this kind of explanation (perhaps in concert with a further modal requirement) be capable of analyzing the relation of metaphysical dependence? Such a theory would not hold that the present depends on the past and the physical laws, since such an explanation would certainly be causal. Yet there are also cases of non-causal explanation that do not seem to be cases of metaphysical dependence.

One might attempt to explain why, in some population, for each male born, there are 1.04 females born. One could attempt to explain this causally, by citing facts about sperm and egg cells, but one could also offer a non-causal explanation that cites evolutionary factors: it is advantageous to the population to have a 1:1 ratio of males to females at reproductive age, and, for various reasons, males are somewhat more likely than females to die before reproducing. This explanation is not causal: rather than citing earlier facts about features of the population and relevant dynamic laws, we cite certain facts about features of the population in the *future* to explain the sex ratio. Yet perhaps one could argue that this evolutionary kind of explanation is simply not a *real* explanation, as the evolutionary explanation appealed to is plausibly a stand-in for a far more complex explanation about *past* populations and their genetic features and reproductive successes.[77]

[77] This example is discussed in Kitcher [1989], and a closely-related example is discussed in Sober [1983].

A clearer example is Hempel's pendulum [1965]: consider that the period of a pendulum is explained by the length of its rod, as well as the formula $t = 2\pi\sqrt{l/G}$. This explanation is not causal: the length of the rod is not a *cause* of its period having a certain value; the pendulum's period is simply a function of the length of the rod and gravity. In this case, there seems to be a non-causal explanation of the period of pendulum from its length and from gravity; there is also (asymmetrical) supervenience of the period of the pendulum on the length of the rod and the force of gravity, but we should not say that the length of the rod, the force of gravity, and the equation above *ground* the pendulum's period. The length of the rod is not metaphysically prior to the period of the pendulum, nor is the period of the pendulum prior to the length of the rod.

We could also consider another kind of example. One might try to explain why a star has ceased collapsing by citing the Pauli Exclusion principle: given constraints on the properties that collections of subatomic particles can have, there is simply no possible state in which the star collapses further. Such an explanation cites certain nomic laws, but it does not refer to any causal or dynamic law; there is no claim about how the system develops given some previous state, but rather a claim that a certain range of states of the system are simply not possible.[78] Yet we would not want to say that the failure of the star to collapse further *metaphysically depends* on the Pauli exclusion principle. Once again, there can be a non-causal explanation, (asymmetrical) supervenience, but not metaphysical grounding.[79]

[78] Note that Lewis [1986b] holds that this is not an example of genuine explanation; rather, stating that the Pauli exclusion principle holds that the star cannot collapse further merely denies that there is any possible explanation for why the star cannot continue to collapse.
[79] Note that the pendulum example and the collapsing star example are also cases of *synchronic* explanation: we need not appeal to some system at different points in time to give an explanation of the *explanandum*. Thus, it does not seem possible to gloss 'non-causal explanation' as 'synchronic explanation.'

This notion of non-causal explanation is not the only option on the table for the defender of the explanation view of grounding. Perhaps the kind of explanation in question should be a version of what Cummins refers to as a *property theory* or what Poland refers to as a *realization theory*. For Cummins, a property theory is best viewed in contrast to a transition theory: a transition theory explains why a system is in the state that it is in, given some previous state; a property theory, however, explains "What it is for S to instantiate P, or in virtue of what does S have P" [Cummins 1985: 15]. For Poland, a realization theory is something that, for some higher-level attributes N, "abstractly characterizes the kinds of attributes that are sufficient for the realization of N and that shows how such attributes can combine to actually constitute N in particular cases." [Poland 1994: 210]. Note that both of these views of explanation are very difficult to state without slipping into locutions such as 'in virtue of' or 'realizes,' and 'constitute,' which are themselves the very target of this discussion. This is not to say that there is no way of making sense of a notion of explanation that is capable of making sense of the dependence relation, but it does not seem that any of the plausible contenders are up to the task.

5.2.4 Against Primitivism

One tempting response to worries about the nature of the dependence relation is simply to take it as primitive. Rosen [2010] and Schaffer [2009] hold that the notion of metaphysical dependence is clear enough to use in without an analysis. After all, we have a reasonably intuitive grasp of what is meant by the kinds of locutions of dependence mentioned above, even if they are difficult to explain in a perspicuous fashion. We agree on certain paradigm cases, such as that the existence of {Socrates} depends on the existence of

Socrates. If we have availed ourselves of the notion of dependence for thousands of years, it would seem odd to now hold that the notion is too mysterious on its own.

While I am rather sympathetic with these remarks, it simply does not follow that we should abandon our pursuit of an analysis of the grounding relation. After all, we have reasonably robust intuitions of what knowledge is, and there is broad agreement about the paradigm cases of knowing and not knowing. Yet this fact does not stop philosophers from attempting to better understand knowledge in other terms (*pace* Williamson). Epistemologists are still in business, and still providing accounts that seek to provide plausible and informative theories of knowledge that could then be used to better understand the concept and settle other debates in and around epistemology.

Furthermore, we might hope that an analysis of the dependence relation in other terms might illuminate why it functions in the way it does. Perhaps a successful analysis would give us some guidance in understanding more contentious examples of grounding; a good account of grounding might then be useful in determining whether, for example, the existence of a heart depends on a body, or if the existence of a body depends on the heart (*inter alia*). More interestingly, an account of grounding could then be used to help settle a debate such as that between the priority monism and pluralism, or between the reductive and non-reductive materialist. If this is too much to ask of our theory, we might also have a more modest task for our analysis of dependence. If the analysis determines why the dependence relation seems vague or contentious in certain kinds of cases, then that too would be an interesting result. Even if the proper analysis of dependence cannot resolve long-standing debates, it could still explain why these debates seem so difficult, and point toward the deeper sources of disagreement that spur these arguments.

