Counterfactual Conditionals Orthodoxy and Its Challenges

Daniel Dohrn

Contents

Introduction: Counterfactual Conditionals in the Philosophy of Language

There is an intense debate in the philosophy of language and linguistics on so-called counterfactual conditionals. I shall introduce what I take to be the standard view:

First, counterfactuals have truth–conditions.
Second, these truth–conditions can be spelled out in terms of possible worlds.
Third, the possible worlds deciding on the truth or falsity of a counterfactual are those that minimally differ from the actual world.

Then I shall point out selected challenges to the standard view. I shall discuss proposals how to deal with these challenges within and outside of the standard semantics. I am especially interested in discussing in how far the standard view can be preserved and amended in a friendly way. I do not aim at comprehensively covering the main topics in the debate. Given how huge and complicated the debate has become, any selection is inevitably idiosyncratic. I admit that my selection could be more balanced. It is guided by my personal interests and my intellectual biography. The most salient lacuna may be that I set aside the highly important debate on counterfactuals and causal modeling. Another imbalance is that I focus on details of particular positions. My defence is that the positions discussed are paradigmatic. Delving into these positions in some detail, I aim at sharpening our sensitivities for the level of details at which the pertinent problems have to be tackled. I hope to give the reader some idea of the complexity of the issues involved and to motivate her to further pursue the topics considered, filling the gaps my discussion has left.

Besides my research interests in the philosophy of language, I am also interested in the history of philosophy. One of the main reasons why I take the history of philosophy to be interesting is that it allows us to appreciate why certain issues rouse to salience as contrasted to other issues that might as well merit our interest; why certain background assumptions and methodological presuppositions became commonly shared as contrasted to others that might as well be true and relevant; and how these salient issues and background assumptions shaped the development of the debate. I shall use this book as an opportunity to deepen my perspective by a historical dimension. I am not only interested in following the main arguments, but also in the meanderings that gave rise to the richness and variety of current debate.

The book is divided into two parts. The first part is mainly reconstructive. I survey the development that has led to the standard account. I construe this development as driven by one main issue: motivating general truth–conditions for counterfactuals. My main aim is to identify the critical breaking points at which the problems outlined in the second part arise. I shall focus on four paradigmatic accounts: Goodman's discovery of what he calls the problem of cotenability, Stalnaker's version of the standard semantics, Lewis's refinements of Stalnaker's version, and eventually Kratzer's integration of the standard account into a more general semantics for modal expressions.[1]

The second part is more original, comprising research on the problems identified in the first part. Again I emphasize that I do not aim at being exhaustive. Rather I focus on highly selective interventions at particular neuralgic points in the debate. My selective interventions illustrate the pertinent problems and exemplary ways of reacting to them. I shall give an overview of my interventions.

[1] Nelson Goodman, 'The Problem of Counterfactual Conditionals', *The Journal of Philosophy*, 44 (1947), 113–128; Robert Stalnaker, 'A Theory of Conditionals', *Studies in Logical Theory. American Philosophical Quarterly Monograph*, 2 (1968), 98–112; David Lewis, *Counterfactuals* (Oxford: Blackwell, 1973); Angelika Kratzer, 'Modality', in *Semantics*, ed. by Arnim von Stechow and Dieter Wunderlich (Berlin: DeGruyter, 1991), pp. 639-650.

I shall start with (2.) the *logics* that was an important achievement of Lewis's version of the standard account. That logics invalidates certain principles which hold e.g. for simple material conditionals, in particular transitivity, strengthening the antecedent, and contraposition. In section (2.1.) I consider an exemplary attempt at restoring these principles. My second intervention concerns the assumption that counterfactuals have *truth–conditions*. In section (2.2.), I consider whether Gibbard's counterexamples, which are purported to show that indicative conditionals lack truth–conditions, can be transferred to counterfactuals. The third challenge to be considered arises from the interaction of counterfactuals and *probability*. In (2.3.) I shall present some problematic intuitions about counterfactuals with 'probably'. I shall discuss two paradigmatic semantics which accommodate such intuitions before introducing my own proposal, which preserves the letter of the standard account. In section (2.4.), I shall address the notorious future similarity objection to Lewis's proposal how to measure minimal divergence from the actual world. Among the ramifications of the future similarity objection, I shall discuss Morgenbesser cases in section (2.4.1.) and problems with further counterexamples to Lewis's similarity metrics in the sections (2.4.2.) and (2.4.3.). The latter counterexamples all deal with Lewis's claim that antecedent worlds easily diverge from but not so easily converge on the actual world. In section (2.5.), I shall further discuss issues about minimally diverging worlds, in particular what happens if these worlds turn out to be relevantly 'deviant' or atypical. I close with discussing intricacies of future-directed 'would' in section (2.6.). As indicated, the thrust of my debate is to see in how far we can preserve the cherished standard semantics as summarized in section (1.) and where the breaking points might be.

1. The Basics
1.1. Goodman and the Problem of Cotenability

A nice starting point for a study on counterfactuals is a google n–gram search. Such a search quickly reveals that the widespread use of the term 'counterfactual' is of recent origin. Two key texts by Chisholm and Goodman mark the beginning of the debate.[2] I shall focus on the latter. Goodman's work was instrumental in spreading the label 'counterfactual' for a kind of expression which Chisholm had still called the 'contrary-to-fact' conditional. Goodman also assembled some key topics, which would shape the debate to come. Goodman's interest was driven by the relevance of counterfactual conditionals to the debate on central terms in the philosophy of science. He pointed out difficulties for understanding the semantics of counterfactuals, and he outlined a paradigmatic way of tackling them, which led to a well-known problem. I shall take a closer look at Goodman's seminal work. Since I am also interested in how the debate historically evolved, in the changing patterns of interest, research programs, and paradigms, I find it important to give the reader a sense of how Goodman's argument evolves up to his famous problem of *cotenability*.

Goodman starts with noting that the analysis of counterfactual conditionals is crucial to understanding natural laws, theory confirmation, dispositions, and causality (p. 113). Useful as counterfactuals seem, they nevertheless posit substantial difficulties, as can be shown by confining attention to genuine counterfactuals, i.e. those with actually false antecedent and consequent. For these, the tempting truth-functional analysis as a material conditional is obviously false. A material conditional is true precisely if either the antecedent is false or the consequent is true. Consider a piece of butter which has always been kept in the refrigerator:

(A1) If the butter had been heated to 150°, it would have melted.
(A2) If the butter had been heated to 150°, it would not have melted.

Obviously, the first is true and the second is false, but the material conditional is true in both cases.

Our intuitions need to be accounted for. Such an account cannot simply consist in collecting empirical evidence, say by heating the butter to see whether it melts or not, as that would lead to losing the contrary-to-fact–status of the counterfactual. Given the unavailability of these alternatives, Goodman claims that we have to address the peculiar *connection* between the antecedent and the consequent. The consequence that there must be some connection between the antecedent and the consequent is tempting but not trivial. As we shall see, Stalnaker rejects it, pointing to counterfactuals which are made true simply because the consequent holds irrespectively of whether the antecedent is true or not.

Goodman suggests that the antecedent has to bear on the consequent, but the relationship rarely is one of logical consequence. For instance, take a dry match which is not struck:

(A3) If the match had been struck, it would have lighted.

(A3) seems true, but the consequent is not logically entailed by the antecedent. The same obviously goes for (A1). There must be something inexplicit that mediates the connection, certain background conditions S like: the match is dry, well-made, there is oxygen, and so on.

The question becomes how to construe their mediating role. One suggestive proposal is that the transition is mediated by logical entailment. The consequent is logically entailed by the conjunction of S&A. In accepting a counterfactual, we do not merely claim that the consequent

[2] Roderick Chisholm, 'The Contrary–to–Fact Conditional', *Mind,* 55 (1946), 289–307; Goodman, 'The Problem of Counterfactuals'.

follows from the antecedent *if* background conditions S are satisfied. Rather we commit ourselves to their being satisfied.

However, how are we to confine S? We cannot simply add everything that is actually true as among these truths is the denial of the antecedent A *the match has been struck*. Actually, it is not the case that the match has been struck. We cannot either state the condition as follows: there is a set of actually true sentences or propositions from which in conjunction with A C follows. For from A and not–A everything follows.

It is not sufficient either to exclude not–A as there are other true sentences whose conjunction entails everything. Take

(A4) If that radiator had been frozen, it would have broken.

Now take empty and thus true generalizations like 'all radiators which freeze without reaching 0°C freeze' and 'all radiators which freeze without reaching 0° don't freeze'. Conjoining these, everything follows, including that the radiator does not break. Another example: assume Jones is not in Carolina. Consider

(A5) If Jones had been in Carolina, he would have been in North Carolina.
(A6) If Jones had been in Carolina, he would have been in South Carolina.

Jones being in Carolina together with 'Jones is not in North Carolina' and 'Jones is not in South Carolina' again entails everything.

It won't work either to require that the conjunction of A and S not entail a contradiction, as we can get the truth of (A5) from A and 'Jones is not in South Carolina' and the truth of (A6) from A and 'Jones is not in North Carolina'. Yet we can well imagine circumstances under which neither of these would be true. It is simply not settled where Jones would have been.

Up to this point, I have only considered the positive requirement that S and A must entail the consequent. Goodman considers adding a negative requirement: there should be no S* either such that S*&A entail non–C without entailing a contradiction. However, there are counterexamples even to this condition.

Consider again

(A3) If the match had been struck, it would have lighted.

(A3) seems true under normal circumstances. Moreover, the following seems false:

(A7) If the match had been struck, it would not have been dry.

Among the candidates for S is also the actual truth that the match did not light. If we conjoin this truth with A and the other background conditions save the match being dry, we get that the match cannot have been dry. For the match not to light, one of the normal conditions under which a match lights when struck must be given up. If we uphold all the other conditions, we get that the match cannot have been dry.

At this point, Goodman sees only one way to ascertain the right result: we must require that no P be part of S such that

If A had been the case, P would not have been the case.

This gives us the requirement of cotenability. The facts S that are conjoined with A must be *cotenable* with A. A must not counterfactually entail that these facts do not obtain. We get the following conditions:

A counterfactual A>>C (If A had been the case, C would have been the case) is true precisely if C is entailed by A in conjunction with a set of true sentences S such that any sentence P that forms part of S is cotenable with A.

A sentence P is cotenable with A precisely if

Not (A>> not–C)

As Goodman admits, this leaves us with a problem rather than a solution:

'Thus we find ourselves involved in an infinite regressus or a circle; for cotenability is defined in terms of counterfactuals, yet the meaning of counterfactuals is defined in terms of cotenability. In other words, to establish any counterfactual, it seems that we first have to determine the truth of another. If so, we can never explain a counterfactual except in terms of others, so that the problem of counterfactuals must remain unsolved. Though unwilling to accept this conclusion, I do not at present see any way of meeting the difficulty.'(p. 121)

This, then, is the notorious problem of cotenability. We cannot determine the truth of a counterfactual without determining which actual truths are cotenable with the antecedent A. Yet we cannot determine which actual truths are cotenable with the antecedent A without considering another counterfactual, and so on.

Goodman has shaped the debate to come not only by leaving us with the conundrum of cotenability. He also raised a number of further important issues. Among these issues is the status of laws. Laws not only are often formulated by exploiting their counterfactual stability, they also figure prominently among the cotenable facts that are used in deriving a consequent like 'the match lights' from an antecedent like 'the match is struck'. It will therefore be important to see what role the distinction of laws and facts has played in the debate. One particular nuance of this distinction concerns antecedents which are unlawful or even impossible in a stronger sense (if such there is). Goodman's example is

(A8) If triangles were squares,…

An account of counterfactuals must take stance on such counterlegals or counterpossibles, as they are called in more recent debate.

As we have seen, Goodman takes a further important theoretical decision. He conceives the relationship between the antecedent, the background conditions, and the consequent as one of entailment. The conjunction of background condition together with the antecedent A logically entails the consequent C. Lewis called the resulting view 'meta-linguistic'.[3] Tempting as it may seem, this decision of Goodman's again is not trivial and might be resisted. Perhaps it is this decision that leads to the cotenability problem in the first place.

[3] Lewis, *Counterfactuals*, section 3.1..

1.2. Minimal Difference/Divergence/Departure: The Stalnaker–Lewis Semantics
1.2.1. Stalnaker

I shall now come to one of the most influential paradigms in the floating world of philosophy, so influential that it is often called the standard semantics for counterfactual conditionals. I refer to the standard semantics as orthodoxy in the title of this book. The first version of the standard semantics is due to Stalnaker, *A Theory of Conditionals*. Stalnaker's declared aim is to provide a general semantics for the *concept* more or less tightly connected to the everyday use of would–conditionals. This semantics is to allow him to single out the contributions of semantics and pragmatics to the truth–conditions of particular counterfactual utterance tokens.

Stalnaker uses an epistemic heuristic to determine the semantics. He asks how we figure out whether a counterfactual is true. Eventually, he settles for the so-called Ramsey test:

'First, add the antecedent hypothetically to your stock of beliefs; second, make whatever adjustments are required to maintain consistency (without modifying the hypothetical belief in the antecedent); finally, consider whether or not the consequent is then true.'(p. 102)

The Ramsey test leads Stalnaker to the following truth–condition:

'Consider a possible world in which A is true, and which otherwise differs minimally from the actual world. *'If A, then B' is true (false) just in case B is true (false) in that possible world.*'(p. 102, m.e.)

I shall call an approach along these lines minimal difference/divergence/departure account. I shall sometimes also talk of closeness or similarity of worlds, worlds being the closer and more similar to each other the less they diverge from each other.

Stalnaker uses Kripkean modal logic to make this truth–condition more precise. Take the set of all possible worlds. Define a relation of accessibility between worlds, which either obtains among all possible worlds or a subset thereof. Add the absurd world at which everything is true. Define a selection function which has a proposition (simplifying, the thought expressed by a declarative sentence, e.g. *that the butter melted* for the sentence *the butter melted*) and a world as its input and a world as its output. The selection function f(A,w) selects for each antecedent A and each world w precisely one world w*. A conditional A>>C is true precisely if the consequent C is true at the selected world w*. The selection function has to satisfy the following conditions, letting w_e be the world at which the conditional is assessed as to whether it is true at w_e, in the normal case the actual world @.

S1. A has to be true at w*.
S2. The selection function selects the absurd world precisely if A is inaccessible from w_e.
S3. If A is true at w_e, $f(A,w_e)=w_e$

This condition S3 ensures that any world is closest to itself in the ordering.

S4. For any w_e and any antecedents A, A*, if A is true in $f(A^*,w_e)$ and A* is true in $f(A,w_e)$, then $f(A^*, w_e) = f(A, w_e)$.

I find it most convenient to express what Stalnaker aims at by this condition as follows:

For any worlds w* and w** that are possible relative to w, either w* or w** is closer to w or both are equally close.

These conditions confine

Stalnaker's Truth–Condition:
A counterfactual A>>C is vacuously true at w precisely if there is no A–world accessible from w.
A>>C is non-vacuously true at world w precisely if the accessible A–world that is minimally different from w is a C–world.

Coming to a critical assessment, I shall take a closer look at the transition between the Ramsey test and Stalnaker's constructive proposal. Stalnaker says:

'The concept of a possible world is just what we need to make this transition, since a possible world is the ontological analogue to a stock of hypothetical beliefs.'(p. 102)

What does Stalnaker mean by an 'ontological analogue'? Stalnaker does not characterize worlds, but we may as a first stab think of them as maximal states of affairs, and of our universe as an example of a world. Assume possible worlds are part of our ontology. Then for any consistent set of beliefs, there is a set of possible worlds at which these beliefs are true.

However, this does not bring home the transition from the Ramsey test to Stalnaker's intuitive truth–condition. It is not clear how a world that minimally diverges from ours save for the antecedent A relates to Ramsey's procedure of hypothetically revising one's beliefs by A. There are several transitions which must give us pause.

First, it is not at all clear in what way Ramsey's procedure invites a minimal revision. One suggestion is that the revision should be conservative in preserving as much of our current belief system as possible. But conservatism is not a matter of course. It would need additional motivation. A more immediately plausible idea is that the revision should preserve just the right beliefs. It should pay due respect to the evidential relationships among beliefs. If the hypothetically accommodated belief sufficiently strongly counts against previously held ones, they ought to be revised, if not, they should be preserved. In sum, it is not obvious in what sense belief revision à la Ramsey is to be minimal.

Even if we grant that the Ramsey test comes with a requirement that the revision be minimal in some sense, it is highly doubtful that this sense is the same in which the world selected by the selection function minimally diverges from the actual world. Stalnaker indicates that pragmatic factors may play a role in both, but that does not forge a suitable connection between the two kinds of minimal revision. In fact, Stalnaker himself provides reasons why such a connection should *not* be expected. In addressing the problem of how to gather empirical evidence for counterfactuals, Stalnaker himself emphasizes the role of those facts shared by the actual world and the antecedent world closest to it that go beyond what we know or believe (pp. 111–112). We invoke closeness to the actual world to take care of facts we do not know. These facts are taken to be the same as in the actual world. But such facts obviously can play no role in the Ramsey test as they do not form part of our belief system.

I surmise that the problem here lies in Stalnaker's lack of recognition for the distinction between counterfactual and indicative conditionals. Stalnaker in *A Theory of Conditionals* does not mention the difference. As a matter of fact, given his endorsement of the Ramsey test as a method of figuring out counterfactuals, it is not fully clear that he can make room for a deep difference among conditionals. However, the difference was later highlighted by Adams using the famous example:[4]

(A9) If Oswald did not kill Kennedy, someone else did it.
(A10) If Oswald had not killed Kennedy, someone else would have done it.

[4] Ernest Adams, 'Subjunctive and Indicative Conditionals', *Foundations of Language*, 6 (1970), 89–94.

Most participants in the debate agree that (given we know that Kennedy was actually killed) (A9) is true but (A10) is false. I find the following example more compelling:

(A11) If Shakespeare did not write Hamlet, someone else did.
(A12) If Shakespeare had not written Hamlet, no one else would have.

(A11) seems true, (A12) seems false. The Shakespeare example is more convincing. It would seem a more tremendous coincidence if another person had written that very same play than if someone else had been poised to kill Kennedy. As a consequence, (A12) seems more obviously false than (A10).

Later Stalnaker saw the requirement to account for the distinction. He suggested that the counterfactual and the indicative conditionals have different felicity conditions.[5] The felicity condition of an indicative is that the antecedent is not ruled out by the common ground in a conversation. For instance, it seems odd to say:

'I know that Shakespeare wrote Hamlet.
(A11) If Shakespeare did not write Hamlet, someone else did.'

At least we need to deal with the intuitive contrast, for instance by stressing '*If*'.
In contrast, it is perfectly fine to say:

'I know that Shakespeare wrote Hamlet.
(A12) If Shakespeare had not written Hamlet, no one else would have.'

While Stalnaker's ingenious approach in terms of minimal departure has some intuitive appeal, so far we do not have a general motivation for it. A more promising line of motivation would be to consider the idea of minimal departure as a constructive solution to Goodman's cotenability problem: Stalnaker does not explicitly mention the problem. Moreover, he explicitly rejects the idea that we have to spell out a connection of relevance between the antecedent and the consequent (p. 101). Stalnaker wants his account to cover counterfactuals in general, including counterfactuals which are true simply because the consequent holds whether the antecedent holds or not, so-called semi–factuals, and counterfactuals which are simply true because the antecedent and the consequent happen to be actually true. In contrast, Goodman held the view that the real category to carve at the joints does not include semi–factuals. Still it seems somewhat plausible to follow the linguistic surface form here, which does not make a difference between counterfactuals and semi–factuals except in the possibility to insert 'even' in the case of the latter.

Another difference between Stalnaker and Goodman is the following: Stalnaker does not commit himself to Goodman's view that the connection between the antecedent and the consequent has to be one of entailment. In contrast, Stalnaker's selection function avoids the commitment to such a connection, which is replaced by truth at a certain world.

Notwithstanding such differences, there surely is a close connection between solving the cotenability problem and the paradigm of minimal divergence. The facts that the antecedent world closest to actuality shares with the actual world together with the antecedent are good candidates for approximating the set of facts from which Goodman's S and S' are to be taken. The closest antecedent world selected by the antecedent A of A>>C easily meets the condition imposed by Goodman on S: no actual P such that A>>not–P forms part of the set of facts from which the consequent is somehow derived, i.e. the set of facts which hold at the world at which

[5] Robert Stalnaker, 'Indicative Conditionals', *Philosophia*, 5 (1975), 269–286.

A and C have to hold for A>>C to be true. Moreover, no facts which meet Goodman's requirement can be expected to fail to be true at that world. The minimum divergence account makes good on the idea that we should assemble the maximum number of actual facts which are eligible for S.

Still the cotenability problem is not sufficient to motivate the minimal divergence approach. The cotenability problem arises from Goodman's meta-linguistic view. According to that view, there must be a relationship of entailment between the antecedent A, some suitable set of propositions S which are true at the actual world and the consequent C. Yet nothing in Stalnaker's approach ensures that the facts which obtain both at the actual world and at the closest antecedent world together with the antecedent entail the consequent. Thus, the account does not solve the cotenability problem. Rather it amounts to an alternative proposal how to mediate the transition from the antecedent to the consequent.