To be sure, if our best attempts at analyzing the dependence relation fail, we could adopt primitivism about grounding as a kind of fallback position. It would be misguided to stop providing philosophical analyses of phenomena in terms of other phenomena just because we cannot explain the grounding relation. Nonetheless, if some account of grounding does seem plausible, and provides guidance in more problematic cases, then such an account would be a welcome improvement over primitivism about grounding.

5.2.5 A Simple Counterfactual Approach

Schaffer, in his [2009], suggests (though does not endorse) a counterfactual strategy for dependence. According to such a strategy, metaphysical dependence is a species of counterfactual dependence:

> (Counterfactual) A depends on B iff, if B were not true, then A would not be true, but if A were not true, then B might still be true.

This proposal could either be construed as an independent account of dependence, or as an additional requirement to shore up the supervenience account. In either case, such an account has some clear advantages: it provides the asymmetry required of the metaphysical dependence relation and it casts metaphysical dependence in a similar light to that of counterfactual dependence. Perhaps the dependence relation is similar to, or even a generalization of, causation, and perhaps (something like) a counterfactual analysis is appropriate for causation. Thus, a kind of counterfactual analysis should be on offer for metaphysical dependence.

Of course, Schaffer does not endorse anything like this counterfactual account of dependence. For one thing, such an account would clearly require the use of counterpossibles, or counterfactuals with necessarily false antecedents: the antecedents of

both the 'might' and 'would' conditionals will often be metaphysical impossibilities. For example, it is quite plausible that {∅} depends on the empty set. It is, however, metaphysically impossible that either {∅} or the empty set does not exist, so both counterfactuals have necessarily false antecedents. Schaffer assumes that such counterfactuals are either vacuous (as Lewis, Stalnaker, and others do), or otherwise unsuitable for use in a theory of dependence. By this point, however, such worries might not move us deeply. I have argued at length that a non-vacuous account of counterpossibles is necessary, and even if the particular account offered here fails, it seems that some non-trivial account is needed to make sense of our linguistic and philosophical practice.

Setting aside worries related to the proper analysis of counterpossibles, the question of whether the simple counterfactual analysis is adequate remains. Consider again the case of Socrates and his singleton. Since they exist in the same possible worlds, the existence of either one supervenes on the existence of the other. But it does not seem that the counterfactual account suggested here will provide the correct result, either.

Consider the first part of Counterfactual: If Socrates were to not exist, then {Socrates} would not exist. This is presumably true, since it seems plausible that if Socrates did not exist, then his singleton would not exist, either. The nearest worlds where Socrates does not exist are presumably worlds where philosophy develops differently, worlds where the history of Greece and Greek culture diverge from ours, and so on; they are not worlds where impure sets get formed differently. Since Socrates does not exist, the set composed of Socrates does not exist, either, as there is no individual to form that set. So far, so good.

Consider now the second part of Counterfactual: If {Socrates} were to not exist, then Socrates might still exist. Is this 'might' counterfactual true? It does not seem so. If {Socrates} were to fail to exist, then Socrates would not exist, either. All of the nearest

worlds without singleton Socrates are worlds without Socrates around to form the set; again, they are worlds with different historical or cultural truths, but not worlds with different mathematical truths. Worlds without the singleton are worlds without the philosopher to create the set, rather than worlds where set theory somehow does not allow one to create a set out of some individual.[80]

Something like Nolan's doctrine of the Strangeness of Impossibility[81] could be driving these intuitions: no impossible world is closer than any possible world. Since there is a sphere of possible worlds where {Socrates} does not exist – the worlds that lacks both Socrates and {Socrates} – we need not consider any of the impossible worlds that contain Socrates but not {Socrates} when evaluating these conditionals. The Strangeness of Impossibility doctrine is quite plausible, but nothing so strong is required here. The only thing to note here is our firm intuitions about the truth values of the counterfactuals being employed: it simply does not seem that, if {Socrates} were to fail to exist, Socrates himself might still exist.

Consider the similarity metric for counterfactuals discussed in chapter 3. According to such a view, a world is nearby to the extent that it matches the laws of the actual world and the facts of the actual world, with greater weight placed on matches of law. Locating the closest sphere of worlds where {Socrates} does not exist means locating worlds without {Socrates}, but that preserve our laws and facts as much as possible. There are, to be sure,

---

[80] There is, to be sure, some disagreement about how to understand the semantics of the 'might' counterfactual used here. For Lewis, a 'might' counterfactual is the dual of a 'would' counterfactual, so to say that, if {Socrates} were to fail to exist, then Socrates might still exist is to say, 'It is not the case that, if {Socrates} were to fail to exist, then it is not the case that Socrates would still exist.' For Stalnaker, a 'might' counterfactual is to be understood as an epistemic modal operating on a 'would' counterfactual, so we should read the sentence as (something like) 'For all I know, if {Socrates} were to fail to exist, then Socrates would still exist.' Both of these might claims seem false.

[81] See [Nolan, 1997]

impossible worlds that lack {Socrates} but contain Socrates, but these worlds would achieve a match in matters of fact (*viz*, facts about the philosopher) at the expense of altering set-theoretic laws about how impure sets can be formed. Thus, *all* the closest worlds that lack {Socrates} will also lack Socrates, and so the 'might' counterfactual is false.

5.3 Counterfactual Exclusion

5.3.1 The Counterfactual Exclusion Requirement

There is an important role for counterfactuals to play in the analysis of dependence, but as we have seen above, it cannot be the one that Schaffer suggests. The failure of the counterfactual approach to dependence seen above could be attributed to the fact that we are simply using the wrong counterfactuals. A counterfactual analysis can, after all, give us the tools required to meet the challenges faced by an account of dependence that relies on supervenience. Recall that metaphysical dependence cannot be analyzed by supervenience alone, since dependence is asymmetric and irreflexive, while supervenience can be symmetric and is reflexive. A counterfactual analysis of the metaphysical dependence relation will allow it to be hyperintensional in a way that supervenience is not. That is, if two propositions P and Q are true in the same possible worlds, then P and Q will supervene on each other, and thus (if the supervenience account of dependence is true) metaphysically depend on each other. But, as I have argued previously, counterfactuals are hyperintensional: counterfactuals with impossible antecedents have non-trivial truth conditions, for example. Thus, if two propositions P and Q are true in the same possible worlds, it does not follow that they would behave the same way as antecedents in counterfactuals.[82]

[82] There are other ways that counterfactuals are hyperintensional. Consider, for example, the pair, 'If I were the President, then I would live in the White House' and, 'If the President

The problem with the counterfactual account sketched above is that, at least in many cases of metaphysical dependence, the more- and less- fundamental entities exist at the same possible worlds. Thus, if either one were to fail to exist, then so too would the other. But it would be a mistake to conclude that this is a problem for every counterfactual account of dependence. Rather than making use of the kinds of counterfactuals mentioned above, the right theory of dependence must use counterfactuals to highlight some other modal property of the entities in question. The requirement I favor can be called *counterfactual exclusion*: rather than ask what would happen if the more fundamental proposition were to be false, we should ask what would happen if only one, but not the other proposition were to hold:

> (Counterfactual Exclusion) A depends on B iff, if exactly one of either A or B were true, then B would be true.

According to Counterfactual Exclusion, for one proposition to depend on another, we must determine what would happen if one of those propositions were true and the other false. In the language of the nearness of worlds, if A depends on B, then the nearest worlds where A is false and B is true are closer (according to some plausible metric of the similarity of worlds) than worlds where A is true and B is false. In such cases, we could say that A is more "modally fragile" than B. This modal fragility introduces the kind of asymmetry and hyperintensionality required for our analysis, and does so in a way that suggests the right kind of modal relationship between the grounded and grounding entities.[83]

---

were me, then I would live in the White House.' Even though the antecedents are logically equivalent, the truth values of these counterfactuals are different.

[83] Asserting that A is more modally fragile than B is not to assert that B *might* exist without A, at least on the most natural reading of 'might.' I do not doubt that there are contexts according to which the nearest sphere of worlds where A is false is one that contains some worlds where B is nonetheless true, but this is not to say that, in ordinary contexts, a claim such as 'If either A or B were to be false, then A would be false' entails 'If A were false, B might be true.' The former counterfactual forces us to evaluate the worlds at a sphere where

How would this account handle the case of the dependence of {Socrates} on Socrates? According to the counterfactual requirement, for {Socrates} to depend on Socrates, it should be that, if only one of the two were to exist, then it would be Socrates that would exist. This seems quite plausible. After all, {Socrates} seems more "modally fragile" that Socrates; it is easier to delete a set from a world than a human being; the world without Socrates would look far more different to us than the world without {Socrates}. Of course, this is no true defense of the claim, and one might hold that there is no real difference in "ease of deletion" or that it is mere bias to hold that differences in how a world "looks" physically are more important than differences in how it looks set theoretically. I will attempt to defend this judgment about similarity at below. Nonetheless, it is plausible that the impossible worlds that differ only with respect to set theory are smaller deviations from actuality than those that differ only with respect to the history of the physical world.

It is also clear that there is no threat of Socrates also depending of {Socrates} according to the counterfactual account of dependence. The counterfactual requirement is not met in this direction: if exactly one of either Socrates or {Socrates} were to exist, it would surely be Socrates, and not {Socrates} that would exist.

5.3.2 Similarity and Counterfactual Exclusion

---

either A is true and B is false, or A is false and B is true; the later counterfactual makes no such demand. To put the point in a more concrete manner, I am *not* committed to the claim that Socrates might have existed without {Socrates}. Claims about what is possible and impossible are understood in the ordinary way, by what is true at various possible worlds. There is no possible world where Socrates exists without {Socrates}. When one evaluates a conditional such as, 'If exactly one of Socrates or {Socrates} were to fail to exist, then it would be {Socrates} that would not exist,' one is forced to evaluate the conditional with a different system of spheres of worlds that contains impossible worlds; we are thus not committed to any corresponding claim about possibility.

This account of dependence faces an immediate challenge: we must be able to provide a reasonable account of the truth conditions of counterfactuals with necessarily false antecedents. Many of the conditionals expressed by the counterfactual exclusion requirement will be counterpossibles. In the case of Socrates and {Socrates}, for example, it is impossible that Socrates should exist without {Socrates} or *vice versa*, since (presumably) the laws of set theory are metaphysically necessary, and would entail that {Socrates} exists whenever Socrates exists.[84] Furthermore, such a theory of the truth conditions of counterpossibles should be at least at least as objective as the metaphysical dependence relation: counterfactuals, after all, are famously vague and context-sensitive, while dependence is not. Even if there are some cases where the facts about dependence seem unclear, it would be a mistake to think that we can make something more or less fundamental merely by re-describing it or raising a question of dependence in a different conversational context. It is, therefore, important that there be a "default" reading of the counterfactuals, and that, at least in most cases, we can use this default metric to analyze the truth of dependence claims.

Such an account was offered in chapter 3; the standard similarity metric for counterpossibles discussed is a descendent of Lewis's similarity metric for counterpossibles, from his [1979]. Lewis's standard metric is generalized so as to make sense of the kinds of laws and matters of fact that are relevant to the similarity of *impossible* worlds. We determine the logical and mathematical laws of a world by determining the best logico-mathematical system for a world, and then hold that a world is nearby to the extent that, first, it matches the logical and metaphysical laws of the actual world; second, that it matches the logical and

[84] The (restricted) axiom of comprehension would guarantee that {Socrates} exists whenever Socrates exists, since Socrates would satisfy the predicate 'is Socrates.'

metaphysical matters of fact of the actual world; and third, that is nearby according to Lewis's metric for the relative nearness of possible worlds. On such a view, no world which contains {Socrates} but not Socrates will be closer than any world that contains Socrates but not {Socrates}.