There is a further potential motive for the minimal divergence account, which presumably had a strong impact on Stalnaker: the convenience of using the logics of possible worlds developed by Kripke and the neat logics for counterfactuals that could be attained by using it. Stalnaker's logics is neat in that it validates inferences which are intuitively valid and invalidates principles which are intuitively invalid. Among the valid inferences are modus ponens and modus tollens:

Modus ponens: P>>Q; P; thus Q
Modus tollens: P>>Q; not–Q; thus not–P

Among the invalid inference patterns are strengthening the antecedent, transitivity, and contraposition.

Strengthening the antecedent: P>>Q; thus P&R >>Q
Transitivity: P>>Q; Q>>R; thus P>>R
Contraposition: P>>Q; thus not–Q>>not–P

Whether conditional excluded middle holds is highly contentious:[6]

(P>>Q)∨(P>>not–Q)
It is not the case that both are false: P>>Q and P>>not–Q.

I shall close with mentioning another issue about Stalnaker's account. Stalnaker takes it to be a largely pragmatic matter how the total ordering is determined, i.e. what the closest antecedent worlds are. This allows Stalnaker to keep the account flexible in dealing with examples which seem highly context-dependent, but it gives rise to concerns about the role of counterfactuals in science that motivated authors like Goodman to be interested in counterfactuals in the first place, for instance their role in supporting laws, causality, and so on. For instance, one would not expect it to be a pragmatic matter whether laws hold. To ensure this, the similarity ordering must be subject to certain objective constraints. Stalnaker's theory allows for such constraints but does not account for them.

I conclude my brief survey of Stalnaker's account. I shall now proceed to Lewis's variation of it.

[6] Robert Stalnaker, 'A Defense of Conditional Excluded Middle', in *Ifs: Conditionals, Belief, Decision, Chance, and Time*, ed. by William L. Harper, Robert Stalnaker, Glenn Pearce (Dordrecht: Reidel, 1981), pp. 87–104; J. Robert G. Williams, 'Defending Conditional Excluded Middle', *Noûs*, 44 (2004), 650–668; Nathan Klinedienst, 'Quantified Conditionals and Conditional Excluded Middle', *Journal of Semantics*, 28 (2011), 49–170.

1.2.2. Sobel and Similarity: Lewis

Lewis's (1973) account is so close to Stalnaker that their minimal divergence approach is usually known as the Lewis–Stalnaker semantics and competes with Kratzer's for becoming largely standard in the philosophical debate.[7] There are some subtle differences, though. Lewis weakens some of Stalnaker's assumptions and advocates stronger assumptions where Stalnaker generously and non-committally invokes pragmatics.

Lewis takes up a line of motivation which is already implicit in Stalnaker. In Stalnaker's account, the *counterfactual conditional* is sandwiched between a logically stronger and a logically weaker expression: the *strict conditional* on the one and the *material conditional* on the other hand. Stalnaker does not exploit the motivational potential of sandwiching, though. Doing so becomes the thrust of Lewis's approach to counterfactuals. Lewis carries sandwiching one step beyond Stalnaker.

As distinguished from Stalnaker, Lewis starts with discussing the alternative of a strict conditional approach to counterfactuals. According to such an account, a counterfactual A>>C is true precisely if the material conditional A⊃C holds necessarily. The approach is most easily stated in terms of possible worlds. A counterfactual is true if all worlds in which A is true are worlds in which C is true. In short, all A–worlds are C–worlds. Lewis uses the logical properties identified by Stalnaker to argue against the strict conditional approach. The latter validates strengthening the antecedent and transitivity. Yet Lewis goes further. He develops a suggestive iterative strategy to plausibilize a closeness ordering of worlds à la Stalnaker.

Consider the following example, which Lewis calls a '*Sobel sequence*':

(A13) If the US had thrown their nuclear weapons into the sea, there would have been war.
(A14) If the US and the other superpowers had thrown their nuclear weapons into the sea, there would have been peace.
(A15) If the US and the other superpowers had thrown their nuclear weapons into the sea and thereby caused a breakdown of global fishery industries, there would have been war.
…

Another example from the literature is:[8]

(A16) If Sophie had gone to the parade, she would have seen Pedro dance.
(A17) If Sophie had gone to the parade and got stuck behind someone tall, she would not have seen Pedro dance.
(A18) If Sophie had gone to the parade and got stuck behind someone tall and used stilts, she would have seen Pedro dance.
…

The general pattern:

A_1>>C
A_1&A_2>>not–C
A_1&A_2&A_3>>C
…
Conjoining additional propositions A_2, A_3 …with an antecedent A_1 is called *antecedent strengthening*. Two things seem to hold for such sequences. Firstly, under certain assumptions they can be prolonged indefinitely. Second, a great many sequences of this form may be true. This is evidence that the strict conditional approach can't be right.

[7] Lewis, *Counterfactuals*.
[8] Anthony Gillies, 'Counterfactual Scorekeeping', *Linguistics and Philosophy*, 30 (2007), 329–360.

Lewis uses the intuition that such sequences can be prolonged indefinitely to argue for a minimal difference approach to conditionals. However the minimal difference account is spelled out, a world at which all the superpowers throw their weapons into the sea may differ more from the actual world than a world in which only the US throw their weapons into the sea, but not the other way round. For any world in which all the superpowers throw their weapons into the sea ipso facto is a world in which the US do. More generally, no world at which a strengthened antecedent $A_1\&\ldots A_n$ is true can diverge less from the actual world than the closest world at which an unstrengthened antecedent $A_1\&\ldots A_{n-1}$ is true.

These observations allow Lewis to state a neat proposal how Sobel sequences can be true. The closest A_1–world is a C–world. The closest $A_1\&A_2$ is less close than the closest A_1–world. Thus, nothing prevents it from being a not–C–world. The closest $A_1\&A_2\&A_3$–world is less close than the closest $A_1\&A_2$–worlds. Thus, nothing prevents it from being a C–world again and so on. Lewis does not claim that the minimal difference–approach is the only one that allows to account for true Sobel sequences, but the possibility of true Sobel sequences adds further constraints to a semantics of conditionals over and above sandwiching between the strict and the material conditional. We get arbitrarily many further sandwiching layers. Sandwiching thus lends further support to the minimal difference–account.

I note, however, that the proposal depends on certain background assumptions. One assumption is that the semantics is static. Though counterfactuals are granted to be somewhat context-sensitive as it is a pragmatic matter how the ordering of worlds according to their closeness is determined, it is a background assumption that context is not so shifty that its shiftiness could account for the truth of Sobel sequences. Even if such background assumptions are granted, Lewis's approach only provides limited support for the minimal difference approach as long as it has not been shown that there is no other truth-conditional account that posits truth–conditions in between the strict and the material conditional and yields true Sobel sequences.

I shall now consider some points at which Lewis parts ways with Stalnaker. They can be divided into issues where Lewis is less committal than Stalnaker and issues where he incurs stronger commitments. The general motive of being less committal is that the account is in danger of being too restrictive and thus to provoke counterexamples. The general motive of being more committal is that the account at one crucial point does not seem to live up to the expectations concerning the role of counterfactuals in science that have been raised by Goodman and others.

I begin with the points at which Lewis generalizes Stalnaker by loosening some of his more restrictive conditions: one is the so-called *limit assumption*: There is a closest antecedent world whenever a counterfactual is true. This assumption may be questioned. For it may be that, for any antecedent world, there are antecedent worlds closer to actuality. Take comparisons of length. A stick which is 99cm long differs less in length from a stick which is 1m long than a stick which is 98cm long. But a stick which is 99,9cm long comes even closer. It is outranked by a stick which is 99.99 cm long and so on. To appreciate the connection to counterfactuals, consider the following question:

(A19) If that stick had not been 1m long, how long would it have been?

Assume I say, guided by the idea of minimal difference in length: 99cm. This answer seems arbitrary, as there are smaller differences, and the same goes for any other answer. The persuasive power of the comparison is limited, though. There are doubts in how far we are guided by considerations of minimal divergence in length in judging (A19). Rather we are at a loss at answering the question. It is not clear how this result relates to the issue of minimally

diverging worlds, but obviously such comparisons are not guided by dimensions made salient in the way the example makes salient comparisons of length.

A second assumption of Stalnaker's which Lewis eschews is the so-called *uniqueness assumption*: the selection function that figures in the truth–condition of counterfactuals always selects one particular world, which is the closest one. Lewis criticizes that worlds may tie in for being closest. As an example, consider a fair coin. The coin is not tossed. But had it been tossed, would it fallen heads or tails? It seems that worlds at which it falls heads and worlds at which it falls tails tie in for closeness. Still, the following seems true (excluding worlds where the coin does strange things like landing on its margin):

(A20) If the coin had been tossed, it would have fallen either heads or tails.

If we adopt the uniqueness assumption and additionally assume that there is a tie between worlds at which the coin falls tails and worlds at which the coin falls tails, (A20) cannot be true. To avoid this result, we should drop the uniqueness assumption.

If worlds can tie in for closeness, the question becomes salient what happens in cases in which the counterfactual is true in some but not all closest worlds. Take the coin example. None of the following seems true:

(A21) If the coin had been tossed, it would have landed heads.
(A22) If the coin had been tossed, it would not have landed heads.
(A23) If the coin had been tossed, it would have landed tails.
(A24) If the coin had been tossed, it would not have landed tails.

As a consequence, it seems true that it is not the case that if the coin had been tossed, it would have landed heads and so on. This conflicts with the following principle as endorsed by Stalnaker:

Not (A>>C) $\supset$ (A>>not–C)

Stalnaker maintains that counterfactuals are neither true nor false if there is no uniquely closest world. However, he proposes a supervaluationist view of such cases.[9] There are several ways of precisifying a counterfactual which is neither true nor false. For instance, one precisification of the coin counterfactuals selects the closest world in which the coin falls heads, another precisification selects the closest world where the coin falls tails as the closest world where the coin is tossed. For all precisifications the following holds. Either the coin would have fallen heads, or it would have fallen tails. In other words, when C and not–C–worlds tie in for being the closest A–world, for any precisification conditional excluded middle holds:

A>>C $\vee$ A>>not–C

One may doubt this idea of precisification, though. There does not seem to be any room for being more precise in the coin example. We have two well-defined possibilities, and it seems arbitrary to privilege one.

In other cases, though, there seems to be room for precisification. Consider the famous Caesar–counterfactuals:

(A25) If Caesar had been in command in Korea, he would have used catapults.
(A26) If Caesar had been in command in Korea, he would have used the atom bomb.

[9] Stalnaker, 'Conditional Excluded Middle', pp. 87–104.

Lewis says:

‘Thus we account for such pairs of counterfactuals as Quine’s
If Caesar had been in command [in Korea] he would have used the atom bomb.
Versus
If Caesar had been in command he would have used catapults. [...]

I could [...] call on context rather to resolve part of the vagueness of comparative similarity in a way favourable to the truth of one counterfactual or the other. In one context, we may attach great importance to similarities and differences in respect of Caesar’s character and in respect of regularities concerning the knowledge of weapons common to commanders in Korea. In another context, we may attach less importance to these similarities and differences, and more importance to similarities and differences in respect of Caesar’s own knowledge of weapons. The first context resolves the vagueness of comparative similarity in such a way that some worlds with a modernized Caesar in command come out closer to our world than any with an unmodernized Caesar[...] My intuition is to explain the influence of context entirely as the resolution influence.’(pp. 66–67)

According to Lewis, sometimes counterfactuals in isolation are vague with regard to the relevant similarity ordering. Then context may fix the similarity ordering. In the case of the Caesar–counterfactuals, we are first at a loss which one of these counterfactuals to privilege, but it is easy to imagine a context in which we have cues. One may say:

‘Caesar had only the limited knowledge of weaponry that was available to a Roman. He at most knew catapults. He would not have known what to do with modern weaponry.

(A25) If Caesar had been in command in Korea, he would have used catapults.’

This seems somewhat plausible.

Or one may say:

‘Caesar was absolutely ruthless in choosing his means. He always used the most effective weapon to bring down his enemies.

(A26) If Caesar had been in command in Korea, he would have used the atom bomb.’

Again this seems acceptable.

The same goes for a famous example used by Lewis against Stalnaker (p. 82):

(A27) If Bizet and Verdi had been compatriots, Bizet would have been Italian.
(A28) If Bizet and Verdi had been compatriots, Verdi would have been French.

Again we may be at a loss which one is true and which is false. But consider the following:

‘How could Bizet have been Italian? Well, assume his ancestors had moved to Italy, and by a tremendous coincidence all the other circumstances that led to Bizet being begot and born remain untouched. Then Bizet would have been a compatriot of Verdi. In other worlds:

(A27) If Bizet and Verdi had been compatriots, Bizet would have been Italian.’

An analogous story could be told for (A28).

An interesting variant of such ties between opposing pairs of counterfactuals is due to Goodman. Goodman deems the following two counterfactuals true:[10]

(A29) If New York City were in Georgia, then New York City would be in the South.
(A30) If Georgia included New York City, then Georgia would be in the North.

However, the antecedents seem to be logically equivalent. Goodman surmises that word order here gives cues how to enrich the antecedent by 'and the boundaries of Georgia were unchanged' and 'the boundaries of New York were unchanged'. In the Lewis–Stalnaker approach, one may say that the antecedent itself provides minimal pragmatic cues as to what the relevant similarity order looks like.

A third point at which Lewis considers loosening Stalnaker's conditions is the centering condition. For any world w, the world minimally differing from w is w itself.

S3. If A is true at w_a, $f(A,w_a)=w_a$

Since Lewis feels less pressure to settle for uniquely closest worlds, he considers weakening this condition in admitting the possibility that there might be worlds other than w which are as close to w as w itself. This seems plausible given the view that the similarity ordering is pragmatically determined. Such a pragmatic view draws on the idea that some respects of similarity matter more for us than others given our current interests. If the difference between w and w* is not relevant to us in determining the similarity ordering, nothing prevents w* from counting as equally close to w as w itself for the purposes of conversation.

As a consequence, Lewis offers two variants of the centering condition:

Strong centering: For any w, there is no other world w* which is as close to w as w.
Weak centering: For any w, w is among the worlds closest to w.

Lewis's amendments of Stalnaker's restrictive conditions lead him to the following truth–condition for counterfactuals:

Lewis's Truth–Condition:
A counterfactual A>>C is vacuously true precisely if there is no accessible A–world.
A>>C is non-vacuously true at world w precisely if there is an A&C–world which differs less from w, on balance, than any A&not–C–world.

Having discussed the points at which Lewis *weakens* Stalnaker's assumptions, I now come to a respect in which Lewis incurs *stronger* commitments than Stalnaker. It is part and parcel to Stalnaker's project to provide general semantic truth–conditions which capture the concept that guides us in using counterfactuals. He maintains that there is such a unique concept underlying all the vagaries of everyday usage. In order to bring out this concept, he aims at precisely locating the point where semantics ends and pragmatics begins. Pragmatics begins when we ask how the similarity ordering of worlds is determined over and above the formal requirements of the ordering.

Stalnaker is mainly interested in the semantics and leaves the pragmatics open. However, this leads to problems for the ambitions many philosophers, including Stalnaker himself, harbour about counterfactuals. We have seen Goodman pointing out to the scientific relevance of counterfactuals for defining natural laws, theory confirmation, dispositions, and

[10] Goodman, p. 121.

causality. Stalnaker himself suggests to define laws by a counterfactual connection between predicates F and G:[11]

For all x, if x were F, it would be G.

Lewis in turn harbours great ambitions of defining causation by counterfactuals, starting with the intuitive:[12]

A is a cause of B if, had A not happened, B would not have happened.

Stalnaker's view on the pragmatics of determining similarity harbours a problem for such ambitions. Scientifically minded philosophers like Quine distrusted counterfactuals due to the seeming arbitariness of their truth. If we can get almost any counterfactual true and false by conjuring up the right similarity ordering, as witnessed by extreme examples like the Caesar–counterfactuals and the Verdi–Bizet cases, this dependence on pragmatic factors like our varying interests threatens to infect anything that is defined by means of counterfactuals.

There are also principled doubts about our readiness to judge overall similarity. As Keynes noted, a

'[...] book bound in blue morocco is more like a book bound in red morocco than if it were bound in blue calf; and a book bound in red calf is more like the book in red morocco than if it were in blue calf. But there may be no comparison between the degree of similarity which exists between books bound in red morocco and blue morocco, and that which exists between books bound in red morocco and red calf.'[13]

Whenever A is more similar to B than to C in one respect and more similar to C than to B in another, incommensurable respect, we cannot rank B and C with regard to overall similarity. Assume that there are only two alternate ways of binding a book bound in red morocco: blue morocco, red calf. How are we to judge the following:

(A31) If the book had not been bound in red morocco, it would have been bound in blue morocco

versus

(A32) If the book had not been bound in red morocco, it would have been bound in red calf.

Again there are principled doubts that the issue is settled by intuitive similarity, but disregarding such doubts, there is also an issue which consequent describes an overly more similar antecedent situation.

Even granting there is incommensurability, we might not feel overly concerned. There seem to be intuitive trade–offs. The rest is vagueness, one might say. Within limits, we just dissolve incommensurabilities by *fiat*. The brusqueness of such acts can be allayed by embedding them into suitable stories. Assume the book is in a library where books are only bound either in red or in blue morocco. In this case, of course the book would have been bound in blue morocco had it not been bound in red morocco. Or take a person's weight and temperature. We are not in a position to say how much temperature outweighs a difference in weight. But there are constraints, for instance due to normalcy conditions. A body temperature of 39° C is counted as a remarkable deviation from a healthy person. So is a weight of 110 kg.

[11] Stalnaker, 'A Theory of Conditionals', p. 110.
[12] David Lewis, 'Causation', *The Journal of Philosophy*, 70 (1973), 556–567.
[13] John Maynard Keynes, *A Treatise on Probability* (London: Macmillan, 1921), p. 36.

So at first glance, probably a difference in weight between 80 and 85 kg is considered less important than a difference in temperature between 36° and 39° C. Furthermore, contexts contribute to resolving vagueness. In a debate on obesity, weight differences may count more heavily towards dissimilarity than differences in body temperature.

Concerns remain. Firstly, in how far do these contextually bound prima facie similarities live up to our expectation that there are similarities 'out there' to track? Secondly, again the question of the grand story rises. *Overall* similarity cannot mean similarity of a whole to another with regard to particular features, for instance similarity of total worlds with regard to the number of atoms in them. Rather it must mean something like similarity *all things considered*. In how far does our limited perspective converge to a picture 'all things considered'? Lewis tries to get rid of concerns about incommensurabilities by imposing strong restrictions on overall similarity. Some respects are more important than others, some do not count at all.[14]

Concerns about the scientific function of counterfactuals are one factor that leads Lewis to taking a closer look at how the similarity ordering is determined. Another factor is the notorious *future–similarity objection*. The future similarity objection springs from a counterexample of Kit Fine's to the Lewis–Stalnaker semantics. Consider:

(A33) If Nixon had pressed the nuclear button, there would have been a nuclear holocaust.

(A33) seems intuitively true. For it to be true, a world at which a nuclear holocaust ensues must be overall more similar to the actual world than a world at which the signal just fizzles out and history converges to the history of the actual world. But this seems unintuitive. Surely a world at which Earth is devastated by nuclear holocaust in the seventies is overall less similar than a world at which the signal fizzles out and everything returns to normal.

To evade this objection, Lewis develops the following strategy: instead of looking for intuitive similarity, he makes overall similarity of worlds a technical notion. He uses intuitive counterfactuals like the Nixon example to reason back to what our criteria of similarity or minimal difference must be like to support them. However, he also takes guidance from relevant theoretical considerations, in particular concerning the distinction of determinism and indeterminism. As a working characterization of determinism, I propose the following:

A world w is deterministic precisely if there cannot be a world w* such that
(i) w* has the same fundamental laws of nature as w,
(ii) w* perfectly matches w in particular matters of fact at some time t,
(iii) w* does not perfectly match w in particular matters of fact at some other time t*.

If two deterministic worlds match in laws and facts at some time, they cannot be different in facts at a different time.

Lewis then goes on to develop overall criteria of similarity for deterministic worlds. Such worlds raise the following problem: if the antecedent A is not actually true, the closest A–world must differ from the actual world @ at least in A. If that world shares the laws of the actual world, as a deterministic world it must differ from @ at all times throughout the history of the universe. Such a world does not seem very similar, even if we disregard intuitive judgements of similarity. In particular, it seems unmotivated to preserve the laws at all costs. Lewis's alternative is that the closest A–world differs in *laws*. However, since laws are involved in explaining any particular fact, it would seem a weird coincidence either if we had large-scale match in fact but substantial mismatch in the laws. To deal with this problem, Lewis introduces

[14] 'To what extent are the philosophical writings of Wittgenstein similar, overall, to those of Heidegger? I don't know. But there is one aspect of comparison that does not enter into it at all, not even with negligible weight: the ratio of vowels to consonants.'(David Lewis, 'Counterfactual dependence and time's arrow', in *Philosophical Papers II* (Oxford: Oxford University Press, 1986), pp. 32–66 (pp. 41–42)).

his much-criticized idea of a 'small miracle'. Though the label is neat, it is somewhat misleading. Lewis does not have in mind something that runs counter to the laws of nature, at least not to the laws of the antecedent world considered. He only claims that the laws of the closest A–world are minimally different from the actual laws in that they allow for the antecedent to become true. Lewis thinks of the difference as an exception clause for the specific highly localized circumstances that lead to A becoming true.[15] To use an example of Lewis, think of some extra neurons firing unexplainably (given the actual laws) in Nixon's brain and making him press the nuclear button immediately afterwards. Lewis requires that small miracles are spatiotemporally closely confined, and they are rather simple. They do not consist of exceptional events of many different kinds.