Consider the world that contains Socrates but not {Socrates}; call this world w_1. This world contains a violation of mathematical law, since in this world, there is no set created out of Socrates; the (restricted) comprehension principle has an exception with respect to Socrates. Nonetheless, this need not be a "large" miracle. There is no reason to think that this violation of the principle of comprehension will require any further changes to set theory in general. And while there is at least one change in the space of mathematical facts (*viz.*, the existence of {Socrates}), it is unclear to what extent this change will ramify through the rest of set theory.[85] However the facts of set theory work out in these worlds, none of the facts about history or culture or physics would change.

Compare this world to the nearest world that contains {Socrates} but not Socrates; call this world w_2. This world will have violations of mathematical law analogous to those mentioned above, since there will be a set such that its sole element does not exist.[86] But this is not the only way in which this world is different from ours: it diverges from our world in all the historical and cultural ways that would come along with the disappearance of the philosopher Socrates. This world would diverge radically from the history of the actual world at 469BCE. Thus, w_2 contains not only a small logical "miracle" required to create a set with a member that does not exist – much like w_1 – but also a very large divergence in

[85] We may remain neutral on the question of whether, say, {{Socrates}} would exist at the nearest worlds where Socrates exists without {Socrates}.
[86] This world might also require iterating this violation of law in such a way as to eliminate the existence of sets such as {{Socrates}}. See the previous footnote.

matters of fact. The worlds, then, are alike with respect to the size of logical miracles, analogous with respect to matches of mathematical fact, but w_1 achieves a match of matters of historical fact that w_2 does not.

In short, the world that contains Socrates but not {Socrates} contains a "small" violation of mathematical law, so as to block for the formation of the set {Socrates}. The world that contains {Socrates} but not Socrates also contains a similar "small" violation of mathematical law, so as to create a set out of non-existing members. At this point, our similarity ledger is even between these two worlds, since their laws depart from the actual world in analogous ways. However, world w_1, like the actual world, is a world that contains Socrates, and has a history influenced by Socrates; world w_2 does not have such a history like ours. Both worlds contain a similar "small" miracle, but w_1 matches the actual world in a way that w_2 does not. Thus, the world w_1 will be closer to the actual world than w_2.

It should be noted that, even if the present analysis of counterpossibles fails, our intuitions about counterfactuals with necessarily false antecedents are robust enough to at least be a guide to dependence. That is, insofar as we are inclined to agree that, if either Socrates or {Socrates} were to fail to exist, then it would be {Socrates} that did not exist, the counterfactual exclusion analysis of dependence is on the right track.

We are now in a position to see how this theory of dependence can handle other examples of grounding. We have an intuition that the empty set grounds the existence of the set composed of the empty set, that $\{\varnothing\}$ grounds the existence of $\{\{\varnothing\}\}$, and so on. The analysis of dependence as counterfactuals supports this judgment: if only one of either $\varnothing$ or $\{\varnothing\}$ were exist, then it would be $\varnothing$ that would exist. The nearest worlds that contain $\varnothing$ but not $\{\varnothing\}$ are closer than any world that contains $\{\varnothing\}$ but not $\varnothing$. This counterfactual, to be sure, is somewhat less intuitive than the analogous counterfactual about

Socrates and {Socrates}. Nonetheless, the default similarity metric might be called on to provide some guidance.

Call a nearby impossible world that contains $\varnothing$ but not $\{\varnothing\}$ w_3; call a nearby impossible world that contains $\{\varnothing\}$ but not $\varnothing$ w_4. Both of these worlds must contain some kind of violation of mathematical law, either to prevent the creation of $\{\varnothing\}$ from $\varnothing$, or to allow a set to exist without one of its elements. Thus, both of these worlds contain an analogous violation of mathematical law. However, the violation of law in w_4 extends much farther than just the creation of some set from a non-existent element. After all, a systematization of the mathematical facts of this world must look quite different from a systematization of our mathematical facts, since our mathematical facts make use of the set theoretic hierarchy, which is itself founded on the null set. Since there is no null set in w_4, some other way of systematizing the natural numbers and other facts related to them is necessary. By comparison, a world such as w_3 will still have different mathematical laws by virtue of lacking $\{\varnothing\}$, but this change need not result in changes to how we understand the set-theoretic hierarchy. Other natural numbers, for example, need not be built from $\{\varnothing\}$, and the non-existence of $\{\varnothing\}$ need not mean that $\{\{\varnothing\}\}$ is problematic. Thus, a world like w_3 will be closer to the actual world than w_4; the existence of $\{\varnothing\}$ is more "modally fragile" than the existence of $\varnothing$. The counterfactual requirement is met, and so it is true that $\{\varnothing\}$ depends on $\varnothing$, and not *vice versa*.[87]

5.4 Explanation and Counterfactual Exclusion

[87] Can we say that $\{\{\varnothing\}\}$ depends on $\{\varnothing\}$? This seems more difficult to show. Perhaps it is still more difficult to systematize the mathematical fact of a world that lacks $\{\varnothing\}$ than it is to systematize the facts of a world that lacks $\{\{\varnothing\}\}$, but it is unclear that this judgment can be sustained. More would have to be said about how simple mathematical facts ground more complex mathematical facts.