Using the notion of a miracle, Lewis has us consider several candidates for the closest world at which Nixon presses the button:

(i) A world which perfectly matches the actual world in laws but differs from it in facts from the very beginning until the end of time.
(ii) a world which perfectly matches the actual world in facts save for the antecedent A but has very different laws.
(iii) a world which perfectly matches the actual world in facts until shortly before the antecedent, then diverges by a small miracle, and then reconverges by a small miracle such as to achieve *perfect* future match in facts.
(iv) a world which perfectly matches the actual world in facts until shortly before the antecedent, then diverges by a small miracle, and then reconverges by a small miracle such as to achieve *approximate* future match in facts.
(v) a world which perfectly matches the actual world in facts until shortly before the antecedent, then diverges by a small miracle without ever matching the actual world again.

Of these worlds, (v) is Lewis's candidate. It underpins (A33). As we have seen, a world which (i) shows mismatch in facts throughout history seems rather dissimilar from the actual world. A world with very different laws (ii) seems even more problematic, given the explanatory role of laws for particular matters of fact. A world that (iii) perfectly reconverges seems a promising candidate. However, Lewis insists that such a world is impossible. Reconvergence by a small miracle is not to be had. The reason is that any event leaves many future traces of different sorts. Here is Lewis's description of these traces:

'[...]Nixon's deed has left its mark on the world[...] There are his fingerprints on the button. Nixon is still trembling[...] His gin bottle is depleted. The click on the button has been preserved on tape. Light waves flew out of the window, bearing the image of Nixon's finger on the button, are still on their way into outer space. The wire is ever so slightly warmed where the signal current passed through it. And so on, and on, and on.'[16]

In Lewis's opinion, a complete cover–up action for these many and varied and spatiotemporally dispersed traces would take not a small but a big, widespread miracle.

What remains is (iv) a world with approximate match. Such a world has some intuitive appeal. For instance, we might think of a small miracle that makes the signal of Nixon's pressing fizzle out while leaving the other traces of the deed. Such a miracle could prevent the nuclear disaster. To get (A33) right, Lewis must insist that approximate match counts for nothing or almost nothing. The justification is his methodology to reason back from the intuitively true counterfactuals to the guiding criteria of similarity.

This, then, gives us Lewis famous lexical ordering of four criteria of similarity:

[15] Lewis, 'Counterfactual Dependence', pp. 54–55.
[16] Lewis, 'Counterfactual Dependence', p. 45.

‘(1) It is of first importance to avoid big, widespread, diverse violations of law [big miracles].
(2) It is of second importance to maximize the spatio-temporal region throughout which perfect match of particular fact prevails.
(3) It is of third importance to avoid even small, localized simple violations of law [small miracles].
(G9) It is of little or no importance to secure approximate similarity of particular fact, even in matters that concern us greatly.’[17]

Our semantics for counterfactuals should not depend on the highly contestable hypothesis that the actual world is deterministic. What about indeterministic worlds? For this case, Lewis replaces the small miracles by a chance event and the big miracles by large-scale counter-entropic developments.

This concludes my discussion of Lewis’s amendment of Stalnaker’s semantics. I shall now briefly look at Kratzer’s general theory of modal expressions.

[17] Lewis, ‘Counterfactual Dependence’, pp. 47–48.

1.2.3. Orthodoxy à la Kratzer

I close the basic presentation of truth-conditional orthodoxy about counterfactuals by taking a quick look at Angelika Kratzer's variant of the minimal difference semantics in terms of possible worlds.[18] This variant has become increasingly influential both in the philosophical and in the linguistic debate. One reason is that the account is highly general. Kratzer starts from a principled analysis of modals like 'could', 'might', probably'. She covers different kinds of modality, epistemic ('perhaps'), deontic ('should'), circumstantial or metaphysical ('sugar is solvable').

Kratzer distinguishes between a *modal base* and an *ordering source*. Both are provided by context or 'conversational background'. The modal base consists of a set of propositions restricting the set of possible worlds to those at which these propositions are true. These are the accessible worlds. The ordering source determines an ordering of these worlds. It provides a set S of propositions. The ordering is determined by the following conditions: the ideal world i is one where all propositions in S are true. A world w* is at least as close to i as a world w** precisely if all propositions in S which are true in w** are true in w*. Kratzer then goes on the define the notion of necessity and graded notions of possibility, which can be used to define the modal auxiliaries in turn. For instance, 'must' is interpreted as quantifying over all possible worlds fixed by the modal base. 'Must A' is true precisely if all worlds in the modal base are A–worlds.

As for a conditional 'If A, C', Kratzer characterizes it as implicitly modalized by a covert modal. Comparable to the role of 'must A', the antecedent A is added to the modal base. It acts as a restrictor on the accessible worlds. All accessible worlds must be A–worlds. These worlds are quantified over. The conditional is true precisely if the consequent C is true in all those worlds which come closest to the world at which the conditional is assessed according to the contextually relevant ordering source.

Kratzer differentiates between several categories of conditionals. The strict conditional results if the modal base and the ordering source are empty in the sense of admitting all possible worlds as accessible. The material conditional results if the ordering source is empty and the modal base contains any proposition that is true at the world of assessment. The counterfactual results from an empty modal base and an ordering source which contains anything that is true at the world of assessment.

Summarizing, Kratzer presents an elegant reformulation of the minimal difference account, which allows her a generalization to all modal expressions. This concludes my brief survey over the basics of the orthodox truth-conditional semantics of counterfactuals. In the second and much longer part of this book, I shall discuss several challenges which arise for the account. I shall mainly use the Lewisian version of the standard account as it is most widespread in the philosophical debates I aim at covering. As indicated at the beginning, my method shall be highly casuistic. I shall outline a general problem and then illustrate it by discussing selected positions on it in some detail.

[18] Kratzer, 'Modality'.

2. Challenges to Orthodoxy
2.1. Logics

I shall start with logics. We have seen that for both Lewis and Stalnaker, the logical properties of counterfactuals play a key role in closing in on the right semantics for counterfactuals. They accept principles as valid and to be respected as far as there are no intuitive counterexamples. As far as there are counterexamples, Lewis and Stalnaker build their view such as to invalidate the corresponding principles. The alternative is that the principles are valid, but the apparent counterexamples can be explained away by a shift in the semantics of the terms used. I shall defend Lewisian orthodoxy against one exemplary way of pursuing this strategy.

In the Stalnaker–Lewis semantics certain logical principles which hold for the material and the strict conditional become invalid: contraposition, strengthening the antecedent, transitivity viz. hypothetical syllogism (CSH). The move has been contested. Berit Brogaard and Joe Salerno aim at defending the validity of these inferences. Here are three examples of intuitively failing inferences (RWH):

`(Reliable John)
[B1] If John had made a mistake, it would not have been a big mistake.
[B2] Therefore, if John had made a big mistake, he would not have made a mistake.

(Wet Match)
[A3] If this match had been struck, it would have lit.
[B3] Therefore, if this match had been soaked in water overnight and struck, it would have lit.

(Hoover)
[B4] If J. Edgar Hoover had been a communist, he would have been a traitor.
[B5] If he had been born a Russian, he would have been a communist.
[B6] Therefore, if he had been born a Russian, he would have been a traitor.´[19]

The standard account, say Brogaard and Salerno, consists of the following conditions:

(TC) 'a subjunctive of the form 'If A had been the case, B would have been the case' is true at a world *w* iff B is true in all the A–worlds closest (or most relevantly similar) to *w.*´
(Vacuous Case) `[...]if there are no closest A–worlds, then vacuously all the closest A–worlds are B–worlds.'
(Context) '[...]closeness is contextually determined, since which worlds are relevantly similar to a given world is a contextual matter.' (p. 40)

Based on these assumptions, Brogaard and Salerno outline a way of interpreting RWH that preserves their validity provided that context is held fixed. The key move is that Brogaard and Salerno only distinguish between possibilities which are closest and possibilities which are inaccessible. This does not follow from the way they describe the standard account. It is a substantial amendment. If we accept it, CSH become valid.

Concerning (Reliable John), Brogaard and Salerno distinguish the following alternatives: for (B1) to be true, John's making a big mistake cannot be among the closest worlds where John makes a mistake. Thus, it must be inaccessible. If context is held fixed, (B2) is vacuously true. The inference is sound. Or, if context is not preserved but changes so as to make worlds in which John makes a big mistake accessible, the inference fails merely due to context failure. Analogously for (Wet match). Concerning (Hoover), there are two cases. Either Hoover's being Russian is considered an *accessible and closest* possible situation in which he is a communist. Then (B4) is wrong. Or it is and remains inaccessible. Hence the inference is valid. Brogaard and Salerno are committed to the following claim: whenever an instance of CSH fails, it fails due to context shift.

[19] Berit Brogaard and Joe Salerno, 'Counterfactuals and Context', *Analysis*, 68 (2008), 39–46 (p. 39).

Brogaard and Salerno direct their criticism against the usual treatment of CSH within the standard account. They refer to Lewis's elaboration of possible–worlds semantics. Their objection is that RWH have not been *shown* to be invalid. For there is the understanding they outline. And given the standard account *above*, they are right. There are versions of *that* account that support their reasoning. However, the *Lewisian version* as it is standard in most contemporary philosophy enshrines stronger commitments. And given these commitments, there are instances of CSH failure that are not due to context shift.

It is crucial how 'closest' in TC is spelled out by Lewis. From the viewpoint of textbook Lewisianism, Brogaard and Salerno cannot rule out instances of CSH failing for the following reason: The antecedent possibility of the second premise, respectively, is accessible *but not as close* as the antecedent possibility of the first premise.

In a Lewisian semantics, it is natural to distinguish two things:

Similarity: An antecedent possibility may be closer than another one; nevertheless both are accessible.
Accessibility: An antecedent is not counted among the accessible possibilities at all.[20]

There is a further difference to textbook Lewisianism: according to Brogaard and Salerno, context decides which possibilities are accessible (p. 40). In Lewis's favourite view, in contrast, we '…call on context … to resolve part of the vagueness of comparative similarity[…]'[21] Lewis concludes: 'My intuition is to explain the influence of context entirely as the resolution influence.'[22] To Lewis, context only resolves the vagueness of comparative similarity.

There might be possibilities which are accessible without being closest. But even given orthodoxy, why not adopt Brogaard's and Salerno's position? The (B2 / A3 / B5) antecedent possibilities are as similar as the (B1 / A3 / B4) ones or inaccessible? To see why not, take a really far-fetched, nomically impossible variant of (Wet Match):

(Wet Match*)
(A3) If this match had been struck, it would have lit.
(B7) Therefore, if the match had been struck and there had been a widespread violation of natural laws guiding the behaviour of valence electrons, the match would have lit.

To Brogaard and Salerno there are two alternatives.

First alternative: The antecedent of (B7) is not considered a relevant possibility at all.
Second alternative: Context shifts such as to make the (B7) antecedent possibility not only relevant but *as close* as the (A3) antecedent possibility.

However, there are conversational situations in which both alternatives are excluded. Imagine two physicists discussing:

Dialogue1
Physicist I: 'If this match had been struck, it would have lit.'
Physicist II: 'But doesn't it follow that if the match had been struck and there had been a widespread violation of natural laws guiding the behaviour of valence electrons, the match would have lit?'

[20] Cf. David Lewis 'Score–keeping in a language game', *Journal of Philosophical Logic*, 8 (1979), 339–59 (p. 43).
[21] Lewis, *Counterfactuals*, p. 66.
[22] Lewis, *Counterfactuals*, p. 67.

Physicist I: ‘No. I see the possibility of a wide-spread violation of natural laws. But it is too outlandish to draw this conclusion.’

In this situation, the antecedent of (B7) is acknowledged as an accessible possibility. It seems preposterous to assume that Physicist II relies on the contextual inaccessibility of a widespread violation of natural laws. It would be as preposterous to insist against Physicist I that *in the course of the dialogue* a gross context shift occurs, which makes the (A3) antecedent possibility as close as the (B7) antecedent possibility.

Brogaard and Salerno might accommodate the Dialogue1 by distinguishing two senses of accessibility:

Accessible1: A possibility is conceivable.
Accessible2: A possibility is counted among the contextually relevant ones.

Physicist I may grant that the (B7) antecedent possibility is conceivable but refuse to count it among relevant worlds. However, firstly Physicist I does not have to be understood in this way. Secondly, once one accepts a similarity–ordering which is more fine-grained than the distinction of being closest and being inaccessible, one simply cannot exclude an interpretation of (Wet Match*) according to which the (B7) antecedent is relevant but less similar than the (A3) antecedent possibility. Since orthodoxy conceptually makes room for the threefold distinction closest–less close albeit accessible–inaccessible, it seems arbitrary to preclude that there are contexts in which the second member of the distinction becomes relevant.

In short, according to orthodoxy, there may well be everyday instances of (Wet Match*) of the following sort:
(i) Both (A3, B7) antecedent possibilities are accessible.
(ii) The (A3) antecedent possibility is closer, i.e. more similar than the (B7) antecedent possibility.
(iii) No context shift.
Still (Wet Match*) intuitively fails as an inference. So within an orthodox Lewisian framework, Brogaard’s and Salerno’s analysis is not successful.

Coming to a different logical issue, Brogaard and Salerno also discuss a counterfactual version of McGee’s counterexample to modus ponens:

(Reagan)
[B8] `If a Republican were to win, then if Reagan were not to win, Anderson would win.
[B9] A Republican will win.
[B10] So if Reagan were not wo win, Anderson would win.´(p. 44)

Brogaard and Salerno are bound to defend the validity of modus ponens. The orthodox view has no difficulties with doing so. One simplified orthodox treatment goes like this: For (B8) to be true, we must consider the world where Reagan does not win closest to the closest world where a Republican wins. The closest world in which a Republican wins is the actual one, as claimed by (B9). In order for (B10) to be true, the next world to the actual world where Reagan does not win must be a world where Anderson wins. Since according to (B8) & (B9), the closest world where Reagan does not win is a world where Anderson wins, (B10) is true provided (B8) and (B9) are. This reasoning generalizes to all cases in which the antecedent possibilities are accessible. The argument is valid as no context shift occurs and the similarity ordering is stable. When the possibility of a Republican winning is inaccessible, (B8) is vacuously true but (B9) cannot be true. When the possibility of Reagan not winning is inaccessible, (B8) and (B10) are vacuously true. Modus ponens is preserved.

Now (B10) is intuitively false. So one better had not to accept (B8) and (B9). Fortunately, one is not obliged to hold (B8) true. In the actual world a Republican (Reagan)

wins. The next world to the actual world where Reagan does not win does not have to be a world where Anderson wins. Rather it might 'mimic the outcome of the polls'.

Contrary to this intuitive treatment, Brogaard and Salerno argue that the contextual condition held fixed for (B8) and (B9) is that a Republican wins:

'To evaluate the first premise we go to the nearest worlds in which a Republican wins. As Reagan was first in the polls and indeed won, the actual world is the only closest world in which a Republican wins. But the Reagan–loser worlds closest to the actual world are still worlds in which a Republican wins. That is because the context holds fixed that a Republican wins. And so, the Reagan–loser world closest to the actual world must be an Anderson–winner world. The first premise comes out true. Further, since a Republican actually won, the second premise is true as well.'(p. 44)

According to Brogaard and Salerno, we hold fixed as part of the context that a Republican wins. Worlds where no Republican wins are inaccessible. If we hold context fixed in this way for (B8) and (B9), there is no reason why not to hold it fixed until (B10). Then (B10) has to come out true. We get the same result provided all antecedent possibilities of (B8)–(B10) are accessible *and* a world where Reagan does not win is inaccessible. In all these cases, we have to accept (Reagan). Accepting (Reagan) is a very counterintuitive result. To avoid it, Brogaard and Salerno add:

'However, if we allow the context to shift from a range of worlds in which a Republican won to a range of worlds in which the outcome of the election mimics the outcome of the polls, then the conclusion comes out false.'(pp. 44–45)

Our intuitive rejection of (Reagan) is explained by a shift between (B9) and (B10). But why should there be such a shift? The dichotomy of being closest and being inaccessible does not leave much space for shifting manoeuvre. The only promising candidate is this: a world where Reagan does not win is promoted from being inaccessible to being accessible. The other way round (B10) would be vacuously true. But it seems ad hoc to insist that, when evaluating (B8), one finds the world where Reagan loses inaccessible, only to find it accessible when assessing (B10). I do not see how Brogaard and Salerno could make further room for 'mimicking the outcome of the polls'. Lewisian orthodoxy can. This is a disastrous result for Brogaard and Salerno.

Are there any reasons to quit the orthodox framework in favour of Brogaard's and Salerno's solution? It might be argued that it is better to have a unified account in which CSH come out valid. But Adams pairs like (A9)/(A10) and (A11)/(A12) show that there are good reasons to treat ordinary indicative conditionals, material conditionals, and counterfactuals differently. There is no reason to assume that they should be subject to the same logics. So being orthodox, we should not wonder if CSH come out valid at least for the material conditional but not for counterfactuals.

Summarizing, I have defended the orthodox account against a revision which promises to preserve certain logical principles invalidated by the standard account. There is no sufficient reason to preserve these principles against our intuitions.

2.2. Challenging Truth–Conditions: Gibbard Cases

I have presented a truth-conditional account of counterfactuals. The idea that counterfactuals have truth–conditions has come under pressure from several sides. I shall discuss one exemplary argument. The argument revolves around transferring a famous and beautiful counterexample to truth-conditional accounts of *indicative* conditionals to counterfactuals. Here is the counterexample:

'Sly Pete and Mr. Stone are playing poker on a Mississippi riverboat. It is now up to Pete to call or fold. My henchman Zack sees Stone's hand, which is quite good, and signals its content to Pete. My henchman Jack sees both hands, and sees that Pete's hand is rather low, so that Stone's is the winning hand. At this point, the room is cleared. A few minutes later, Zack slips me a note which says 'If Pete called, he won,' and Jack slips me a note which says 'If Pete called, he lost.' I know that these notes both come from my trusted henchmen, but do not know which of them sent which note. I conclude that Pete folded.'[23]

Consider the two conditionals

(C1) If Pete called, he won
(C2) If Pete called, he lost

Both are uttered in the same context. Both Jack and Zack seem perfectly right to utter these conditionals from their point of view. Moreover, Gibbard observes that none of the speakers is 'mistaken about something germane'.[24] Yet the two conditionals seem mutually inconsistent: *if Pete called, he lost* entails *it is not the case that, if Pete called, he won*, and vice versa. Gibbard wonders how it can be that two speakers can utter mutually inconsistent conditionals without being mistaken. Moreover, the situation seems symmetrical. At least a symmetrical but less beautiful example can easily be forged. Here is one due to Bennett:[25]

In a drainage system, there are three gates, top gate, left gate and right gate. Precisely if top gate and left gate are open, water will flow through left gate. Precisely if top gate and right gate are open, water will flow through right gate. If top gate is open, either left gate or right gate might be open, but not both. If top gate is closed, left gate and right gate might both be open. Righty knows additionally that right gate is open. She utters 'If top gate was open, water flew through right gate only'. Lefty knows additionally that left gate is open. She utters: 'If top gate was open, water flew through left gate only'. In fact top gate was closed.

I will come to other symmetric examples in a moment.

From his observations, Gibbard draws the lesson that indicative conditionals have no truth–conditions. To him, it cannot be that two speakers assert true but inconsistent sentences in the same context. It cannot be either that at least one of the two sentences is false as that would require at least one of the speakers to be mistaken about something germane. Moreover, as the symmetric cases show, both sentences would have to be true, or both would have to be false; it cannot be that one sentence is false and the other is true.

I shall very briefly survey some potential reactions before proceeding to counterfactuals. One reaction that promises to preserve truth–conditions is an account of assertions of conditionals as *conditional assertions* only in a situation in which A is true. By a conditional 'if A, C' one asserts C in case A. If A is not true, one does not assert anything. Thus, since Jack and Zack did not assert anything, there are no assertions that could conflict. I shall discuss an

[23] Alan Gibbard, 'Two recent theories of conditionals', in *Ifs: Conditionals, Belief, Decision, Chance, and Time*, ed. by William L. Harper, Robert Stalnaker, Glenn Pearce (Dordrecht: Reidel, 1981), pp. 211–47 (p. 231).
[24] Gibbard, p. 231.
[25] Jonathan Bennett, *A Philosophical Guide to Conditionals* (Oxford: Oxford University Press, 2003), p. 256.

account along these lines in section (2.3.1.2.). Here I only point to one problem: assume Jack says 'it is not the case that if Pete called, he won'. Jack seems to have asserted something true regardless of whether Pete did call or not.

Another proposal would be to withdraw to conditions of assertion instead of truth–conditions: we do not ask: when is a conditional true? We ask: when is it adequate to assert it? The proposal requies us to specify conditions of assertability. Such a view clashes with our disposition to call conditionals true and false, respectively. Perhaps such judgements should not be taken too seriously, but they show at least that some revision of our everyday ways of speaking may be unavoidable if truth–conditions are eschewed.

My topic in this book are counterfactuals. Thus, instead of further pursuing ways to deal with Gibbard's challenge for indicative conditionals, I shall proceed to discussing whether the challenge transmits to counterfactuals. Adam Morton and Dorothy Edgington disagree as to whether there are counterfactual Gibbard cases. I shall argue against Morton that there may be such cases, and against Edgington that the standard truth–conditional analysis of counterfactuals can account for them. The general lesson to be drawn is twofold: First, the Gibbard phenomenon should be kept free from theoretical bias. Second, if a truth-conditional analysis succeeds in counterfactual cases, it might as well succeed with regard to indicative Gibbard conditionals.

Trying to get Gibbard's Riverboat example in focus, Jonathan Bennett complains that '[...]the pure signal of [Gibbard's] argument has sometimes been invaded by noise coming over the wall from the subjunctive ['would'] domain.'[26] I shall present an argument to the contrary. Proper attention to counterfactual Gibbard cases helps to purify the signal from theoretical bias. The case of counterfactual Gibbard conditionals should be reopened; notwithstanding the intense debate between Adam Morton and Dorothy Edgington, the crucial issue remains unsettled.

Edgington argues that in some cases, we can proceed from indicative Gibbard conditionals to corresponding counterfactuals. Here is Morton summarizing the example:

'There is a disease *D*, vaccines *A* and *B*, and a side–effect *S*. Neither *A* nor *B* alone completely prevents *D*. If you've had *A* and you go on to get *D* you get *S*; but if you've had *B* and you go on to get *D* you don't get *S*. If you've had both *A* and *B* you don't get *D* and so don't get *S*. There are two observers *X* and *Y* and a patient, Jones. *X* knows hat Jones has had *A* and thus is justified in believing that *if Jones gets D he will get S*. *Y* knows that Jones has had *B* and thus is justified in believing that *if Jones gets D he will not get S*. Each of their beliefs is justified by what they know. They contradict one another, but learning the whole truth will not show that one is right, since the whole truth includes the fact that Jones has had both *A* and *B* and thus will not get *D*[…] Edgington goes on to consider a dead–Jones case. Suppose Jones is run over by a bus before there is any chance of his getting *D*. Then, she argues, *X* can say *if Jones had got D he would have got S* and *Y* can say *if Jones had got D he would not have got S*. As a result, 'at that time the Gibbard phenomenon applies each has adequate reason for his opinion, and the world rules out there being an objectively correct opinion, for it rules out Jones' getting the disease'.'[27]

Edgington builds Gibbard counterfactuals which correspond to indicative Gibbard conditionals:

(C3) If Jones gets D he will get S.
(C4) If Jones gets D he will not get S.
(C5) If Jones had got D he would have got S.
(C6) If Jones had got D, he would not have got S.

[26] Bennett, *Conditionals*, p. 83.
[27] Adam Morton, 'Can Edgington Gibbard Counterfactuals?', *Mind*, 106 (1997), 100–105 (pp. 101–102), quoting Dorothy Edgington, 'On conditionals', *Mind*, 104 (1995), 235–330 (p. 319).

Edgington accepts, Morton rejects that there are Gibbard counterfactuals. Morton's argument builds on the Lewisian standard analysis of counterfactuals. Bennett agrees with Morton, using a similar argument.[28] Edgington replies to Morton by arguing against the standard Lewisian analysis of counterfactuals.[29] Yet a dialectically much stronger point can be made (and then turned against both Morton/Bennett and Edgington): even if the standard analysis is accepted, Morton cannot establish that Edgington is wrong; there might be Gibbard counterfactuals. To make this point, I shall develop my own argument against Morton.

Morton summarizes what he takes to be necessary and sufficient conditions for indicative Gibbard conditionals:

'The facts are symmetrical between them, in that there are equally good reasons for thinking that one is true as that the other is. So one is true iff the other is. Call this *Symmetry*. But they contradict one another: if getting *S* is a consequence for Jones of getting *D* then escaping *S* is not a consequence. Call this *Contradiction*. So the one is true iff the other is not. But these two biconditionals are contradictory. (Note that they can be contradictory even if the sentences they discuss have no truth–value.) So we had better not give any truth–values.'[30]

In order to be Gibbard cases, Edgington's counterfactuals must conform to *Symmetry* and *Contradiction,* says Morton:

'*Symmetry* and *Contradiction* can be produced for the counterfactuals without specifying anything that makes them cease to hold for the indicatives. Thus Edgington has to make two claims about the dead Jones case. First, that the case can be spelled out so that there are no further facts which favour one counterfactual over its contrary which do not also favour one of the indicatives in the live Jones case over its contrary[…] And second that in the situation thus spelled out either of the two counterfactuals is true iff the other is false.'[31]

In Morton's view, the acceptance of Gibbard counterfactuals that correspond to indicative Gibbard conditionals depends on two claims:

Counterfactual Symmetry: there are no further facts which favour one counterfactual over its contrary which do not also favour one of the corresponding indicatives over its contrary.
Contradiction: one of the two counterfactuals is true iff the other is false.

As we will see, however, the acceptance of Gibbard counterfactuals does not commit one to any of these claims.

Morton uses *Counterfactual Symmetry* and *Contradiction* in arguing as follows: purported Gibbard counterfactuals either fail to satisfy *Symmetry* or *Contradiction*.

Consider the *first alternative*: Gibbard counterfactuals fail to satisfy Counterfactual Symmetry. Morton's argument is that there normally are facts which break the symmetry. They are relevant to making one counterfactual false and the other true without doing the same for the indicative counterparts. There are different ways of further fleshing out the situation depicted by Edgington. For instance, it might be that Jones almost missed his appointment to get A, but only extraordinary circumstances could have prevented him from getting B.[32] According to Morton, such circumstances make worlds in which Jones failed to get A closer than worlds in which Jones failed to get B. If Jones had got D, that would be because he had not got A. Yet he would still have got B before. Thus (C5) if Jones had got D, he would have got S. The additional assumptions act as tie–breakers in the case of counterfactuals but not in

[28] Bennett, *Conditionals*, p. 242.
[29] Dorothy Edgington, 'Truth, Objectivity, Counterfactuals and Gibbard', *Mind,* 106 (1997), 107–116 (pp. 112–113). Cf. the critical review of Edgington's argument in Bennett, *Conditionals,* pp. 255–256.
[30] Morton, 'Gibbard', p. 101.
[31] Morton, 'Gibbard', p. 102.
[32] Morton, 'Gibbard', p. 102.

the case of the corresponding indicative conditionals. Then the former violate *Symmetry* while the latter do not. *Counterfactual Symmetry* does not hold. Thus, the additional facts which hold in the scenario disqualify purported counterfactual Gibbard cases.

The *second alternative* is that Contradiction fails: Morton acknowledges that Edgington's example can be fleshed out such as to yield perfectly symmetrical counterfactuals. For instance, Jones might have been administered both A and B before his birth.[33] While in asymmetric cases, Morton takes it for granted that only one counterfactual is true and the other is false, in symmetric cases, he does not take this for granted. In these cases, *Contradiction* becomes crucial. According to Morton, the symmetric cases miss this second condition of Gibbard cases. From 'it is not the case that if Jones had got D, he would have got S', it does not follow that 'if Jones had got D, he would not have got S'. It may simply be that *both are false*.

Coming to my criticism, I shall first address *Contradiction.* I doubt that Contradiction is a necessary condition for Gibbard cases. We have seen Morton noting: '…if getting *S* is a consequence for Jones of getting *D* then escaping *S* is not a consequence. Call this *Contradiction*. So the one is true iff the other is not.' *Contradiction* as Morton has it implies that one of the conditionals is true *iff* the other is not (not (A→C)↔(A→not–C)). As Morton acknowledges, what we need for Gibbard cases is the following: 'if getting S is a consequence for Jones of getting D then escaping S is not a consequence'. But to satisfy this requirement, we do not need *Contradiction*. We need only that the two conditionals cannot both be true; one is true *only if* the other is not:

Weak Contradiction: Not both 'If D, S' and 'if D, not S'

This is exactly the principle employed by Bennett. He calls it

'Conditional Non–Contradiction: not((A→C)&(A→not C))'[34]

And it is plausible that counterfactuals meet this requirement![35] Gibbard cases only have to satisfy the weaker principle.[36] Thus, violating *Contradiction* does not disqualify a pair of counterfactuals from forming a Gibbard case.

Having discussed *Contradiction*, I shall now discuss *Counterfactual Symmetry*. Counterfactual Symmetry is not sufficient to bring home Morton's point as it does not bear on perfectly symmetric cases. Still it is worth discussing. By Morton's lights there is a key difference to indicative Gibbard conditionals. Fleshing out the story will very often lead to one of the counterfactuals being true, the other false. We have to ask which is the slightest departure from actuality that leads to Jones getting the disease. Morton's examples of fleshing out the story make one candidate vivid: Jones has only got *one* of the vaccines. He almost has missed getting A. So we might reckon a situation in which he has not got A closer than a situation in which he has not got B. Given some auxiliary assumptions, Y's counterfactual is true but X's is not. In a genuine Gibbard situation, Morton argues, there is no comparable way of fleshing out the story such that one of the *indicative* conditionals comes true but the other does not. Thus, Morton gives reasons to assume that we often have Gibbard lookalikes in which both opponents are equally entitled to their conditionals but these conditionals are false. The counterfactuals violate Symmetry.

---

[33] Morton, 'Gibbard', p. 103

[34] Bennett, *Conditionals*, p. 84.

[35] Thus, Edgington's and Morton's discussion whether conditional excluded middle holds seems beside the point (Edgington, 'Truth', p. 114).

[36] But what if someone insists on the stronger principle? I reply with a question: Why does one need the stronger principle in contrast to the weaker one, why is it essential to the Gibbard phenomenon?

In reply, in order to hear the pure voice of the Gibbard phenomenon, we should get rid of theoretical noise. Adopt for a moment a stance of theoretical innocence: The standard analysis of counterfactuals has not yet been established. Considering the symmetric dead Jones scenario, would anyone unaffected by theory surmise where to look for the additional facts that make X's and Y's counterfactuals false? I imagine that she would be in the very same position we are when considering the case of Jones being alive: Untainted by theory, what we know is that X's and Y's epistemic standing is symmetric. Both protagonists are epistemically blameless; they are not mistaken about anything germane. And we know that they are ignorant about some matter that is relevant to their conditionals: Jones has got both vaccines. There might be further matters which are relevant. It is open what consequence for the truth–value of the conditionals should be drawn. *Weak Contradiction* indicates that they cannot both be true. In the case of indicative conditionals, it is a mystery how one or both could be false, and which fact could make them false.[37] Yet the only difference to the case of counterfactuals is that a well-established standard analysis provides an answer how the counterfactuals can be false. Without the standard analysis, no one would have thought of this answer. We do not (yet) have an equally well-established answer for indicative conditionals. But nothing precludes that such an answer might be established, and that it tells us where to look for the facts that make indicative Gibbard conditionals false.

One might feel hesitant to endorse this diagnosis as long as no indication has been given where to look for additional facts that could make indicative Gibbard conditionals true or false. Thus, I mention some examples of a truth-conditional analysis: one is the classical horseshoe analysis of the indicative conditional.[38] Considered as material conditionals, both X's and Y's verdicts come true simply in virtue of their antecedents being false. X and Y are ignorant but not mistaken about this fact. As in the counterfactual case, there is an unknown additional fact that decides the truth–value of the conditionals. There might be other candidates for such facts. I do not want to advocate the horseshoe analysis; after all, it violates *Weak Contradiction*.

An alternative example is Stalnaker's view: the upshot is that the truth–value of the indicative Gibbard cases is evaluated in the same way as the truth–value of the counterfactual ones: We ask which world is more similar, needs a smaller departure from the actual world, a D&S–world or a D&not–S–world.[39] Under symmetry all conditionals are indeterminate as the selection function does not select a unique closest world.[40] Still conditionals have truth–conditions. This completes the parallel between indicative and counterfactual conditionals. In both cases, there is an intuitive puzzle how to decide between conflicting conditionals; in both cases, one person may be perfectly entitled to endorse one of the conflicting conditionals, while another person may be perfectly entitled to endorse the other. The only difference is that in the

[37] DeRose alleges that in the Riverboat example, the only fact eligible for making one conditional false is Pete having the lower hand (Keith DeRose, 'The Conditionals of Deliberation', *Mind*, 119 (2010), 1–42 (p. 24)). He rules out this candidate, though, as one may hold onto the conditional (C1) upon learning that Pete probably has the lower hand. I cannot do justice to this argument here but only outline a strategy to begin with. As a first step, offer an explanation why one may *falsely* deem Pete's losing hand a fact that makes the conditional false: normally, learning about Pete's hand leads to *retracting* the conditional; yet this is not due to the latter's falsity but rather because it has become pointless, the antecedent situation being too improbable; in contrast, one may also choose to *give* it a point ('To be sure, given the distribution of cards, it is improbable that Pete will win; yet he will never call unless he has the winning hand; so *if* he calls, he will win.'). In a second step, the analogy with counterfactuals is exploited to show that the facts making a conditional false might be far from obvious; instead of a simple, well–confined matter of fact (Pete's losing hand), we might have to explore a complicated configuration of facts playing some complex semantical role (forming the world closest to actuality or the like). By the way, the very asymmetry exploited in DeRose's argument (one conditional being less well founded) sheds doubt on the Riverboat example as a paragon of the Gibbard intuition, in contrast to Edgington's more symmetric one (cf. William G. Lycan, *Real Conditionals* (Oxford: Clarendon Press, 2005), p. 169).

[38] Cf. Lycan, *Real Conditionals*, p. 171, on Lewis's view.

[39] Cf. Lycan, *Real Conditionals*, pp. 171–172.

[40] Cf. Bennett, *Conditionals*, p. 183.

counterfactual case, there is a well-established truth-conditional analysis, in the indicative case, there is none.

While these considerations may be sufficient to shed doubts on Morton's argument, the main intuitive difference remains: at least one in a pair of purported Gibbard counterfactuals is false. In contrast, Gibbard says on the indicative conditionals:

'[…]one sincerely asserts something false only when one is mistaken about something germane. […] Neither [of the protagonists] has any relevant false beliefs, and indeed both may well suspect the whole relevant truth.'[41]

The requirement that both opponents may *suspect the whole relevant truth* must not be read too strongly. If X and Y had reasons to suspect that Jones has had both A and B, their entitlement to their conditionals would be in danger. If Zack and Jack suspected that Sly Pete knew his opponent's hand and had the losing hand himself, they would refrain from asserting their conditionals. Jack might uphold his conditional if pressed, but there would be no point in asserting it. Thus, what remains is the requirement that the protagonists not be *mistaken* about relevant facts.

Bennett adds the requirement that both opponents are fully entitled to their conditionals: 'I stress fully entitled; these acceptances are intellectually perfect.'[42] This requirement must not be read too strongly either. 'Intellectually perfect' cannot mean that the positions of X and Y could not be epistemically superior with respect to assessing their conditionals, be they indicative or subjunctive. Surely it would in a way be better if they knew that Jones has had A *and* B. But this does not impair their entitlement. What remains is the requirement that both speakers must be entitled to their conditionals.

In sum, we have two requirements for Gibbard cases: none of the two speakers must be mistaken about anything relevant to the truth of their utterances. Both must be justified in their utterances.

Given these requirements, the question becomes: are X and Y mistaken about anything germane in uttering their counterfactuals, or is there anything that impairs their justification? We may first ask: do they have any relevant false beliefs? Of course, in the Lewisian standard analysis, at least one of the counterfactuals is false. So at least one of X and Y has relevant false beliefs, namely the counterfactuals themselves. Yet since it is open whether indicative Gibbard conditionals are false just as their counterfactual versions, it would beg the question to use the falsity of the counterfactuals as a reason against their being Gibbard cases. Bennett emphasizes: 'Gibbard must mean that one sincerely asserts something false only if one is mistaken about some relevant *nonconditional* matter of fact.'[43] X and Y do not seem mistaken about some relevant non-conditional matter of fact.

Morton insists that there are further facts which may decide the truth–value of the counterfactuals but do not bear on the truth–value of the indicatives. Perhaps X and Y are mistaken about them, or their ignorance undercuts their justification. To assess this hypothesis, we may ask: should X and Y scrutinize these facts before endorsing their counterfactuals? The answer is no. Morton fails to explain an important finding of Edgington's: in her example, X's and Y's transition from the indicative to the subjunctive conditionals seems perfectly smooth. There is no additional condition for this transition besides Jones being run over by a bus; nor is there anything patently irrational or illegitimate about it.

We can account for the smooth transition within the standard Lewisian analysis of counterfactuals. According to Lewis's standard criteria, some minimal divergence from

[41] Gibbard, 'Two Theories of Conditionals', p. 231.
[42] Bennett, *Conditionals*, p. 83.
[43] Bennett, *Conditionals*, p. 84.

actuality brings about Jones getting D; under determinism, it amounts to a small miracle.[44] From our better-informed perspective it seems difficult to say whether Jones would have got S had he got D. Any small divergence that brings about Jones getting D must interfere with A and B having their normal effect. We cannot tell in how far this leads to Jones getting S or not. However, the smooth transition in Edgington's example indicates that in contrast to us, *X* and *Y* do not have to bother. Why not? None has reason to assume that Jones has got both vaccines. In order to account for the smooth transition, we must regard X and Y as perfectly vindicated in neglecting this possibility, just as they justifiably neglect it when endorsing the indicative conditionals.

It is important to appreciate Edgington's choice of example. She has Jones run over by a bus in order to make getting D a suitable counterfactual scenario. X and Y know that John has been run over by a bus. Jones having had both vaccines would also be sufficient for ensuring D to be contrary–to–fact. Yet if X and Y knew that Jones has had both vaccines, their reasons to accept their counterfactuals would not be sufficient, just as they would not be sufficient to accept the indicative variants. The smooth transition to the counterfactuals can be explained as follows: to get a justified take on the counterfactual case, X and Y only need to consider the minimal departure from actuality undoing John's having run over by a bus and making him contract D. Since they have no reason to suspect that John has got both drugs, they do not have to take into account a departure that undoes this fact.

In sum, even granted that one of two Gibbard counterfactuals has to be false, nothing we are told about indicative Gibbard cases excludes that the same goes for them. As a consequence, assuming X's and Y's counterfactuals are false, they testify against Gibbard's claim that 'one sincerely asserts something false only when one is mistaken about something germane'. To be mistaken about something germane would require to be mistaken about some fact which one should know or suspect before incurring a commitment to a conditional. X and Y do not seem mistaken in this way, and still one of the counterfactuals they accept may be false. The same criticism applies to Bennett's claim ('→' stands for the indicative conditional):

> '[...]we saw how one person can be perfectly entitled to accept A→C and to accept A→not C; but this cannot happen with A>>C and A>>¬C. [...]Never can both be true or fully acceptable, as conflicting indicatives in a Gibbardian stand–off can be.'[45]

Indeed, it might be impossible for both counterfactuals to be true; but each might nevertheless be endorsed in isolation with perfect entitlement, be *fully acceptable*, just as the conflicting indicatives. This counts against the alleged difference between the indicative conditionals and the counterfactuals: not being mistaken about something germane and being fully entitled is compatible with indicative Gibbard conditionals being false, just as at least one of the corresponding counterfactuals is.

I ponder a potential reply: perhaps the crucial issue is not what X and Y *should* know but simply what they *do not* know. The question becomes: Is it irreconcilable with the Gibbard phenomenon that some protagonist in a Gibbard case is ignorant about non-conditional facts bearing on her conditional? The answer must be no. For any indicative Gibbard cases, there are relevant facts which the protagonists do not know. If they were to take into account these facts, they would not utter their conditionals.

Taking stock, the Gibbard intuition should be kept free from theoretical presuppositions. And it should be ensured that the relevant symmetries of epistemic position obtain. Instead of Morton's *Symmetry* and *Contradiction*, Gibbard cases are subject to the following requirements:

---

[44] Lewis, 'Counterfactual Dependence', pp. 47–48, 59.
[45] Bennett, *Conditionals*, p. 242, notation adapted.

Epistemic Symmetry:
X and Y are in a symmetric evidential situation such that no one *is mistaken* about anything germane.
X is entitled to accept a conditional 'if A, C' (respectively the corresponding counterfactual, if applicable).
Y is equally well entitled to accept a conditional 'if A, not–C' (respectively the corresponding counterfactual, if applicable).

Weak Contradiction:
'If A, C' and 'if A, not–C' cannot both be true; the corresponding counterfactuals cannot be either.

If Gibbard scenarios are understood in this way, they are reconcilable with a truth-conditional analysis of counterfactuals. This gives rise to an important dialectical point: As far as Gibbard cases are concerned, a truth-conditional analysis of indicative conditionals may fare as badly or as well as the truth-conditional analysis of counterfactuals.