This theory of dependence is not complete as it stands. The counterfactual exclusion requirement tracks a relation that I have referred to as "modal fragility," which is not the same as dependence. That is, counterfactual exclusion is only useful in determining which of some pair of propositions can more easily be true without the other; this seems to correspond to which proposition of this pair is more basic than the other. But one proposition can be more basic than the other without the more basic proposition grounding the less basic proposition: consider the propositions *that the moon exists* and *that 2 + 2 = 4*. Perhaps we could wed the supervenience requirement to counterfactual exclusion so as to capture the idea that the grounded and grounding propositions are properly related to one another: the supervenience requirement, after all, would rule this out as a case of grounding, since the existence of the moon does not supervene on the truth of 2 + 2 = 4. Thus, the supervenience requirement handles the above case, since one can alter whether the moon exists without altering whether 2 + 2 = 4. However, we can consider another pair, such as *that 2 + 2 = 4* and *that properties are universals* (assuming that properties are universals). The supervenience requirement is met, because there is no possible world that differs with respect to the truth of arithmetic but does not differ with respect to the truth of the nature of properties; the counterfactual exclusion requirement is met since, plausibly, the nearest impossible world where one of these is true but not the other is a world where properties are not universals, but 2 + 2 = 4. Thus, according to a hybrid theory of dependence that combines supervenience with counterfactual exclusion, the truth of properties as universals depends on the truth that 2 + 2 = 4, but this seems obviously false. If supervenience cannot be called upon to solve this problem, how are we to understand these examples?

This challenge can be met, but only with some additional machinery. It is not enough that the grounded propositions are more "modally fragile" than the grounding

propositions, and adding a supervenience requirement does not seem to help. In the kinds of cases described above, the counterfactual exclusion account goes wrong because the relation between propositions in the grounding relation needs to be stronger than that of relative modal fragility; the dependence relation must also capture the intuition that the less fundamental proposition is true *because* (in some sense of because) of the more fundamental propositions. The notion of explanation discussed above in section 1.3 capture this sense of "because." Thus, our final analysis of metaphysical dependence is a combination of counterfactual exclusion with explanation:

> (Counterfactual Exclusion With Explanation) A depends on B iff, a) if exactly one of either A or B were true, then B would be true, and b) A is explained by B.

This approach combines the counterfactual exclusion requirement with the modal requirement; the explanation requirement provides the requisite "because" link between the *explanandum* and *explanans*, while the counterfactual exclusion requirement provides the right kind of modal link between the two.

Consider once again the case of Socrates and his singleton. We have seen above that the counterfactual exclusion requirement is met: if only one of Socrates or his singleton were to exist, then it would be Socrates, and not {Socrates}, that would exist. We have also seen that the explanation requirement is met as well: describing the world in terms of physical objects and their properties, and using those to then derive truths about set theoretic entities is more economical than a description of the world in terms of set theoretic entities that attempts to derive facts about physical objects from facts about sets. Thus, the present theory preserves the result that {Socrates} depends on Socrates.

Such an approach provides an answer to the worry that complex truths about properties should not turn out to be grounded in simple truths about arithmetic. Even

though the truth of some theory of properties is more modally fragile than some simple truth of arithmetic, there is no relation of dependence between them because truths about arithmetic do not explain truths about the nature of properties. That is, even if the counterfactual part of the analysis of dependence is met in this case, the requirement than the more-fundamental proposition explains the less-fundamental proposition is clearly not met.[88]

Analyzing dependence as counterfactual exclusion and explanation is also an improvement over an account of dependence as explanation alone. Recall that at least some of these accounts wrongly hold that the present depends on the past and the laws, since the past and the laws explain the present. This is not a problem for dependence as counterfactual exclusion and explanation: though the state of the world at some time t_1 is explained by the laws and some previous state at t_0, the counterfactual exclusion requirement is not met. Is the world where the state of the world at t_1 does not obtain, but the laws and the state of the world at t_0 are held fixed closer than the world where state of the world at t_1 is held fixed, but the state of the world at t_0 or the dynamic laws are different? It seems that there is no basis to favor one result over the other. Assuming that determinism is true, the first scenario represents a failure of some state of the world and dynamic laws to produce some subsequent state of the world, while the second scenario represents a failure of some later state of the world and the dynamic laws (run backward) to produce a previous state of the world, and thus both worlds represent the same kind of failure of some state of the world and dynamic laws to derive other states of the world.[89] Furthermore, both of these

[88] The same reply can be given to the challenge that the truth of *that 2 + 2 = 4* should not ground the truth *that the moon exists*.

[89] I assume that, when we hold fixed the state of the world at t_1 but change the state of the world at t_0 *or* the dynamic laws, we will, at least in the default context, change the state of the

worlds would provide the same degree of match of matters of fact, either from t_1 into the future, or from t_0 into the past. Since the counterfactual exclusion requirement is not met, we need not say that the some state of the world depends on a previous state and the laws.

Non-causal versions of the explanation theory seem to be committed to the claim that facts about the period of a pendulum are grounded in facts about the length of the rod, the force of gravity, and math, or that facts about the collapse of a star are grounded in its state and the Pauli exclusion principle. The present theory does not face these difficulties because, once again, the counterfactual exclusion requirement is not met. There is no reason to think that that there is any clear answer about what would be the case if we were to either change the period of a pendulum but hold fixed facts about the length of the rod and the relevant laws, or change the length of the rod or relevant laws but hold fixed facts about the period of the pendulum: there seems to be no reason to favor one kind of change over another. Similarly, there is no reason to think that we should change facts about the present state of the collapsing star and hold fixed the Pauli exclusion principle and the star's previous state, rather than changing facts about the previous state of the star and hold fixed the Pauli exclusion principle and the star's present state. This proposal, then, represents an advance over our previous theories of dependence in so far as it gets the central cases right and avoids the counterexamples discussed above.

5.5 Physicalism

One of the central examples of metaphysical dependence is the dependence of the mind on the body. The Counterfactual Exclusion with Explanation account faces at least

---

world at t_0 and hold the laws fixed. On the current way of talking about the comparative similarity of worlds, gratuitous changes in laws represent larger divergence from the actual world than changes of matters of fact.

two problems with respect to the treatment of this kind of case. First, when we say that the mind depends on the body, we are making a counterfactual exclusion claim that, if only one of either the mental facts or physical facts were to obtain, then it would be the physical facts that would obtain. In the language of the similarity of worlds, we should say that the worlds that contain our mental facts, but not physical facts are not as similar to the actual world as the worlds that contain our physical facts, but not our mental facts. But how can one argue for this claim? Perhaps zombie worlds are closer than ghost worlds, but this would mean rejecting the claim that ghosts are metaphysically possible while ghosts are not or the claim that metaphysically impossible worlds are more distant than metaphysically possible worlds. Though one certainly could reject either of these claims, doing so would represent a significant cost for the theory.

Another worry arises from the multiple realizability of mental states. For example, some mental state might depend on a corresponding physical state, such as when my pain depends on my brain being in a certain physical state. But the moral of the multiple realizability arguments is that a creature might experience pain, even though it does not have the particular brain state associated with the way that some other creatures experiences pain; a mental state such as pain could be realized in the way that it actually is in humans (say, as c-fiber stimulation), or it could be realized in other ways in non-human animals or aliens. But recall the counterfactual exclusion claim: if my pain depends on my brain state, it would follow that, if I were to either to have the particular brain state that I have and not be in pain or be in pain but not have the particular brain state, then I would have the brain state but not be in pain. Is this true? Not according to the multiple realizability argument: after all, I might still be in pain but be in some *other* brain state, *e.g.* if I were a Martian experiencing Martian-pain. Such a world is presumably closer than one in which I have the brain state

that I actually have, but am somehow not in pain. If this is the case, then the counterfactual exclusion analysis fails, because my experience of pain would not depend on my particular brain state.

Maybe such a change in my biological make-up is a great change indeed, and would result in a drastically different kind of world with different kinds of laws related to physics and biology. If I am essentially the kind of being that I am, and thus necessarily experience pain in the way that I do, then such a world is metaphysically impossible. Thus, even though *pain* could be realized in many different ways, *my pain* (or, more generally, the mental state of any being capable of mental states) could only be realized in the way that it actually is. Worlds where I experience pain without c-fiber stimulation, then, are quite distant indeed, and could well be more distant than worlds where I experience c-fiber stimulation without experiencing pain. Perhaps an account along these lines could ultimately be developed and defended, though these essentialist claims do not strike me as highly plausible, and I do not attempt to mount such a defense at present.

If this theory of dependence is to be extended to such cases, then the counterfactual exclusion requirement would have to be changed in some way. Either the requirement would have to be re-written in such a way as to get around these problems, or the similarity metric associated with these counterfactuals will have to allow that (say) zombie worlds will be closer than ghost worlds and that worlds where pain is instantiated by some other physical system are quite distant. Both ways of amending the theory could be spelled out, but would significantly complicate the theory. Further work along these lines is necessary before such a theory of dependence can be applied to the case of mental states depending on physical states.

5.6 Conclusion

It is possible to analyze the relation of metaphysical dependence in terms of counterfactuals and explanation. Thus, when we say that A depends on B, we say that, first, that if only one of either A or B were to be true, it would be B that would be true, and second, that the truth of B explains the truth of A. Note that this account puts a large weight on our ability to understand and evaluate these counterfactuals, as well as the notion of explanation. At least in many of the paradigmatic cases of dependence, however, our intuitions about counterfactuals and explanation are firm enough to bear this weight. In other cases, the standard similarity metric for worlds discussed in chapter 3 might be relied upon to illuminate the truth of these counterfactuals.

On the account of similarity proposed in chapter 3, the notion of lawhood plays a very important role. In addition, the notion of explanation is required to make sense of many of these cases of dependence. Thus, it is plausible then debates about grounding and dependence will turn out to be debates about laws or debates about explanation. Arguments about reductionism in mind or morality, or about priority in mereology, then, are really arguments about how the laws function or about what the facts of explanation are: evaluating these claims involves determining what the best systematization of the laws are in this world, or in some impossible world, and determining what role these laws may play in providing explanations for the various things that are true at this world.

REFERENCES

References

Albert, D. (2000). Time and Chance. Cambridge, Harvard University Press.

Anderson, A. a. B., ND (1975). Entailment: The Logic of Relevance and Necessity. Princeton, Princeton University Press.

Armstrong, D. (1978). A Theory of Universals. Cambridge, Cambridge University Press.

Armstrong, D. (1983). What Is A Law Of Nature? Cambridge, Cambridge University Press.

Bennett, J. (1984). "Counterfactuals and Temporal Direction." Philosophical Review **93**(57-91).

Bennett, J. (2003). A Philosophical Guide to Conditionals. Oxford, Oxford University Press.

Bennet, K. (forthcoming) "Building and Causing."

Berto, F. (2009). Impossible Worlds and Propositions: Against the Parity Thesis. Philosophical Quarterly.

Bricker, P. (2006). "Absolute Actuality and the Plurality of Worlds." Philosophical Perspectives **20**(1): 41-76.

Brogaard, B. a. S., Joe (2008). "Remarks on Counterpossibles. "

Brogaard, B. a. S., Joe (2008). "Why Counterpossibles are Non-Trivial." The Reasoner **1**(1): 6-8.

Chalmers, D. (2002). Does Conceivability Entail Possibility? Conceivability and Possibility. T. a. H. Szabo Gendler, J. Oxford, Oxford University Press**:** 145-200.

Chalmers, D. (2004). "Epistemic Two-Dimensional Semantics." Philosophical Studies **118**(1): 153-226.

Chalmers, D. (2009). Propositions and Attitude Ascriptions: A Fregean Account.

Chisholm, R. (1946). "The Contrary-to-Fact Conditional." Mind **55**(220): 289-307.