In my view, considering counterfactuals may be the best way to argue for *Epistemic Symmetry* and *Weak Contradiction* as requirements for genuine Gibbard cases. Since there is a well-established truth-conditional analysis, it becomes obvious that both protagonists can be perfectly justified to endorse their conditionals although they are false. Where there is no such analysis, this tends to be obscured. So contrary to Bennett's complaint, the news coming over the wall from the subjunctive domain help to overcome a theory–bias which distorts the Gibbard intuition. It is ironic that Edgington's discovery of Gibbard counterfactuals can be turned against the lesson 'no truth–conditions' she herself draws from Gibbard's examples.

2.3. Probabilities
2.3.1. Proposals in the Literature
2.3.1.1. Schulz's Arbitrariness Account

One big issue for the theory of counterfactuals is the role of probability. I shall discuss a range of accounts which approach this topic. One subcase of probabilistic counterfactuals are counterfactuals about lotteries. Lotteries pose great challenges to epistemology. Yet intuitions on lottery counterfactuals are also notoriously puzzling. In particular, they seem in tension with the standard account of counterfactuals.

I shall start with discussing an approach recently developed by Moritz Schulz.[46] To dissolve problems with lottery counterfactuals, Schulz comes up with a new semantics for counterfactuals. The new semantics dissolves the tension discerned in the standard account and nevertheless preserves many of the attractive features of the latter.

In discussing Schulz's approach, I first briefly summarize the problem and the solution à la Schulz. Second, I discuss some uncertainties about Schulz's presentation of the problem and propose an amendment. Instead of a principle based on the theoretical notion of subjunctive credence, one should rather use more basic linguistic evidence. Third, I present counterevidence to Schulz's solution: it does not square well with the embedding behaviour of counterfactuals. Embedding evidence is contestable. Yet the evidence I provide just testifies to an independently motivated aspect of counterfactual reasoning: it tracks a natural distinction between what does and what does not follow from the antecedent plus background facts.

Schulz addresses a problem of lottery counterfactuals. In the simplified standard account, a counterfactual is true iff all relevant antecedent worlds are consequent worlds. Relevance in turn is usually spelled out in terms of closeness or similarity to an evaluation world. Take a fair lottery with a great many tickets. Anna does not buy a ticket. What about the following counterfactual?

(D1) #If Anna had bought a lottery ticket, she would have lost.

It seems somewhat inappropriate to utter (D1). Perhaps (D1) can be felicitously uttered when one wants to stress how unreasonable it would have been for Anna to buy a ticket. But in many contexts, (D1) seems infelicitous. In the standard (Lewis–Stalnaker) account of counterfactuals, this is explained as follows. (D1) is obviously not true, and so one had better not assert it. For (D1) to be true, all the contextually relevant worlds where Anna buys a ticket have to be worlds where she loses. Contextual relevance is standardly spelled out in terms of closeness or similarity to the world from which the conditional is evaluated (usually the actual one). The relevant worlds are usually partitioned as follows (though this does not directly follow from the standard account): Anna has precisely one particular ticket t and ticket 1 wins, Anna has ticket t and ticket 2 wins... Hence besides all the relevant worlds where she loses there is one relevant world where she wins. (D1) is not true.

To bring out the puzzle, Schulz asks what degree of credence one should place in (D1). He contends that credence should abide by an intuitive constraint. Provided there is no inadmissible evidence, your rational credence *Cr(...)* in a counterfactual $P >> Q$ should equal the objective probability *Ch(...)* of the consequent given the antecedent (measured by the proportion of *Q*–worlds among the relevant P–worlds):

(Credence) $Cr(P>>Q|Ch(Q|P)=x)=x$.

[46] Moritz Schulz, 'Counterfactuals and Arbitrariness', *Mind*, 123 (2014), 1021–1055.

Applying this constraint yields high credence in (D1). The chance of Anna's ticket losing given she buys one is high. Yet applying the standard account, one knows that (D1) is false. How can one place high credence in what one knows for certain to be false, given the standard account?

Schulz's solution is to introduce a new truth–condition for counterfactuals:

A counterfactual *P>>Q* is true iff Q at some P–world which is *arbitrarily selected* from the relevant, e.g. most similar P–worlds.

One gets precisely the desired epistemic profile: (D1) is true provided the arbitrarily selected world where Anna buys a ticket is one where she loses. The ratio of losing and winning worlds which are available for the arbitrary selection precisely corresponds to the objective probability of losing as contrasted to winning in the lottery. Since the probability of Anna losing is very high, so should one's credence in (D1) be. However, one cannot know (D1) to be true, just as Anna cannot know in advance that her ticket will lose when she buys it. Hence it seems epistemically irresponsible to assert (D1). As a result, we get an explanation why it seems inappropriate to utter (D1).

The evidence revisited

While I agree that there is a puzzle around, I feel uncertain about the manner it is presented. In particular, I doubt that (Credence) is intuitive upon closer inspection. (Credence) is motivated by a parallel to Lewis's

(Principal Principle): $Cr(P|Ch(P|E)=x)=x$.

Your credence in P should equal the chance of P given your total evidence *E*, provided the latter is admissible. It is notoriously difficult to characterize admissibility, but, as a first stab, your total evidence should only inform you about the chance of P, not independently about whether P is the case. (Credence) is not as immediately compelling as (Principal Principle).

The problem of this parallel can be illustrated by an example where probabilities are more volatile than in the lottery case:

(D2) There has been a storm in the North Sea on May 15.

Assume you have good reasons to think that, as distinguished from a lottery, the probability of a storm in the North Sea constantly changes throughout history until May 15. Moreover, all your evidence with respect to (D2) tells you that, on May 9, the chance of a storm was 30%. Then your credence in (D2) should be 0.3. Now assume you get additional information that there has been no storm. This is clearly relevant evidence which should change your credence in (D2) to close to 0. Conditional credence behaves similarly. Consider the conditional probability of a storm in the North Sea on May 15 given there was a storm in the Irish Sea ten hours before. Assume all you know is that it tends to be very volatile, but on May 9, it was 60%. Then your credence in a storm in the North Sea on May 15 *conditional* on a storm in the Irish sea ten hours before should be 0.6. Now additionally assume that, May 15 having passed, you know that there was no storm either in the North Sea or in the Irish Sea. Then your conditional credence will presumably be undefined.

Consider

(D3) There would have been a storm in the North Sea on May 15 if there had been a storm in the Irish Sea ten hours before.

What does (Credence) tell you? Your only relevant information is that at some time the conditional probability of there being a storm in the North Sea given a storm in the Irish sea was 0.6. According to (Credence), your credence in (D3) should be 0.6. But once you know that there has been no storm, it is not obvious that your credence should be guided by the probability on May 9 just because you happen to know it. The motive for doubt is that the conditional probability is so volatile. In the lottery cases, this problem of (Credence) does not become manifest as probabilities are not volatile.

There are several ways to defend (Credence) against these concerns. One may deny that the conditional probability on May 9 is the probability that figures in (Credence). But then the latter probability becomes mysterious. One may deny that (Credence) is applicable when you know there has been no storm either in the North Sea or the Irish Sea. For the present conditional probability of a storm is undefined. But then the question becomes why the same does not hold for (D1) as you know that the antecedent is contrary–to–fact. Another way to defend (Credence) would be to exploit the admissibility constraint from (Principal Principle). Perhaps evidence that the conditional probability is volatile is inadmissible. But firstly, I feel uneasy about (Credence) even when you know the conditional probability at some point long before antecedent time but are completely in the dark as to whether this probability is stable. Secondly, we lack an explanation why the admissibility constraint behaves so differently in the case of counterfactuals and actual events.

I do not want to deny that something like (Credence) may be true, nor that it yields the intuitive results for (D1). I just want to highlight disanalogies between (Credence) and (Principal Principle). These disanalogies are significant. (Principal Principle) guides reasoning where one only knows the chances of an event and not whether it actually will occur or has occurred. In contrast, counterfactual reasoning (at least of the 'had'–'would'–sort) is mostly used to reason about events one knows not to have occurred. These disanalogies should prevent one from simply adopting (Credence) as a general principle.[47]

Instead of (Credence), I suggest to make do with linguistic intuitions, which have a less problematic standing. The following seems intuitively acceptable:

(D4) If Anna had bought a lottery ticket, she would probably have lost.

But at least judging from its surface form, in asserting (D4), one seems to claim that (D1) is probably true. So one seems to accept that something one knows to be false is probably true. The result is very close to Schulz's original puzzle, but rests only on elementary linguistic data. One gets an expression which captures some of the ideas behind (Credence) if one inserts one particular probability. Assume there are 10,000 tickets in the lottery. Then intuitively,

(D5) If Anna had bought a ticket, it is 99.99 percent probable that she would have lost.

Instead of putting the evidence in terms of credence 0.9999 in (D1), one may put it in terms of our willingness to accept (D5).

Even stronger evidence in favour of Schulz's account can be derived from the more cautious

(D6) If Anny had bought a lottery ticket, she would perhaps have lost.

Taken at face value, the epistemic modal adverb 'perhaps' here seems to be used to express the epistemic possibility that (D1) is true. This epistemic possibility is ruled out by the standard

[47] Sarah Moss, 'Subjunctive Credences and Semantic Humility', *Philosophy and Phenomenological Research*, 87 (2013), 251–78, avoids this problem by exclusively focusing on future–directed subjunctive conditionals.

account (what we know about the lottery situation) but ensured by Schulz's account. Unless there is a convincing rival explanation, (D4) and (D6) strongly support Schulz's account.

I have argued that the linguistic intuitions mentioned have a less problematic standing than (Credence). However, I admit that they come with uncertainties of their own that would be avoided by basing the argument on (Credence). For instance, one may doubt that, in accepting (D4), one accepts that (D1) is probable. The same for (D6). Indeed I will come up with a somewhat diverging reading. Still it seems to me that (Credence) is not compelling enough as a general principle to bear the weight of Schulz's argument.

Counterevidence to Schulz's account

I shall now discuss some range of data which seem to speak against Schulz's account. Schulz's solution owes its special charm to the combination of definite truth or falsity with unknowability: a lottery counterfactual like (D1) is definitely true/false albeit principally unknowable. Now we can usually play through the truth of some statement *hypothetically* even when we are not in a position to know it. In this vein, we may try to hypothetically consider what follows from the truth of (D1). Hypothetical reasoning of this sort is expressed by 'assume', 'suppose', 'under the hypothesis' and the like. Unfortunately, these expressions interact in a very intricate way with conditionals; we get enmeshed in the notorious embedding problem. I am well aware that using evidence from embedding, of which we already have seen an example in Brogaard and Salerno's discussion of (Reagan), is deeply problematic. I do not want to incur a commitment to the general possibility of embedding, or one particular analysis of embedded conditionals. But I do not think either that embedding intuitions can be simply dismissed.

I shall present some particular examples where I have rather clear intuitions. These intuitions are in tension with Schulz's account. Someone who wants to defend the latter faces the challenge to explain their precise profile. Simply pointing to other problems with embedding is not sufficient as an explanation. Moreover, I shall argue that the best explanation invokes some independently motivated picture of the role of counterfactual reasoning. So even if one is reluctant about embedding, this picture motivates general doubts about Schulz's solution.

There is a further reason why one should not easily dismiss the embedding evidence I am going to present. Implicit embedding is almost ubiquitous in philosophical reasoning. I implicitly used it in presenting (D1) embedded in a *hypothetical* lottery scenario where some fictional person Anna does not buy a ticket. The cases to come are different only in that they focus on the embedding.

I observe that it often seems quite natural to embed certain counterfactuals into hypothetical or suppositional reasoning. Especially amenable to such an embedding are counterfactuals which may well be true for all we know, even given Schulz's account. I shall consider two embedding examples. Both elicit significantly different intuitions, contrary to what one should expect given Schulz's account. The first example is a dispute between Galilei and an Aristotelean physicist. Both disagree about free fall in a vacuum. Galilei may say:

'Consider two bodies of different mass in a vacuum and without any disturbing influence. Perhaps they would fall at the same speed, perhaps they would fall at different speeds. Assume the following:

(D7) If two bodies of different mass had been dropped, they would have fallen at different speeds.

Given this assumption, the Aristotelean theory may be right. Given the opposite assumption, it is false.'

The ruminations of my imagined Galilei seem perfectly in order. He considers two different hypotheses viz. epistemic possibilities of what happens in all relevant antecedent situations.

I shall now compare the Galilei case to a lottery case. I take it for granted that, in a double slit experiment, it is perfectly indeterminate whether an electron passes through one slit A or the other slit B. Nothing hinges on this particular example. If there are perfectly indeterministic processes (which should not be precluded for semantic reasons), any of them would do as an example. Consider two physicists faced with a concrete double slit experiment. One of them says:

'Consider the experimental setting within a longer time period Δt. As you can see, no electron has passed. But what if things had been otherwise? Note that I do not want you to consider what would have happened if an electron had passed through slit A. I want you to consider just what would have happened if an electron had passed through one of the two slits.[48] Assume the following:

(D8) #If an electron had passed through one of the two slits within Δt, it would have passed through A.'

I take it that (D8) is odd in the context of hypothetical reasoning. One is tempted to reply: 'I cannot assume this! It is indeterminate which slit the electron would have passed!'

Here is the most plausible explanation: one can only *assume* ('assume the following') what one takes to be a genuine epistemic possibility. In an appropriate context, we are willing to consider very far-fetched possibilities. Yet we cannot make room for the assumption that (D8) is true. This is evidence that, under the common assumption that the path of the electron is genuinely indeterminate, there is no epistemic possibility that (D8) is true. But in Schulz's theory, there is a salient epistemic possibility that (D8) is true. One of the worlds to be arbitrarily selected is a world where the electron passes through A. So the best explanation of why (D8) is odd conflicts with Schulz's theory.[49]

One may object that the uneasiness about (D8) is not semantic but rather pragmatic in nature: for an epistemic possibility to be taken seriously as a hypothesis, what is taken to be epistemically possible should not be unknowable for principled reasons. Since one can never know (D8), (D1), and the like for principled reasons, it does not make sense to consider them as hypotheses. But it may be interesting to consider epistemic possibilities even if what is possible is unknowable in principle. A more serious *ad hoc*–concern is why the physicist would want to consider (D8).[50] Perhaps she uses intuitive counterfactuals to assess some Everettian many worlds–approach to quantum physics.

---

[48] There is a temptation to resolve embedded conditionals by non–literal interpretations. In problematic cases, the request to *assume that P had been the case if Q would have been* may be taken non–literally as a request to reason what would have been the case if P *and Q* had been the case. Where there is such a danger, it seems natural to explicitly rule out the non–literal interpretation.

[49] What if the assumption is interpreted as a metaphysical possibility? Then we get a nested counterfactual like

(B8) If a Republican were to win, then if Reagan were not to win, Anderson would win.

In Schulz's account, such an interpretation again would face difficulties explaining the difference between (D7) and (D8). In the standard account, we get an explanation: there might be a closest metaphysically possible world where (D8) is true, but that world would be one with very different natural laws and therefore irrelevant to the double slit experiment.

[50] Why did I use the physicist instead of a lottery case? In the latter, it will be very difficult to preclude the reading 'If she had bought a ticket *and lost*…' from influencing intuitions due to its pragmatic significance.

One may deny that Schulz is committed to the truth and falsity of (D8). Yet there seems to be such a commitment provided the relevant antecedent worlds are divided up in worlds in which the electron goes through slit A and worlds in which it goes through slit B. One may deny that they are divided in this way, but this seems unmotivated unless one brings up physical reasons for dividing up the relevant worlds differently, which rather invalidate the concrete example than the overall objection.

I shall not further discuss concerns about the embedding evidence, arguing instead that the difference between the Galilei and the double slit case sits well with some platitudes about the role of counterfactuals: in a counterfactual, one makes a supposition that one normally takes to be contrary–to–fact as the antecedent sentence is not true. Then one considers what follows from this supposition together with certain background facts (those 'cotenable' with the antecedent). The standard analysis is one way of spelling out what the background facts are. In the Galilei case, it follows from the antecedent etc. whether the two bodies would or would not fall at the same speed. In the double slit case, it does not follow whether the electron passes one slit or the other. For (D8) to be true, this would have to follow.[51] Counterfactual reasoning tracks a joint-carving distinction which is deeply rooted in our world view, the distinction between outcomes which follow and outcomes which do not follow from the antecedent. Schulz's solution blurs this distinction. The arbitrary selection process fixes an outcome where the antecedent plus background facts does not fix it.

In using the notion of counterfactual entailment (what 'follows' from the antecedent), I am well aware that this notion is problematic, in particular when we are in an indeterministic setting. The problems are not fully dissolved by the standard account. But the standard truth–condition that all relevant antecedent worlds have to be consequent worlds comes reasonably close to a notion of entailment that is characterised by validity in all possible worlds. My alternative to Schulz's account to come will weaken this condition for some counterfactuals (those read non-maximally). But it preserves the idea of entailment. The idea is that counterfactual entailment (all relevant antecedent worlds are consequent worlds) is approximated (the exceptions among the antecedent worlds do not matter).

Before coming to my own proposal, I shall discuss a further solution to the problems with lottery conditionals.

[51] The relevant notion of entailment should allow for indeterminism. Perhaps it can be spelled out in probabilistic terms.

2.3.1.2. Barnett's Suppositional Account

I shall now turn to a second proposal that promises to account for probabilistic counterfactuals.

In a series of articles, David Barnett has developed a highly original general theory of conditionals.[52] The grand aim is to reconcile two main rivals: a suppositional and a truth-conditional view[53]. When he extends his approach to counterfactuals, Barnett boldly combines a probability-based view, which characterizes counterfactual reasoning by the probabilistic relationship between the antecedent and the consequent, with a truth-conditional view.[54] He aims at integrating as well the insights of a Lewis–Stalnakerian nearness analysis as the virtues of the traditional metalinguistic approach according to which the truth of a counterfactual depends on the antecedent entailing the consequent given certain further assumptions. In sum, if Barnett is successful, he overcomes the main boundaries by which the philosophical debate has been marked so far. While I confine my critical discussion to counterfactuals, I shall give some hints how they might spell trouble for his suppositional view in general.

I shall focus on Barnett's 2010 paper. Barnett's method is uncommon. He introduces a semantics for an artificial expression 'zif'. After stipulating 'zif', he forwards a challenge: 'Anyone who rejects that *zif* would have been *if* faces the obvious challenge: to find a relevant difference between our entrenched practices with 'if' and our inchoate practices with 'zif' .' Since 'zif' is alleged to be 'if', I will translate Barnett's 'zif' claims to 'if'–claims where appropriate. Barnett stipulates some rules of 'zif'. Barnett calls the consequent what is *stated* by a conditional statement and the antecedent what is *supposed* by the conditional statement:

'Zif Probability A zif–statement is *n% probable* iff what is stated by the statement is made *n*% probable by what is supposed by it.
[...]
Zif Truth A zif–statement is *true* iff what is supposed by the statement entails what is stated by it.
Zif Falsity A zif–statement is *false* iff what is supposed by the statement is inconsistent with what is stated by it.'(p. 279)

These conditions solve the problem with lottery counterfactuals. We reject that

(D1) If Anna had bought a lottery ticket, she would have lost.

For (D1) is not true and we should not accept or assert what is not true.
Yet we accept

(D4) If Anna had bought a lottery ticket, she would probably have lost.

This seems perfectly vindicated by Barnett's conditions. 'Probably' may simply indicate that some threshold of probability is met. It may be further specified:

(D5) If Anna hat bought a lottery ticket, it is 99.99 percent probable that she would have lost.

However, this utility in solving the lottery issue is outweighed by the problems of the theory. I shall start with one problem. Take an everyday counterfactual which under certain circumstances seems perfectly true:

(D9) If I had got up 5 minutes earlier, I would have caught the train

[52] David Barnett, 'Zif is If', *Mind*, 115 (2006), 519–565; David Barnett, 'The Myth of the Categorical Counterfactual', *Philosophical Studies*, 144 (2009), 281–96.
[53] Barnett, 'Zif is If', p. 521.
[54] David Barnett, 'Zif Would Have Been If: A Suppositional View of Counterfactuals', *Noûs*, 44 (2010), 269–304.

(D9) seems true given I missed the train only by less than five minutes. Yet of course, my getting up 5 minutes earlier does not entail my catching the train. Thus, our explanation for why (D1) is unacceptable overgeneralizes to (D9). Thus, Barnett's theory can be expected to be highly counterintuitive from the outset, and indeed it turns out to unduly tax credulity.

I shall critically assess Barnett's theory in more detail, arguing for the following claims:

Barnett fails to provide an adequate closeness constraint (as in the standard account) for everyday counterfactuals, and it fails to do without such a constraint.
Since Barnett's view does not fare better with his own prime example than the standard possible worlds approach, he does nothing to rule out the latter.
His further linguistic evidence does not withstand critical scrutiny.
It is completely open how to modify Barnett's overall suppositional approach to indicatives such as to integrate his view of counterfactuals.

Barnett fails to provide an adequate closeness constraint (as in the standard account) for everyday counterfactuals.