Cohen, D. (1988). "The Probem of Counterpossibles." Notre Dame Journal of Formal Logic **29**(1): 91-101.

Dunn, J. M. (1993). "Star and Perp: Two Treatments of Negation." Philosophical Perspectives **7**: 331-357.

Dunn, J (2010). "Fried Eggs, Thermodynamics, and the Special Sciences." British Journal for the Philosophy of Science **62**: 71-98

Elga, A. (2001). "Statistical Mechanics and the Asymmetry of Counterfactual Dependence." Philosophy of Science **68**: 313-324.

Fine, K. (1975). "Critical Notice: Counterfactuals." Mind **84**: 451-458.

Friedman, M. (1975). "Explanation and Scientific Understanding." Journal of Philosophy **71**: 5-19.

Goodman, J. (2004). "An Extended Lewis/Stalnaker Semantics and the New Problem of Counterpossibles." Philosophical Papers **33**(1): 35-66.

Goodman, N. (1947). "The Problem of Counterfactual Conditionals." Journal of Philosophy **44**(5): 113-128.

Hall, N. (2000). "Causation and the Price of Transitivity." Journal of Philosophy **97**: 198-222.

Hall, N. (2003). "Two Concepts of Causation." in J. Collins, N. Hall and L. Paul (eds.), Causation and Counterfactuals. Cambridge, Mass.: MIT Press, 225-276.

Handfield, T. (2004). "Counterlegals and Necessary Laws." Philosophical Quarterly **54**(216): 402-419.

Hawthorne, J. (2005). "Chance and Counterfactuals." Philosophy and Phenomenological Research **70**(2): 396-405.

Hempel, C. (1965). "Aspects of Scientific Explanation" in Aspects of Scientific Explanation and Other Essays. New York, Free Press.

Hintikka, J. (1962). Knowledge and Belief. Ithaca, Cornell University Press.

Hofweber, T. (2009). "Ambitious, Yet Modest Metaphysics." In Metametaphysics: New Essays on the Foundations of Ontology, D. Chalmers, D. Manley, and R. Wasserman (eds.) Oxford University Press.

Horgan, T. (1993). “From Supervenience to Superdupervenience: Meetings the Demands of a Material World.” Mind 102: 555-86.

Jackson, F. (1977). "A Causal Theory of Counterfactuals." Australasian Journal of Philosophy **55**(1): 3-21.

Kallestrup, J. (2008). "Conceivability, Rigidity, and Counterpossibles." Synthese.

Kim, S. a. M., Cei (2006). "Counterfactuals as Short Stories." Philosophical Studies **129**(1): 81-117.

King, J. (2007). The Nature and Structure of Content. Oxford, Oxford University Press.

King, J. (2007). "What in the World are the Ways Things Might Have Been?" Philosophical Studies **133**(3): 443-453.

Kitcher, P. (1989). Explanatory Unification and the Causal Structure of the World. Scientific Explanaiton. P. a. S. Kitcher, W. Minneapolis, University of Minnesota Press.

Kment, B. (2006). "Counterfactuals and Explanation." Mind **115**: 261-310.

Kment, B. (2006). "Counterfactuals and the Anlysis of Necessity." Philosophical Perspectives **20**: 237-302.

Krater, A. (1977). "What 'Must' and 'Can' Must and Can Mean." Linguistics and Philosophy **1**(3): 337-355.

Kratzer, A. (1986). "Conditionals." Chicago Linguistics Society.

Kvart, I. (1986). A Theory of Counterfactuals. Indianapolis, Hackett.

Leftow, B. (2006). "Impossible Worlds." Religious Studies **42**: 393-402.

Lewis, D. (1970). "General Semantics." Synthese **22**(1): 18-67.

Lewis, D. (1970). "How To Define Theoretical Terms." Journal of Philosophy **67**(13): 427-446.

Lewis, D. (1973). Counterfactuals. Oxford, Blackwell.

Lewis, D. (1973). "Counterfactuals and Comparative Possibility." Journal of Philosophical Logic **2**(4): 418-446.

Lewis, D. (1979). "Counterfactual Dependence and Time's Arrow." Nous **13**(4): 455-476.

Lewis, D. (1982). "Logic for Equivocators." Nous **16**(3): 431-441.

Lewis, D. (1983). "New Work for a Theory of Universals." Australasian Journal of Philosophy **61**: 341-377.

Lewis, D. (1986). On the Plurality of Worlds. Oxford, Blackwell.

Lewis, D. (1986). "Causal Explanation." In Philosophical Papers, Vol 2. Oxford University Press.

Lewis, D. (1994). "Humean Supervenience Debugged." Mind **103**: 373-390.

Lewis, D. (2000). "Causation as Influence." Journal of Philosophy **97**: 182-197.

Loewer, B. (1996). "Humean Supervenience." Philosophical Topics **24**: 101-126.

Ludlow, P. (2000). Interpreted Logical Forms. The Pragmatics of Propositional Attitude Reports. K. Jaszczolt. Oxford, Elsevier Science**:** 31-42.

Lycan, W. (1994). Modality and Meaning, Springer.

Mancosu, P. (2010). Beyond Unification. The Philosophy of Mathematical Practice. P. Mancosu. Oxford, Oxford University Press**:** 151-178.

Mares, E. (1997). "Who's Afraid of Impossible Worlds?" Notre Dame Journal of Formal Logic **38**(4): 516-526.

Mares, E. (2004). Relevant Logic: A Philosophical Introduction. Cambridge, Cambridge University Press.

Maudlin, T. (2007). The Metaphysics Within Physics. New York, Oxford University Press.

McDaniel, K. (2007). "Extended Simples" Philosophical Studies **133**(1):131-41.

Naylor, M. B. (1986). "A Note on David Lewis's Realism about Possible Worlds." Analysis **46**(1): 28-299.