I shall illustrate that, as contrasted to the standard account, Barnett fails to subject counterfactuals to an appropriate closeness constraint. I shall show where this failure leads to difficulties for the account: to begin with, consider the role of probabilities. Most everyday counterfactuals are not true but only probable according to Barnett's criteria. One important question is what probabilities are in counterfactual contexts. We have seen that Schulz's parallel between the Principal Principle, which deals with probabilities, and his principle *Credence* was problematic as it was difficult to reconcile with the volatility of probabilities. Barnett presents his own take on probabilities in the context of counterfactuals. In order to account for suppositional probability ascriptions, Barnett introduces *Conditional Counterfactual Probabilities*:

> 'CCP's appear to measure the *stability* of features and connections in the world. Suppose for illustration that a large number of children have been surveyed and that 95% of them like candy. The question arises whether this statistic reflects a relatively stable connection between being a child and liking candy, or whether it is purely accidental.
> [...]
> The relatively stable connections give way to ones that are more stable, more general, and more basic, until ultimately we reach the brute stabilities, including the fundamental laws of nature.' (p. 278)

I find Barnett's conception of probability in terms of stability difficult to understand. The more stable a connection between A and B is, the more probable A given be seems to be. This view leads to a dilemma. The first alternative is that stability is something along the following lines: a feature or connection is the more probable the higher the proportion of worlds at which it holds.[55] This is an insufficient basis for assessing probabilities of counterfactual suppositions. Consider:

(D9) If I had got up 5 minutes earlier, I would have caught the train

Assume that my probability of reaching the train on the counterfactual supposition that I get up 5 minutes earlier is high. But it does not entirely owe this to the stability of features in the world; the accidental fact how far from the station I actually am plays a crucial role. Even very stable relationships may fail to hold in arbitrarily many metaphysically possible situations. What is

[55] Cf. Edgington, 'On Conditionals', p. 308.

responsible for the probability of a counterfactual is not their stability tout court but their stability relative to sufficiently *close* situations.

Thus, the second alternative is to impose some nearness constraint on probability. Probabilities are assessed given things are as they actually are as far as compatible with the antecedent. But in contrast to the standard analysis of counterfactuals, closeness or preservation of actual facts is not built into 'zif' by default. Nor is it implicit in Barnett's notion of probability. Thus, Barnett's semantics is fatally incomplete. If we supplement it, we get something akin to the standard account, which Barnett wanted to overcome.

A second point where the failure of providing a closeness constraint leads to implausible results can be illustrated by a flash drama of Barnett's, which he uses precisely as a test for whether the semantics of counterfactuals is subject to such a constraint:

Dialogue2
'Smith: [D10] Zif she hadn't stepped on that mine, she would have made it across.
Jones: I doubt it. For suppose that she hadn't stepped on that mine. We must ask ourselves: what is the mostly likely way for this to have come about? Perhaps the initial conditions of the universe had been different; in which case it is highly unlikely that she, or this minefield, would ever have existed[...]
Smith: You are extremely uncharitable. Was it not obvious from our context that what I *meant* was that, [D11] zif she hadn't stepped on that mine *and things had been as similar as possible to actual, up to that point*, she would have made it across?
Jones: Well, in *that* case, she probably *would* have made it across. From now on, please *say* exactly what you *mean*.'(p. 285)

According to Barnett, this dialogue shows that Smith made an implicit closeness supposition over and above the explicit supposition. Since he did not made this supposition explicit, he can be chastized for talking loosely by Jones. To Barnett, this shows that a closeness constraint is not built into the semantics of counterfactuals.

However, replacing 'zif' by 'if', I find Jones' reaction not merely uncharitable but very odd. The possibility of the initial conditions of the universe being different seems simply irrelevant. In order to evaluate Smith's statement, we have to consider the actual situation modulo the soldier not stepping on a mine (however this is to be cashed out). Jones' move bringing into play weird antecedent situations is only saved from outright infelicity by our willingness to accommodate even very outlandish possibilities once they are brought up. Jones insistence that one should say exactly what one means leaves us clueless how to abide. The dialogue counts against Barnett's analysis of 'if' rather than supporting it. This again is evidence that Barnett would have to add a default closeness constraint to accommodate intuitions and to get a neat conception of probability for counterfactuals.

Here is another point at which the lack of an adequate closeness constraint proves fatal: Barnett suggests that *instead* of an implicit closeness supposition, some zif–statements may be subject to a subjunctive free–will supposition

'zif the soldier had *freely chosen* to step just to the left of where she actually *freely chose* to step, the events leading up to this choice would probably have been *just as they actually were*, for there is no reason to think they would have been different, and there is some reason to think they would have been the same.'(p. 287)[56]

Yet Barnett here misses 'zif' and tacitly replaces it by 'if'. Without a closeness constraint that privileges the way things are, *nothing* ensures that things 'would probably have been just as they actually were'.[57] Barnett's probabilities are based on empirically encountered regularities. Yet individual matters of fact are not fully fixed by regularities. For instance, regularities do not fix whether a coin that was actually tossed fell heads or tails. Whether such a process is

[56] Cf. Igal Kvart, 'Counterfactuals', *Erkenntnis*, 36 (1992), 139–179 (p. 141).
[57] Cf. David Lewis, 'Humean Supervenience Debugged', *Mind*, 103 (1994), 473–490 (p. 480).

deterministic or indeterministic, we have to add particular matters of fact which are not explicitly mentioned. The closeness constraint guides us in telling which particular matters of fact to add.
Barnett provides eight clues where an alien linguist examining the use of 'zif' can see that the standard account of counterfactuals does not apply, neither to 'zif' nor to 'if'. I shall critically discuss these purported clues.
Clue #1 is Barnett's argument against building a closeness constraint into the meaning of the counterfactual:

'the outsider might investigate whether explicitly adding a nearness–condition to the antecedent of a zif–statement has any effect on our evaluation of the statement. On the nearest–world hypothesis, it should not. [...]
[D12] Zif hamsters had wings, everything else would be as similar as possible to actual.
[D13] Zif hamsters had wings and everything else were as similar as possible to actual, everything else would be as similar as possible to actual.'(p. 288)

(D13) is necessary, (D12) is not, says Barnett. To Barnett, this is not reconcilable with (D12) being subject to an implicit semantic closeness constraint. If there were such a constraint, (D12) should also be necessary.

In order for Barnett's argument to succeed, the standard nearness truth–condition for counterfactuals would have to be:

For a world of evaluation w, a counterfactual A >> C is true iff C at some A–world which is closer *to actuality* (*our* world) than any A–world such that not–C.

But in fact, the standard truth–condition is this:

For a world of evaluation w, a counterfactual A >> C is true iff C at some A–world which is closer *to w* than any A–world such that not–C.

Here is the rub: to check whether, for any possible world w, (D12) is true, we have to consider the closest world w* to w at which hamsters have wings and everything else is as similar as possible wo *w*, not to actuality as in (D13). No wonder that (D13) is necessary, while this is not guaranteed for (D12).

I note that it might be metaphysically impossible that hamsters have wings, but this does not change the result. Assume it is impossible. Then, in the standard account, (D12) is simply necessarily true as far as a world at which hamsters have wings is inaccessible from any possible world.

Ad clue #2: One of the alleged virtues of Barnett's approach is that it accounts for counterpossibles, counterfactuals with impossible antecedents such as:

'[D14] Zif the truths of fundamental physics were discoverable by a priori conceptual analysis, particle accelerators would be superfluous.

[...]we judge some zif–statements to be about impossible scenarios, and our confidence in such statements is sometimes low and sometimes high. This does not comport with the hypothesis that zif–statements with impossible antecedents are vacuously true (or vacuously false).'(p. 289)

Yet Barnett does not say how *conditional counterfactual probabilities* may apply to impossible situations. What we would need is a detailed account in how far the supposed

impossible circumstances interfere with stable features of the world and in how far they do not, as it is given e.g. by Nolan's closeness account of impossible worlds.[58]

Without any hint as to how to deal with probabilities in this case, what remains in Barnett's account is that counterfactuals with metaphysically impossible antecedents are true iff the antecedent logically entails the consequent; and they are false iff the antecedent is inconsistent with the antecedent. Nothing in between. We are left without any clue how to deal with (D14). No advantage compared to the standard account according to which all counterfactuals with impossible antecedents are vacuously true.

Since Barnett's view does not fare better with his prime example than the standard possible worlds approach, he does nothing to rule out the latter.

Barnett's Clues #3–6 are derived from four principles that Barnett takes to hold for categorical statements and but not for 'zif':

'Clue #3: However confident one is that *S*, one should be at least as confident that there is an answer to the question of whether *S*.
Clue # 4: On the supposition that there is no answer to the question of whether *a* is *F*, one should have zero confidence that *a* might be *F*.
Clue #5: However confident one is that *S*, one should be equally confident that it is true that *S*.
Clue #6: Intuitively, it cannot be objectively incorrect to assign probability 1 to a categorical statement *and* objectively incorrect to assign probability 0 to the statement.'(pp. 290–292)

There are some doubts about these purported platitudes. For instance, the plausibility of Clue#4 depends on what is meant by 'there being an answer'. It the phrase is supposed to mean that there is a fact of the matter, Clue #4 sounds somewhat plausible. Less so for any meaning which includes somehow our capacity of giving an answer. Even in the first reading, it is not a matter of course that one ought to have zero confidence rather than refraining from forming any credential attitude. As for Clue #6, it is not clear how to deal for instance with statements about chancy future developments. Such statements may have an objective probability in between 1 and 0.[59]

Even granting the platitudes, the question becomes why they should fail for counterfactuals. Barnett's argument entirely rests on applying the analysis of 'zif' to one example, not on any further piece of independent evidence:

(D15) If there were a Goldilocks girl, she would like candy.

To Barnett, (D15) clearly shows that the four principles do not hold for 'if': (D15) is neither true nor false; the antecedent does neither entail nor contradict the consequent. Thus, there is no answer to the question whether (D15) is true. The girl might like candy and she might not like candy. Yet assessing the relevant probabilities gives rise to a high confidence that the girl would like candy.

Now there is no reason within the standard analysis why #3–#6 should not hold for counterfactuals. As a consequence, Barnett's argument depends on his analysis of (D15) being superior to the standard analysis. Let us compare Barnett's results to the standard analysis. It does not sound that implausible that (D15) is neither true nor false. How can the standard

[58] Cf. Daniel Nolan, 'Impossible Worlds*:* A Modest Approach', *Notre Dame Journal of Formal Logic*, 38 (1997), 535–572.
[59] Ad Clue #6: Take objective chances. We are going to throw a coin. It seems objectively correct to assign a credence of 0,5 to the statement that it will fall heads. It is objectively as incorrect to assign probability 1 as probability 0. One may deny that 'the coin will fall heads' is a categorical statement, but some authors would in general assign a statement on a chance device like a fair coin 'the coin falls heads' a chance of 0.5.

analysis accommodate that? Lewis might have pointed out that (D15) is vague. Contrary to the first appearance, it is very different from everyday counterfactuals. In contrast to 'If I had got up earlier today…' which solidly hooks into a concrete actual situation, (D15) does not give us enough to envisage a concrete scenario. For instance, when and where does Goldilocks live? In Lewis's default nearness analysis, a small miracle or inconspicuous divergence from actual facts would have to bring about the antecedent. But where is this divergence to be located? What does it look like?

Here is Lewis on vagueness:

'Thus we account for such pairs of counterfactuals as Quine's

If Caesar had been in command [in Korea] he would have used the atom bomb.
Versus
If Caesar had been in command he would have used catapults. [...]

I could [...] call on context rather to resolve part of the vagueness of comparative similarity in a way favourable to the truth of one counterfactual or the other.'[60]

In the same vein, (D15) may call for further ways of cashing out the story. In some of them it comes out true, in some it comes out false. Of course, there are also problems with (D15) being a fictional character, which I disregard here.

Besides vagueness, there are further alternatives for interpreting (D15) within the standard account: We may reckon a world where a Goldilocks girl likes candy more similar than a world where she does not. For instance, we may say that the latter world instantiates less high probability properties; after all, girls usually like candy.[61] Then (D15) comes true.

An alternative way of dealing with (D15) would be to insist that worlds where the girl likes candy and worlds where she does not are equally close. Then the Goldilock case resembles chancy situations the paradigm of which is the throwing of a dice. In the Lewisian standard analysis, both 'If a dice had been thrown, it would have landed six' and 'If a dice had been thrown, it would have not landed six' are false. Analogously, both 'would like candy' and 'would not like candy' turn out false. In contrast, if we accept Stalnaker's uniqueness assumption that there is precisely one closest world to be selected by the selection function, (D15) becomes neither true nor false, just as Barnett has it. In sum, there is plenty of room for reconciling the standard account with any intuitions one might have.

Still, we might feel inclined to ascribe a high probability to a Goldilocks girl liking candy. This is reflected in our accepting as true

(D16) If there were a Goldilocks girl, she would probably like candy

and

(D17) If there were a Goldilocks girl, she might/might not like candy

Here we indeed have our original problem with lottery counterfactuals. Assume that (D15) comes out false or indeterminate in the standard analysis. For instance, let there be many candidates for closest worlds where Goldilocks likes candy and some worlds where she does not. Still (D16) seems plausible. However, we should not accept Barnett's account only because it allows to deal with such lottery cases. After all, we have seen, and we will see that there are

[60] Lewis, *Counterfactuals*, pp. 66–67.
[61] J. Robert G. Williams, 'Chances, Counterfactuals, and Similarity', *Philosophy and Phenomenological Research*, 78 (2008), 385–420.

alternatives. Barnett's proposal must be judged by its other merits or demerits compared to those of its rivals.

Barnett's further linguistic evidence does not withstand critical scrutiny.

I shall come to Barnett's further evidence. Consider

Clue #7
'[...]there is no need to qualify the proposition *that Jones is the murderer* –by, say, 'probably', 'definitely', or 'possibly' in order for a categorical statement of it to be significant. By contrast, subjunctive contents stated relative to subjunctive suppositions do require qualification for their statements to be significant.'(p. 295)

Barnett's idea must be the following: given the semantics of 'zif', either a counterfactual is *definitely* true in virtue of entailment, or it is only *probably* true.

I disagree with Barnett's claim. A normal categorical statement like

(D18) Jones is the murderer

does not need qualification. A counterfactual

(D19) The glass would have shattered if dropped

does not either. We usually utter statements of both kinds without qualifying them. Yet in both cases, we tend to be in a quandary when pressed. Someone might respond: 'Definitely, probably, or possibly?' Normally, on the one hand we will feel committed by our previous utterance to accept one of the options; but on the other hand, we will often hesitate. 'Definitely' might sound too strong, 'probably' too weak. This, I guess, is due to a certain vagueness and intransparency of the threshold of certainty or vindication at stake in our attitudes such as belief or explicit acceptance and the threshold required by the qualification. At least there is no indication that counterfactuals are more in need of qualification than categorical statements. The claim is obviously an artefact of zif–semantics.

There is a further problem: according to Barnett 'definitely' requires entailment. Thus, were 'zif' if, the following should be infelicitous:

Dialogue3
Al: '(D9) Had I got up five minutes earlier, I should definitely have reached the train.'
Bo: 'Definitely? After all, five minutes is not much, and the way is far.'
Al (who happens to be a sprinting champion): 'Definitely!'

(D9) obviously does not meet the requirement of entailment. Still this dialogue sounds perfectly in order. In contrast, what would definitely sound odd is the ziffy:

Bo: #'Come on, what about a sudden volcano eruption or a break in natural laws? You should mind your words. Just add 'probably' (and a nearness constraint) instead of this conceited 'definitely'!'

Furthermore, even when we hesitate to call a chancy counterfactual definitely true, we might not hesitate to call it *true* when qualified:

Dialogue4
Ed: '(D19) The glass would have shattered if dropped.'

Ella: 'Is that so?'
Ed: 'Well, that much is *true*:
(a) it would *probably* have shattered if dropped / (b) (D20) if the glass had been dropped, there would *definitely* have been a high chance of its breaking when dropped.'

Although Barnett might be able to accommodate (a), he seems unable to accommodate (b). Again the antecedent of (D20) does not entail the consequent.

Compare the evidential impact of these simple examples to Barnett's:

Dialogue5
'Ella: Suppose the glass had fallen!
Ed: It definitely would have shattered.
Ella: Well, I hate to be a stickler, but I don't think it's right to say that it *definitely* would have shattered. For, as unlikely as it sounds, a perfect gust of wind could have brought the glass to a gentle landing. [...] just think of a couple of the ways that the glass could have fallen. It could have fallen due to a subtle difference in the initial conditions of the universe, say, one that led to your reactions being a bit slower than they actually were. This difference could also have led to the existence of a perfect gust of wind. Another way that the glass could have fallen is for there to have been a subtle difference in the laws of nature, say, one that led to the glass's accelerating slightly faster than it actually did.'(p. 280)

I think our intuitive grip on such an example is loose. Again the most plausible diagnosis is that we are willing to consider even far-fetched possibilities once they are brought up. This does not mean that they are relevant from the outset. Note that by Barnett's lights, Ella could as well appeal to a huge difference in laws of nature. According to the Lewisian standard analysis, such circumstances are too far-fetched to count as closest antecedent worlds. The Lewisian standard closeness conditions eschew them as well as a subtle difference in the initial conditions of the universe and Ella's subtle difference in the laws of nature as far as the latter has no role in bringing about the antecedent as distinguished from Lewis's default small miracle. Yet Lewis's standard analysis cannot make as short work with certain very improbable chance processes, such as perhaps the sudden gust of wind. They are candidates for closest antecedent worlds. There is a huge debate on the issue.[62] Here I think the standard analysis is perfectly in tune with our intuitions. We tend to neglect certain chance processes.[63] But when we are pressed, we are in a quandary as to how to deal with them.

I doubt that there is a more eligible way to handle Barnett's extremely artificial dialogue than by dismissing some circumstances and feeling in a quandary when others are raised to salience. If I feel any intuitive pull, then it is to accept Ed's initial statement as *perfectly in order*. This is what Barnett denies and what is accounted for by the standard analysis. The quandary created by the dialogue only testifies to our willingness to accommodate far-fetched possibilities, not to their playing a role for the truth of counterfactuals in normal contexts. One proposal for how to account for accommodation is that raising far-fetched possibilities is incompatible with the normal contexts under which counterfactuals are assessed. Just as there are counterfactuals like the Caesar examples which cannot be assessed outside of a strong context, there may be contexts which interfere with our confident assessment of counterfactuals. Bringing up far-fetched possibilities may create such contexts.

Clue #8 provides further linguistic evidence. 'When' and 'where' denote times and places. For a counterfactual to denote anything, it would have to denote a situation. Yet it does not denote anything, says Barnett. To mark the difference, Barnett notes there are six places where 'probable' can be inserted into a counterfactual:

(D21a) It is probable that hamsters would fly, zif they had wings

[62] Cf. Williams, 'Chances'.
[63] Alan Hajek, 'Most Counterfactuals are False', unpublished Manuscript.

(D21b) It is probable, zif hamsters had wings, that they would fly
(D21c) Zif hamsters had wings, it is probable that they would fly
(D21d) Zif hamsters had wings, that they would fly is probable
(D21e) That hamsters would fly, zif they had wings, is probable
(D21f) That hamsters would fly is probable, zif they had wings.(p. 297)

In contrast, there are only four places where 'probable' can be inserted into 'when' or 'where'–statements:

(D22a) It is probable that I will live where Sharon lives
(D22b) It is probable, where Sharon lives, that I will live
(D22c) Where Sharon lives, it is probable that I will live
(D22d) That I will live where Sharon lives is probable.

The remaining two combinations are awkward, to say the least:

(D22e) Where Sharon lives, that I will live is probable
(D22f) That I will live is probable, where Sharon lives. (p. 298)

Coming to my criticism, it is not obvious that the purported differences have anything to do with the issue of denotation. I do not deem them very significant anyway. Note that for instance the German equivalent of (21a), *Es ist wahrscheinlich, dass Hamster flögen, wenn sie Flügel hätten*, allows 5 variants at best (no equivalent to 21d). Moreover, we may try the following instead of (D22e) and (D22f):

(D22e´) Where Sharon lives, there that I will live is probable / at that place that I will live is probable.
(D22f´) There/at that place that I will live is probable, where Sharon lives.

This is not elegant, but can one be sufficiently confident that it is infelicitous to build a deep distinction between 'if' on the one and 'where' and 'when' on the other hand on this verdict? At least concerning (D22e´) I have got mixed reactions from native speakers.

Barnett gives further purported evidence that 'if'–sentences do not denote a situation: '…whereas "the time when Sharon leaves" and "the place where Sharon lives" are grammatical, "the hypothetical situation zif Sharon had left" is not.'(p. 298)[64] Again it is not obvious that this observation has anything to do with the issue of denotation. 'When' and 'where' can be used in direct questions, 'if' can only be used in indirect questions like: 'I ask you if…'. According to the only available account of the relationship between such questions and conditionals, 'if' highlights a positive answer to an indirect question.[65] This observation can be accommodated by a denotational view of conditionals: the positive answer to an inexplicit indirect if–question determines the situation of which the consequent is to hold for a counterfactual to be true.

It is completely open how to modify Barnett's overall suppositional approach to indicatives such as to integrate his view of counterfactuals.[66]

Even granting that Barnett's account is adequate for counterfactuals, we may ask how it fits into his general picture of supposition. According to Barnett, 'zif' is generally to be used to make a suppositional statement. Then so it should be used for counterfactuals. As we have seen,

[64] Cf. Barnett, 'Zif is If', pp. 528–529.
[65] William S. Starr, 'What If?', *Philosophers' Imprint*, 14 (2014).
[66] Cf. Barnett, 'Zif is If'.