Nolan, D. (1997). "Impossible Worlds: A Modest Approach." Notre Dame Journal of Formal Logic **38**(4): 535-572.

Poland, J. (1994). Physicalism: The Philosophical Foundations, Oxford: Clarendon.

Priest, G. (1979). "The Logic of Paradox." Journal of Philosophical Logic **8**: 219-241.

Priest, G. (1990). "Boolean Negation and All That." Journal of Philosophical Logic **19**(2): 201-215.

Priest, G. (1997). "Sylvan's Box: a Short Story and Ten Morals." Notre Dame Journal of Formal Logic **38**(4): 573-582.

Priest, G. (2006). In Contradiction: a Study of the Transconsistent. Oxford, Oxford University Press.

Prior, A. (1957). Time and Modality. Oxford, Clarendon Press.

Rescher, N. a. B., Robert (1979). The Logic of Inconsistency: A Study in Non-Standard Possible-World Semantics and Ontology, Rowman and Littlefield.

Restall, G. (1997). "Ways Things Can't Be." Notre Dame Journal of Formal Logic **38**(4): 583-596.

Restall, G. (2000). An Introduction to Substructural Logics. London, Routledge.

Rosen, G. (1990). "Modal Fictionalism." Mind **99**(395): 327-354.

Rosen, G. (2006). The Limits of Contingency. Identity and Modality. F. MacBride. Oxford, Oxford University Press**:** 13-39.

Rosen, G. (2010). Metaphysical Dependence: Grounding and Reduction. Modality: Metaphysics, Logic, and Epistemology. B. a. H. Hale, A. Oxford, Oxford University Press**:** 109-136.

Routley, R. a. M., RK (1983). Relevant Logics and its Rivals. Atascardero, CA, Ridgeview.

Sainsbury, R. (2007). Option Negation and Dialethias. The Law of Non Contradiction. B. Priest, Armour-Garb. Oxford, Oxford University Press.

Salmon, N. (1984). "Impossible Worlds." Analysis **44**(3): 114-117.

Salmon, W. (1984). Scientific Explanation and the Causal Structure of the World. Princeton, Princeton University Press.

Schaffer, J. (2000). "Trumping Preemption." Journal of Philosophy **9**: 165-181.

Schaffer, J. (2004). "Counterfactuals, Causal Independence, and Conceptual Circularity." Analysis **64**(4): 299-309.

Schaffer, J. (2009). “On What Grounds What.” In Metametaphysics: New Essays on the Foundations of Ontology, D. Chalmers, D. Manley, and R. Wasserman (eds.) Oxford University Press.

Schaffer, J. (2010). "The Least Discerning and Most Promiscuous Truthmaker." Philosophical Quarterly **60**: 307-324.

Scheffler, ?. (1957). "Explanation, Prediction, and Abstraction.” British Journal for the Philosophy of Science **7**:293-309

Schneider, B. (2010). "A Puzzle About 'Because'." Logique et Analyse **53:** 317-43

Shoemaker, S. (1998). "Causal and Metaphysical Necessity." Pacific Philosophical Quarterly **79**: 59-77.

Sober, E. (1983). “Equilibrium Explanation” Philosophical Studies **43**:201-10

Sorensen, R. (2002). The Art of the Impossible. Conceivability and Possibility. T. a. H. Gendler, John. Oxford, Oxford University Press**:** 337-368.

Stalnaker, R. (1968). A Theory of Conditionals. Studies in Logical Theory. N. Rescher, Blackwell**:** 98-112.

Stalnaker, R. (1980). "In Defense of Conditional Excluded Middle."

Stalnaker, R. (1987). Inquiry. Cambridge, MIT Press.

Stalnaker, R. (1991). "The Problem of Logical Omniscience, I." Synthese **89**(3): 425-440.

Stalnaker, R. (1996). "Impossibilities." Philosophical Topics **24**(1): 193-205.

Stalnaker, R. (1999). Context and Content. Oxford, Oxford University Press.

Stalnaker, R. (2003). Ways a World Might Be: Metaphysical and Anti-Metaphysical Essays. Oxford, Oxford University Press.

Trogdon, K. (2010). "Intrinsicality for Monists (and Pluralists)." Australasian Journal of Philosophy **88**(3): 555-558.

Vander Laan, D. (1997). "The Ontology of Impossible Worlds." Notre Dame Journal of Formal Logic **38**(4): 597-620.

Vander Laan, D. (2007). Counterpossibles and Similarity. Lewisian Themes: The Philosophy of David K Lewis. F. J. a. G. Priest. New York, Oxford University Press**:** 258-276.

Van Fraassen, B. (1977). “The Pragmatics of Explanation.” American Philosophical Quarterly **14**:143-50

Varzi, A. (1997) “Inconsistency Without Contradiction”, *Notre Dame Journal of Formal Logic*, **38**: 621–39.

Wasserman, R. (2006). "The Future Similarity Objection Revisited." Synthese **150**(1): 57-67.

Weatherson, B. (2001). "Indicative and Subjunctive Conditionals." Philosophical Quarterly **51**(203): 200-216.

Wierenga, E. (1998). "Theism and Counterpossibles." Philosophical Studies **89**(1): 87-103.

Williamson, T. (2007). The Philosophy of Philosophy. Oxford, Blackwell.

Yablo, S. (2009). "Semantic Arithmetic." Forthcoming.

Yagisawa, T. (1987). "Beyond Possible Worlds." Philosophical Studies **53**: 175-204.

Zagzebski, L. (1990). What if the Impossible Had Been Actual? Christian Theism and the Problems of Philosophy. M. Beaty. South Bend, University of Notre Dame Press**:** 165-183.

Zalta, E. (1997). "A Classically-Based Theory of Impossible Worlds." Notre Dame Journal of Formal Logic **38**(4): 640-660.