Barnett denies that 'zif' denotes a situation; yet he accepts that there is denotation in play: 'conditional denotation'. An antecedent situation is denoted provided there is one.[67] If A is false, nothing is denoted:

'Joe says, [D23] 'Zif the Pope visited yesterday, then we will have a good year'. The outsider responds, 'What do you mean *then* we will have a good year? There is no *then*, because there was no visit by the Pope'. To which Joe responds: 'Surely you must recognize the possibility that you are wrong—that the Pope did in fact visit yesterday. *Suppose* this is so. *Then* we will have a good year. When I say 'then', I only *aim* to be talking about a situation in which the Pope visited yesterday *conditional* on there being such a situation. No Pope, no aim.''[68]

According to Barnett, 'then' in (D23) denotes something *conditionally*. Yet putting into abeyance my above criticism of Clue #8, I do not see any reason why the linguistic evidence for Clue #8 does not apply as well to indicative suppositions.[69] According to Barnett, these suppositions *do* denote something (albeit conditionally). This is incompatible with the lesson Barnett draws from his Clue #8. Thus, either the evidence for the claim that counterfactuals do not denote or the suppositional account has to go.

How can Barnett's template for indicatives be transferred to counterfactual situations? Consider

(D24) If the pope had visited yesterday, *then* we would have a good year.

Straightforward application of the template for indicatives gives: When the antecedent is false, there is nothing to be denoted by 'then'. 'No pope, no aim'; nothing to be aimed at; still there is 'an *absent* attempt at reference rather than a *failed* attempt.'[70] How are we to understand an act which amounts to nothing but an *absent attempt* at denoting whatever 'then' is to denote? What does it mean to aim at something when it is at the same time conveyed that there is nothing to be aimed at? The only way of making sense of such an act is to make the absent attempt parasitic on the success case: A obtains; at least it is somehow open whether A obtains. As Joe responds: 'You must somehow recognize the possibility that you are wrong.' Thus, the problem of accounting for suppositional statements when A is false becomes more grievous in the counterfactual case. It would seem odd to say that for any *genuine* counterfactual (with actually false antecedent), 'then' fails to denote; there is nothing but an *absent attempt* at reference. There is nothing for 'then' in (D23) to stand for.

I shall consider a further argument of Barnett's in favour of his suppositional account of indicative conditionals. The opponent of a suppositional account is faced with the following problem: she is committed to making sense of the question what the truth–value of 'zif A, C' is under the supposition

'[...]that it is false that A. This amounts to a request to evaluate whether C while supposing not just that A but also that it is not the case that A. And this is a request that we cannot satisfy. Hence our response: 'We are at a loss as to how to respond, for we are unable to evaluate the statement under the supposition that it is false that A'.'[71]

[67] Cf. Edgington, 'Conditionals'.
[68] Barnett, 'Zif is If', p. 529.
[69] Consider the indicative: (D21a) It is probable that hamsters fly, zif they have wings… There is no 'situation zif Sharon has left'.
[70] Barnett, 'Zif is If', pp. 529–530. Edgington has it that, when A is false, nothing is asserted (Edgington, 'Conditionals', p. 289). In contrast, Barnett insists: '[...]one who asserts that, zif *A*, *C*, asserts something –namely, that *C*– regardless of whether *A*.'(p. 543) Regardless of whether A, C is asserted *under the supposition that A*.
[71] Barnett, 'Zif is If', p. 536.

The proponent of a suppositional account avoids this problem: under the supposition that it is false that A, her suppositional statement 'zif A, C' is null and void.

Before coming to the point of interest, I shall make a quick a comment on this argument: I do not think that the opponent of the suppositional account really has a problem here. For instance, in Stalnaker's view, the presupposition of the indicative conditional is that the antecedent situation is an open possibility. One may retort to Barnett that the supposition that it is false that A is incompatible with the antecedent being an open possibility.

However, I did not bring up Barnett's argument to rebut it. Rather my aim is to use it in a criticism of his suppositional account of counterfactuals: applying Barnett's own way of putting supposition, supposing A while *presupposing* that it is not the case that A seems precisely to be what the suppositional template demands when we evaluate a counterfactual. What remains is that nothing seems ever to be stated by a genuine counterfactual, not even C *under the supposition* that A. So for any counterfactual, we 'are unable to evaluate the statement under the (pre)supposition that it is false that A'.

A similar problem: 'When we believe under a supposition, we aim at the truth, but we are only committed to this goal *on the condition that the supposition obtains*.'[72] If this move were transferred to counterfactuals, it would seem that one incurs no commitment at all by them. So it remains open how to accommodate *counterfactual* suppositions within Barnett's overall approach.

There are further difficulties of transferring the suppositional view of indicatives to counterfactuals. A 'zif'–statement 'zif A, C' is true iff C *is true* on the supposition that A. Provided we take this as a model for counterfactuals as well, C would be true on the supposition that A iff A entails C; then the probability of C being true on the supposition that A is 1. 'Zif A, C' is n% probable iff: C *being true* on the supposition that A is n% probable.[73] But for the counterfactual, the condition of C being true on the supposition that A is that A entails C. If A does not entail C, the probability that it does entail C is 0. So how can the probability of 'zif A were the case, C would be', i.e. of C being true on the supposition that A ever be different from 0 or 1?[74]

Negation causes trouble, too: 'A statement that it is not the case that, zif *A*, *C* is a statement of a unique thing—that it is not the case that *C*—within the scope of the supposition that *A*.'[75] If we apply this to counterfactuals, from our accepting 'It is not the case that if the coin is/were thrown, it will/would fall heads' it seems to follow that if the coin is/were thrown, it will/would not be the case that it falls heads. So it will/would not fall heads. But we deny that if the coin is/were thrown, it will/would not fall heads.

In sum, while Barnett's approach provides a solution to the problem of probabilistic counterfactuals, his overall theory is highly implausible.

[72] Barnett, 'Zif is If', p. 542.
[73] 'How likely is it to be true that, zif this fair coin is flipped, it will land heads?' To which we respond: 'Fifty percent'(Barnett, 'Zif is If', p. 540)
[74] For a parallel cf. DeRose, 'Conditionals', pp. 12–13.
[75] Barnett, 'Zif is If', p. 546.

2.3.2. A New Proposal: Non–Maximality

Lottery counterfactuals are notoriously puzzling. They seem in tension with the standard account of counterfactuals. Firstly, some counterfactuals which are not true according to the standard account become true when 'probably' is inserted. Secondly, we assign high credence to lottery counterfactuals which are clearly false according to the standard account. So far there is no universally accepted solution to the problem.

I have criticized the extant proposals of Schulz and Barnett. I present a new approach, which does with a minuscule amendment to the standard account. Just as descriptions, conditionals are homogeneous and non-maximal. Homogeneity: some conditionals are neither true nor false if not all relevant (closest) antecedent worlds are consequent worlds. Non–maximality: in certain contexts, not all relevant antecedent worlds have to be consequent worlds for the utterance of a conditional to say something true. Lottery contexts exclude the non-maximal reading, but they are compatible with explicitly weighing the proportion of consequent worlds among the relevant antecedent worlds. This is what happens in the problematic counterfactuals.

The puzzle of lottery counterfactuals

The interaction between conditionals and probability is difficult to understand. We have seen that the following is infelicitous:

(D1) #If Anna had bought a lottery ticket, she would have lost.

Yet in contrast to (D1), (D4) seems perfectly acceptable in many contexts:

(D4) If Anna had bought a lottery ticket, she would probably have lost.

'Probably' may be further specified.

(D5) If Anna hat bought a lottery ticket, it is 99.99 percent probable that she would have lost.

For now I focus on (D4). It is not so easy to understand how (D4) can be acceptable. The interaction between 'probably' and 'would' is intricate. Here are two possibilities to understand (D4): in one alternative, (D4) simply says that (D1) is *probably true*. Yet judging from the standard account, we can know for sure that (D1) is *not* true. In a second alternative, (D4) says that, in all relevant worlds where Anna has bought a ticket and the draw has taken place, the probability of her having lost is high. Yet in most relevant worlds, the probability of her having lost after the lottery draw is 100%, whereas in some, it is 0%.[76] There are further alternatives how to read (D4) which come closer to the intuitive result, e.g. the probability of Anna's ticket losing given she buys (precisely) one at some suitably chosen points in time prior to the draw. The relevant probability may be probability at the evaluation world (the actual one) or probability at the individual relevant antecedent worlds. Alternatively, we may count the proportion of worlds where Anna loses among the relevant worlds where she buys a ticket. The latter alternatives come close to the intuitive results, but they need motivation. The account to come provides one.

A new proposal: non–maximalility
conditionals display homogeneity

[76] I assume that the past is no longer chancy.

My alternative proposal elaborates on the hypothesis that conditionals resemble descriptions in displaying homogeneity and non–maximality. To introduce the hypothesis, I start with evidence about '*would*' conditionals. There is some reason to think that stressing 'would' or inserting 'definitely' makes a difference to how we understand conditionals. Sarah Moss has observed that the following sounds marked:[77]

(D25) #It is not the case that Anna would have lost if she had bought a ticket.

But it seems perfectly all right to say

(D26) It is not the case that Anna *would*/would definitely have lost if she had bought a ticket.

The difference calls for an explanation. Here is a tempting suggestion: '*would*' conditionals stress that one has to take into consideration all relevant antecedent worlds. They ALL have to be consequent worlds. Yet it is not clear from the outset how to integrate this idea into the standard account. According to the latter, any normal 'would' conditional has one take into account all relevant antecedent worlds. The difference in felicity between (D25) and (D26) remains mysterious.

At first glance, Schulz's arbitrariness account can make room for a more significant difference, as the following proposal shows: '*would*' tells one to consider all relevant worlds *and not only an arbitrarily selected one*. But this change in the semantics seems somewhat arbitrary. The point of arbitrary selection is that we perform a test on all relevant antecedent worlds. One is randomly selected as representative. Why add an extra device which requires to consider any particular relevant antecedent world? The only explanation I can imagine is that '*would*' conditionals correct shortcomings of the arbitrarily selected world as a representative of the relevant worlds. But then there should also be advantages to compensate the shortcomings of arbitrary selection compared to taking into account all relevant worlds, pragmatic or epistemic or whatever. The proponent of the arbitrariness account may be challenged to say more about these advantages. Again I cannot rule out that this challenge can be met. But I do not think it has so far been met.

In an alternative understanding, '*would*' conditionals are strict conditionals: all antecedent worlds in a contextually provided modal horizon have to be consequent worlds, not only the closest ones. If this alternative is to explain the difference, the unstressed 'would' conditional had better not be a strict conditional.[78] This option results in a surprisingly deep semantic difference between 'would' and '*would*'.

A radically different approach to 'definitely'/'*would*' is to treat them as epistemic modals, expressing certainty as in:

Dialogue6:
Al: 'Berlin is bigger than Madrid.'
Bo: 'Definitely?'
Al: 'Definitely.'

But (D26) cannot be interpreted as expressing a lack of certainty concerning (D1). For instance, had Al been uncertain in Dialogue6, he could not have expressed his uncertainty by:

…
Bo: 'Definitely?'

[77] Moss, p. 2.
[78] Cf. Gillies, 'Counterfactual Scorekeeping'.

Al: ‘No. ?It is not the case that Berlin is definitely bigger than Madrid’.

I shall explore an alternative explanation of the difference between (D25) and (D26) which I find more appealing than the ones mentioned so far. But I have to add further evidence first. There is a close similarity between the findings on counterfactuals and the behaviour of indicative conditionals and future-directed declarative sentences. Just like (D1), the following often seems inappropriate to sincerely assert:

(D27) #If Anna buys a lottery ticket, she will lose.

After Anna has bought a ticket
(D28) #Anna’s ticket will lose.

Just as (D26), the denial of the above sentences sounds odd:

(D29) #It is not the case that, if Anna buys a lottery ticket, she will lose.
(D30) #It is not the case that Anna’s ticket will lose.

But the stressed versions seem much better:

(D31) It is not the case that, if Anna buys a lottery ticket, she WILL/will definitely/must lose.
(D32) It is not the case that Anna’s ticket WILL/will definitely/must lose.

To explain these data, I shall draw on the suggestion that indicative conditionals display semantic *homogeneity*. To get a grip on this notion, consider a case where homogeneity is uncontentious:

Dialogue4
Talking about books in a library (half of the books are in Dutch)
Al: (D33) #The books are in Dutch.
Bo: (D34) #It is not the case that the books are in Dutch.
Alternatively Bo: (D35) Not *all* the books are in Dutch.

We feel that Al’s utterance of (D33) is weird, but we hesitate to call it false. This can be explained as follows: it is commonly accepted that an incomplete description like (D33) tends to be read as homogeneous.[79] An incomplete description ‘the *F* are *G*’, read as homogeneous, is true precisely if all *F* which are contextually maximally salient are *G*. It is false precisely if no maximally salient *F is G*. When only some of the maximally salient *F* are *G*, the description is neither true nor false. In contrast, a universally quantified sentence ‘all *F* are *G*’ is to be read non-homogeneously; it is true precisely if all *F* are *G* (perhaps in some contextually restricted domain) and false otherwise.

Given close connections between descriptions and conditionals, it is tempting to assume that conditionals also display homogeneity.[80] The difference between (D25) and (D26) and the difference between (D33) and (D35) can then be explained as follows: a conditional in a context displays *homogeneity* provided the following holds: it is true precisely if all relevant

[79] Manuel Križ, and Emanuel Chemla, ‘Two Methods to Find Truth–Value Gaps and their Application to the Projection Problem of Homogeneity’, *Natural Language Semantics*, 23 (2015), 205–248; Manuel Križ, ‘Homogeneity, Non–Maximality, and *all*’, *The Journal of Semantics*, 33 (2016), 493–539.
[80] On general connections between descriptions and conditionals Maria Bittner, ‘Topical Referents for Individuals and Possibilities’, *SALT*, 11 (2001), pp, 36–55; Philippe Schlenker, ‘Conditionals as Definite Descriptions’, *Research on Language and Computation*, 2 (2004), 417–462.

antecedent possibilities are consequent possibilities, false precisely if none of them are, otherwise gappy. Homogeneity is ruled out by '*would*' counterfactuals and musty indicatives. They are false unless all relevant antecedent worlds are consequent worlds. To derive an explanation of the intuitive differences, I need a further substantial assumption: we are hesitant to accept an outer negation as true if the negated sentence is not false: 'Intuitively, a sentence [*it is not the case that S*] will be true exactly when S is false.'[81] It is not fully clear whether the assumption generalises, but at least for descriptions ('it is not the case that the books are in Dutch') and conditionals, it seems fairly plausible. (D25), (D33) , (D34) are cases where the negated sentence is neither true nor false, and thus we are hesitant to accept the outer negation. In contrast, (D26), (D35) are acceptable because the negated sentences are clearly false.

(D28) can be treated in the same way. We read it as dealing with an open future. Anna might lose and Anna might win. I remain neutral in what sense the future is open, epistemically or metaphysically. In uttering (D28), one presupposes that Anna has a ticket. The utterance is true precisely if all relevant future situations are situations where Anna loses, false if none of them are, otherwise neither true nor false. Again the musty version ('Anna must lose/will definitely lose') removes homogeneity: all future situations have to be situations where Anna loses.

Conditionals display non–maximality
non–maximality in descriptions

Homogeneity is closely associated with a closely related phenomenon: non–maximality. Normally an incomplete description is taken to select precisely the contextually most salient individuals satisfying the descriptive condition expressed in the subject noun phrase. However, often descriptions tolerate exceptions among the contextually maximally salient individuals:[82]

All the professors except Smith smiled and then left, leaving Smith behind.
(D36) The professors smiled.
(D37) ?The professors smiled and then (all) left the room.

One may try to explain the felicity of (D36) by domain restriction, i.e. some domain of quantification being restricted to the smiling professors. As a consequence, the utterance of (D37) should also be felicitous. 'Then (all) left the room' would quantify over the restricted domain. To account for the difference between (D36) and (D37), *'*the professors*'* in (D36) must not be read as *all the professors in a contextually restricted domain, excluding Smith,* but as allowing for exceptions from a set of contextually most salient professors (including Smith).[83]

Here is a first take on the example: on the one hand, the maximal reading stands out as a point of departure. The maximal reading selects precisely the contextually most salient individuals. There are means of enforcing a corresponding universal quantification over a contextually restricted domain (*all*). On the other hand, examples like (D36) provide evidence that many contexts do not only privilege a certain *maximal set* of most salient individuals which satisfy some description. These contexts also fix a range of *tolerable departures* from the maximal set. Within that range, it does not matter whether all individuals in the maximal set

[81] Gennaro Chierchia and Sally McConnell–Ginet, *Meaning and Grammar. An Introduction to Semantics* (Cambridge/Mass.: MIT Press, 1999), p. 76.
[82] Example from Križ, 'Homogeneity', p. 498.
[83] I have encountered the tendency to draw a parallel to generics. The parallel is limited: firstly, definite descriptions are not the standard way of expressing generic statements. Secondly, the general criteria for generics (cf. Sarah–Jane Leslie, 'Generics. Cognition and Acquisition', *The Philosophical Review*, 117 (2008), 1–47) seem replaced by something more context–sensitive in the case of non–maximal descriptions. Conditionals seem more amenable to a generic or habitual reading ('If it rains, the streets are wet'), but it remains to be seen in how far non-maximality accounts for this reading.

satisfy the condition imposed by the predicate, or whether there are some exceptions. In the example of the smiling professors, (D36) is acceptable if it only matters that *almost* all maximally salient professors smiled, i.e. Smith not smiling is a tolerable exception. Context determines how many of the professors have to smile for (D36) to be acceptable.

There are competing analyses of non-maximal descriptions. The most advanced proposal by Križ bases non–maximality as a pragmatic phenomenon on homogeneity as a semantic phenomenon:[84] in order to be assertable in a situation, a sentence *S* has to address a contextual *issue*. The issue comes with a contextually relevant partition of possible worlds that are of current interest. A necessary condition for *S* to address the issue is that no cell in the partition at issue contain both a world where *S* is true and a world where *S* is false. But there may well be a cell in the partition which contains worlds where *S* is true and worlds where *S* is not true. A homogeneous description 'the *F* are *G'* is true precisely if all *F* are *G*, false precisely if no *F* is *G*. Otherwise it is neither true nor false. When a homogeneous description is felicitously uttered, the final non-maximal meaning is computed as follows: the utterance presupposes that there is a unique cell in the partition at issue which contains some possible world where the description is true (all *F* are *G*) and no possible world where the description is false (no *F* is *G*). The actual world is claimed to fall into this unique cell. If it does, the utterance is true. The tolerable exceptions are determined indirectly: worlds with a tolerable number of exceptions (*F* that are not *G*) are lumped together in one cell with worlds where all *F* are *G* without exception. Worlds with too many exceptions are lumped together with worlds where the description is false.

I add two important qualifications: firstly, once an exception has been mentioned, it cannot be neglected. In the example of the smiling professors, asserting (D36) is inappropriate once Professor Smith has been mentioned:

Dialogue7
Al: 'Smith didn't smile.'
Bo: (D36) ?'But the professors smiled.'

Secondly, we know from epistemology that lottery contexts do not tolerate exceptions. Anna cannot know that her ticket will lose, however minuscule the probability is that it won't. Any particular outcome counts. For this reason, often lottery contexts are not hospitable to reading descriptions non-maximally. For instance, normally it seems irresponsible to say about a fair lottery:

(D38) #The tickets will lose.[85]

There are lottery contexts, broadly conceived, where a non-maximal reading seems in order. In these contexts we do not attend to the chanciness of the outcome. The more we focus on a lottery aspect, the greater the difficulty will be to enforce a non-maximal reading.

Non–maximality in conditionals

I shall now consider the proposal that counterfactuals also display homogeneity and non–maximality.[86] As we have seen, homogeneity is the semantic phenomenon that there is a third

[84] See also Križ's, 'Homogeneity', criticism of Sophia Malamud, 'The Meaning of Plural Definites: A Decision–Theoretic Approach', *Semantics&Pragmatics*, 5 (2012), 1–58.
[85] (D38) seems odd regardless of whether 'the tickets' refers to all tickets in the lottery or to some salient subset.
[86] Discussing homogeneity about conditionals enmeshes us in the debate on conditional excluded middle (CEM). While I grant that the issue is not yet settled, I note that none of the alternative accounts of lottery conditionals discussed so far supports CEM.

option between truth and falsity: sometimes it is indeterminate whether a conditional is true or false. Non–maximality is the pragmatic phenomenon that, for a conditional to say something true in a context, not all but only *sufficiently many* relevant antecedent worlds have to be consequent worlds. I follow Križ in assuming that homogeneity and non–maximality are closely related. Non–maximality can only arise where there is a space of indeterminacy which could be filled. A conditional that is neither true nor false can nevertheless be used for truly uttering that a significant proportion of relevant antecedent worlds are consequent worlds.

Before coming back to the lottery cases, I shall take a look at some related puzzles which can be neatly dissolved by invoking non–maximality. This provides further evidence for the non-maximal reading of conditionals, including counterfactuals. In particular, I shall consider the interaction of ordinary 'would' and 'might' conditionals. Let there be a delicate china plate. We are inclined to assent to

(D39) If the plate had been dropped, it would have shattered.

Yet applying lessons from quantum physics, we are also inclined to accept:

(D40) If the plate had been dropped, it might have flown off sideways.

It has been noted, however, that (D39) and (D40) cannot be freely combined. The sequence (D39) to (D40) seems all right:

(D39) If the plate had been dropped, it would have shattered;
but (D40) (if the plate had been dropped,) it might have flown off sideways.

Yet the reverse sequence feels odd:[87]

(D40) If the plate had been dropped, it might have flown off sideways;
but (D39) #(if the plate had been dropped,) it would have shattered.[88]

The asymmetry between (D39)–(D40) and (D40)–(D39) is difficult to explain if one endorses the duality of 'would' and 'might': a 'would' conditional is true precisely if the corresponding 'might not' counterfactual is false. Schulz's arbitrariness account can explain the asymmetry as follows: the duality of 'would' and 'might' is rejected. For a 'might'–conditional to be true, just one relevant antecedent world has to be a consequent world. The sequence (D39)–(D40) can be consistently uttered. (D39) is very probably true. In any normal context, this provides sufficient ground to utter (D39). But (D40) is true as well, as there is a relevant antecedent world where the plate flies off sideways. The infelicity of the reverse sequence (D40)–(D39) can be pragmatically explained by findings from epistemology. Raising a relevant alternative where a belief is false changes the stakes for the belief to count as known, and thereby also raises the stakes for asserting it. Uttering a 'might not'–counterfactual (or something that entails it, like (D40)) raises the possibility that the corresponding 'would' counterfactual is false, namely if the arbitarily selected antecedent–world is a world where the consequent is false.

[87] Keith DeRose, 'Can It Be That It Would Have Been Even Though It Might Not Have Been?', *Philosophical Perspectives*, 33 (1999), 385–413.

[88] Perhaps the latter sequence can be uttered with a '*would*' conditional. This seems surprising given the idea that the '*would*' conditional enforces homogeneity and thus rules out non–maximality. But I guess that the effect is an indirect one: by ruling out homogeneity, one makes clear that one excludes the situations where the plate flies off sideways as irrelevant.

A rival explanation of the asymmetry is that the set of relevant antecedent worlds which have to be consequent worlds underlies contextual shifts. There are several possibilities how this shift works. One alternative is to claim that a 'might'–conditional tends to enlarge the range of accessible worlds as long as the latter does not include an antecedent–cum–consequent world. This claim can be implemented within a *strict conditional approach* to subjunctive conditionals.

My preferred alternative invokes non–maximality: a 'would' conditional sometimes leaves room for inexplicit exceptions among the relevant antecedent worlds (i.e. for worlds where the plate is dropped and flies off). Yet once an exception has been explicitly mentioned (by uttering the 'might'–conditional), it has to be taken into account.

To get a better feeling for the linguistic data, it may help to consider combinations of stressed '*would*' and 'might' conditionals. Take

Dialogue8
Al: (D39) 'If the plate had been dropped, it would have shattered.'
Bo: 'But would it definitely have shattered?'
Al: 'No. I admit that,
(D40) if the plate had been dropped, it might have flown off sideways.
Hence
(D41) it is not the case that, if the plate had been dropped, it *would* / would definitely have shattered.'

I note that a perfectly analogous dialogue could be run with indicative conditionals.

One challenge to the strict conditional theory is to account for Bo's question. If (D39) is read as a strict conditional, what could the horizon of assessing Bo's 'definitely' be? If it is the same one as in (D39), the question has already been answered by Al's (D39). If the horizon is enlarged, why so, and how far is it enlarged? Here is how non–maximality explains the results: (D39), read as homogeneous, is indeterminate because not all relevant antecedent worlds are consequent worlds. Still it is used by Al to convey that a contextually significant proportion of relevant antecedent worlds are consequent worlds (where the plate shatters). In contrast, once the exceptions (where the plate flies off) are made explicit as in (D40), homogeneity is precluded. The non-homogeneous '*would*' conditional negated in (D41) accordingly takes into account the exceptions, therefore it is false and thus to be negated.

I shall present a further piece of evidence. We use counterfactuals far more generously than one would expect from the standard account. One may dismiss these ways of using counterfactuals as loose and non-literal and hence irrelevant to a systematic account of counterfactuals. Having encountered a widespread tendency to do so, I do not want my case to depend on them. Still they might appear in a new light when taking into account non–maximality:

Teacher, having experienced that pupils sometimes start quarrelling when he leaves them alone, being asked whether to join for a coffee pause:
'I can't.
(D42) The children would quarrel if I left them alone.'

If (D42) is read literally, it provides evidence for a demanding non-maximal reading.[89] Only a certain proportion of relevant worlds where the children are left alone have to be worlds where

[89] When I introduced this example in an earlier version of this chapter, one referee wrote that (D42) is just false. Surprisingly, the referee did not express doubts that one might utter (D42) in the scenario considered. So the question is: given (D42) does not sound hyperbolic, metaphorical or otherwise non–literal ('the children would kill each other!'), how can we account for the teacher's use of (D42)?

they quarrel to influence the teacher's decision. Otherwise it would seem irresponsible for the teacher to utter (D42) literally. If one harbours doubt about the example, one might also ask oneself whether it is really so different from the (D39). Moreover, one should not forget that non–maximality arguably is a pragmatic phenomenon. We do not have to grant that (D42) is true independently of pragmatics, just that it is used to say something true.

I shall refrain from assembling further evidence for non–maximality and close this section with sketching an analysis of non-maximal conditionals. I have briefly summarized the most advanced proposal to base non–maximality in descriptions on homogeneity by Križ. I shall now provide an informal sketch what the analogue for conditionals might look like, though there may be other ways to flesh out my overall proposal. In Križ's account, the issue addressed comes with a partition of possible worlds. As for counterfactuals, we must be wary of confusing this partition with the possible worlds relevant to evaluating a counterfactual. I suggest a slight amendment of Križ's model. Normally, the issue is to find out how things are. Thus the relevant partition does not divide metaphysical but epistemic alternatives. Often the former can replace the latter, but here is a case where they cannot: assume the issue is whether a description 'the actual samples from the mine are gold' is true or not. Given the rigidifying 'actual', there are no metaphysical possibilities covering the two alternatives. There either is no metaphysically possible world where the actual samples are gold, or there is no metaphysically possible world where they are not. But there is an epistemic possibility that the actual samples are gold, and there is an epistemic possibility that they are not. Hence the relevant partition should be one of epistemic possibilities. In the same vein, the partition at issue when judging counterfactuals and claims to metaphysical modality is one of epistemic alternatives. But since we are to settle explicitly modal questions, the epistemic alternatives to be partitioned concern what the relevant metaphysical possibilities are. To put it otherwise: the worlds relevant to evaluating a counterfactual are metaphysically possible worlds (*m–worlds*), as it is usually assumed. But in order to figure out what the relevant worlds among the metaphysically possible worlds are, we have to consider several epistemic possibilities (*e–possibilities*) what the evaluation world viz. the actual world is like.

With this amendment in place, Križ's model can be transferred to counterfactuals: when a counterfactual is felicitously uttered, the contextual issue must come with a partition of e–possibilities where no cell contains both an e–possibility where all relevant antecedent m–worlds are consequent worlds and an e–possibility where none of them are. For the counterfactual to be used to assert something true, there must be a unique cell in the partition which contains only e–possibilities where sufficiently many relevant antecedent m–worlds are consequent worlds. I refrain from imposing the condition that there must be an e–possibility that *all* relevant antecedent m–worlds are consequent worlds. I do not see why we need this condition in the case of descriptions, and it will lead to unnecessary qualms in the counterfactual case: often we may well be in a position to rule out as an e–possibility that all relevant antecedent m–worlds are consequent worlds. The option of non–maximality would be of very limited avail if we could not use a counterfactual in that case. In sum, in certain contexts, a counterfactual can be used to say something true precisely if there is a unique cell in the contextual partition of epistemic possibilities which contains only epistemic possibilities where sufficiently many contextually relevant antecedent m–worlds are consequent worlds (sufficiently many being measured by some contextual threshold).

There are three things which are settled by context in this model: firstly, context determines the partition of salient epistemic possibilities (i.e. possibilities what the actual world is like). Secondly, context determines the relevant antecedent m–worlds for each of these e–possibilities. Applying the standard account of counterfactuals, these may be the antecedent m–worlds closest to the world from which the counterfactual is evaluated. Thirdly, context determines the threshold of how many relevant antecedent m–worlds have to be consequent worlds for the counterfactual considered to say something true.

I shall ponder in how far the proposal can be transferred to indicative conditionals. There are two difficulties. The first is that the debate on indicative conditionals does not converge towards a standard analysis. There are many competing approaches around. Just to give an example how the account might be transferred to indicatives, I shall settle for one exemplary proposal which is especially amenable to my treatment, but which I cannot properly defend here. The indicative conditional is interpreted by a necessity operator scoped over a material conditional.[90] In my version, the necessity operator is context-sensitive and ranges over epistemic possibilities. This proposal is attractive because it preserves on the one hand the connection to the material conditional; on the other hand it allows to add an additional aspect of contextual relevance, which may be used to avoid the unpleasant result that a conditional is true simply because its antecedent is false.

The second difficulty with indicative conditionals is specific to my approach: while in the case of a counterfactual, the contextual partition of *epistemic* possibilities (i.e. epistemically possible scenarios) could be kept separate from the relevant antecedent possibilities, interpreted as *metaphysical* possibilities, in the case of an indicative, I see no way of keeping them apart. One may bite the bullet and propose the following simple condition: an indicative conditional can be used to assert something true precisely if sufficiently many contextually salient antecedent e–possibilities are consequent possibilities. Again the threshold what counts as 'sufficiently many' is determined by context.

Non–maximality and grading: might, probably, definitely

I now come to a decisive step towards accounting for the lottery evidence. In the standard view, there is a fixed threshold which is either met or not: all relevant antecedent worlds have to be consequent worlds for a conditional to be true. The non-maximal reading at least sometimes requires a more differentiated take, making room for exceptions among the relevant antecedent worlds. Sometimes, for instance in assessing the plate counterfactual (D39), tolerable exceptions may simply not come into view. Yet at other times, for instance in the case of the smiling professors (D36), we may have to take what I call *the grading perspective*. In that case, grading involves three things: firstly to calculate the actual proportion of smiling professors among the most salient professors, secondly to figure out the contextual threshold for that proportion which allows (D36) to be truly uttered, thirdly to figure out whether the actual proportion meets the threshold.

While the non–maximality reading relies on implicit grading, it would be useful to have expressions which make grading explicit. I suggest that 'probably', used in the consequent of a conditional, is one of these expressions. I support my point by locating 'probably' on a scale of related expressions. All these expressions can at least sometimes be read as epistemic modals.[91] This may even be their primary meaning. In the context of a subjunctive, their contribution is peculiar. At one end of the scale is 'might'. 'Might', construed as the dual of 'would' (equivalent to 'not would not') displays a peculiar transition from an epistemic modal to some special use in counterfactuals: on the one hand, the expression is used as an epistemic modal to express claims to epistemic possibility. The basic idea is that, in uttering 'it might be that *P*', one conveys that there is an epistemic possibility that *P*. On the other hand, there is a genuine use in counterfactuals. 'Might' is plausibly construed as weakening 'would' as far as possible within the confines of semantic homogeneity, i.e. the range where a counterfactual is neither

[90] Cf. David Chalmers, 'Frege's Puzzle and the Objects of Credence', *Mind*, 120 (2011), 587–635; Jonathan Ichikawa, 'Quantifiers, Knowledge, and Counterfactuals', *Philosophy and Phenomenological Research*, 82 (2011), 287–313; Daniel Rothschild, 'Do Indicative Conditionals Express Propositions?', *Noûs*, 47 (2013), 49–68.

[91] Overview in Kai von Fintel and Anthony Gillies, '"Might" Made Right', in *Epistemic Modality*, ed. by Andy Egan and Brian Weatherson (Oxford: Oxford University Press, 2011), pp. 108–130.

true nor false: there are *some* relevant antecedent worlds which are consequent worlds. At the other end of the scale there is 'definitely'. 'definitely' also works as an epistemic modal. Roughly, 'definitely P' can be used to rule out the epistemic possibility of not–P. Again there is a genuine use in the context of a counterfactual: 'definitely' or '*would*' make plain that all relevant antecedent worlds without exception are consequent worlds.

I propose that 'probably' displays a perfectly analogous pattern. It has an epistemic meaning, but often its contribution to counterfactuals is peculiar. Endowed only with '*would*' and 'might', we would lack a device which makes explicit that a significant proportion of relevant antecedent worlds are consequent worlds. We require more than just that some relevant antecedent world is a consequent world, but we do not require that all of them are. Instead, the requirement is that some contextual threshold below 100% is met. 'Probably' serves the task. We get an order of counterfactuals according to their increasing strength (the stronger ones entailing the weaker ones):

(D43) If the plate had been dropped, it might have shattered.
(D44) If the plate had been dropped, it would probably have shattered.
(D45) If the plate had been dropped, it *would*/would definitely have shattered.

In the reading I propose, (D40) conveys that some relevant world where the plate has been dropped is a world where it shatters, (D44) conveys that *most* of them are, (D45) that all of them are. Concerning the plate scenario, (D40) is true but too modest, (D44) is true and perfectly informative, (D45) false provided the plate might have flown off sideways. All expressions considered also have other readings where 'might' and co. more clearly function as epistemic modals. I note that, although 'Would'/'definitely' as used in (D44) forms part of the grading scale, it at the same time works as a precisification of 'would' by removing homogeneity: the '*would*'–counterfactual is true precisely if all relevant antecedent worlds are consequent worlds. This observation will become significant.

Homogeneity, Non–Maximality and Lotteries
Grading Lotteries

In how far may the account developed in the last sections help us with lottery conditionals? I summarize the evidence to be explained. We reject

(D1) #If Anna had bought a lottery ticket, she would have lost.

We accept

(D4) If Anna had bought a lottery ticket, she would probably have lost.
(D5) If Anna hat bought a lottery ticket, it is 99.99 percent probable that she would have lost.

I begin with the standard truth–condition for (D1): all closest worlds where Anna buys a ticket have to be worlds where she loses. This explains why we reject (D1). However, we also have seen that the negation of (D1) behaves strangely:

(D25) ?It is not the case that Anna would have lost if she had bought a ticket.

The '*would*' version sounds better:

(D26) It is not the case that Anna *would*/would definitely have lost if she had bought a ticket.

I have taken the general contrast between 'would' and '*would*' as evidence for a homogeneous reading of counterfactuals. Where there is homogeneity, there might also be non–maximality: not all relevant antecedent worlds have to be consequent worlds, but some contextual threshold has to be met. While the non-maximal reading is inexplicit, there are means of explicitly *grading* the proportion of consequent worlds among relevant antecedent worlds. In contrast to (D39), the *lottery* feature of counterfactual (D1) precludes a non-maximal reading of (D1), at least as long as the lottery aspect is salient. Since any single ticket counts, there just is no contextual cut–off which privileges some threshold of sufficiently many tickets below 100% of the tickets. Still the asymmetry between (D25) and (D26) testifies to the presence of homogeneity. This can be explained as follows: in everyday counterfactuals like the plate counterfactual (D39) homogeneity and non–maximality prevail. Rarely will all relevant antecedent be consequent worlds. We expect non–maximality as the default case. This is why the denial of the '*would*'–version generally sounds better than the denial of the 'would'–version, even in the lottery case. Hence (D25) sounds worse than (D26). Lottery contexts impose additional demands on everyday reasoning. Non–maximality is not simply absent. It is to be ruled out by certain regimentations: firstly, one cannot simply rely on a rough and ready practice of ignoring exceptions. Secondly, one cannot simply rely on an implicit threshold what counts as 'close enough' to 100%. However, what is not excluded is the differentiated grading perspective as far as it works as follows: the proportion of consequent worlds among relevant antecedent worlds is *explicitly* graded. I have suggested that there are several expressions which allow grading, one of them being 'probably'. I have located 'probably' within a scale of related expressions. This scale can be applied to lottery counterfactuals:

(D46) If Anna had bought a ticket, she might have lost.
(D4) If Anna had bought a lottery ticket, she would probably have lost.
(D47) If Anna had bought a ticket, she *would* / would definitely have lost.

'Probably' here also relies on a contextual standard for what counts as sufficiently probable. But in contrast to the non-maximal reading of counterfactuals, it does so explicitly: (D4) is lexically different from (D1). We do not have a maximal and a non-maximal reading of the same sentence.

I draw a parallel to a descriptive lottery case: The following is marked

(D48) #The tickets will lose.

There is no non-maximal reading of (D48). Yet the following is fine:

(D49) Most of the tickets will lose.

Just as 'probably', 'most' also invokes a contextual threshold, but again it does so explicitly, in contrast to a description read non-maximally.

The results attained so far allow a more differentiated take on credences. I have rejected Schultz principle (Credence) and suggested that one should only settle for the linguistic evidence given by (D1) and (D4) I shall now present a more positive reaction to Schulz's claim that one should place high credence in (D1). Consider the following dialogue:

Dialogue9
Al: 'what is your credence that
(D1) Anna would have lost if she had bought a ticket?'
Bo: (D4) 'If Anna had bought a lottery ticket, she would probably have lost.
But

(D26) it is not the case that Anna would definitely have lost if she had bought a ticket.'

Instead of (D4), Bo may also use the more specific

(D5) If Anna had bought a lottery ticket, it is 99.99 percent probable that she would have lost.

The intuition that we should assign high credence to (D1) can be accounted for by a full explanation of Dialogue9.

Dialogue9 testifies to a certain vagueness in the request of telling what one's credence in (D1) is. One natural reaction is to give a differentiated set of answers which cover salient specifications of the request. The salient specifications can be derived from the two tendencies in our evaluation of (D1). Firstly, the lottery context drives us towards reading (D1) like a '*would*' counterfactual. Thus, one option of settling the request is to specify it as: is the requirement that *all* relevant antecedent worlds are consequent worlds satisfied? We have seen that denying the corresponding '*would*' counterfactual is preferred to denying (D1). In uttering (D26), Bo both clarifies the question (are ALL relevant antecedent worlds consequent worlds?) and answers it. However, uttering (D26) covers only one clarification.

If Bo were only to utter (D26), her reaction would be somewhat uncooperative. Bo uses '*would*' to clarify the issue addressed. But there are other clarifications to heed. Bo's utterance is naturally supplemented by a different way of precisifying the request for credences. The regimentation that comes with a lottery context excludes non–maximality but leaves the option of explicit grading. The proportion of consequent among relevant antecedent worlds is determined and ranked according to some scale, e.g. 'might', 'probably', and 'definitely'. We have seen that all these expressions allow a transition from an epistemic modal to a peculiar use in grading the proportion of consequent among closest antecedent worlds. I propose that there is a similar transition for credences from the epistemic realm to the grading perspective, which requires to measure the proportion of consequent among closest antecedent worlds. This transition accommodates the natural tendency to read a request for credences as a request for *counting* by a suitably fine-grained measurement scale.

'What is your credence?' is naturally understood as a request to *count.* 'Credence' is a mass term amenable to a 'how much' question: 'what credence' is naturally read as 'how much credence'. There is a transition from such a 'how much' question to a 'how many' question by including a unit of measurement:[92] how much credence, measured by the most salient measurement scale, i.e. how many percent credence do you assign to (D1)? We have got accustomed to this paraphrase, at least in a philosophical context. However, to provide an answer to the question thus reformulated, we have to come up with a suitable measurement scale. Even when we are somewhat clueless about the most salient scale, we are very willing to accommodate the request by looking for a scale in the neighbourhood. The scale made salient by the grading perspective is the proportion of consequent among the relevant antecedent worlds.

Consider again the pair:

(D4) If Anna had bought a lottery ticket, she would probably have lost.
(D5) If Anna hat bought a lottery ticket, it is 99.99 percent probable that she would have lost.

In Dialogue9, Bo asserts (D4) in order to convey that she has counted as requested and found that the proportion of consequent worlds among closest antecedent worlds is high according to some contextual standard.

[92] Karin Koslicki, 'The Semantics of Mass–Predicates', *Noûs,* 33 (1999), 46–91 (p. 75).

The intuition that we ought to assign high credence in (D1) is explained in the same way as Bo's asserting (D4) in Dialogue9. In both cases, the grading perspective makes the proportion of consequent among closest antecedent worlds the most eligible scale of measurement. By using (D4), one expresses the same as when one says that one's credence in (D1) should be high. I have noted that Bo may also use (D5) instead of (D4) in Dialogue9:

(D5) If Anna had bought a lottery ticket, it is 99.99 percent probable that she would have lost.

I propose that by asserting (D5) one expresses the same as when one says that one's credence in (D1) is 99.99%: one has counted the consequent worlds among the relevant antecedent worlds as requested and found the proportion to be 99.99%. The explanation can be easily transferred to related intuitions about indicative conditionals.

A Non–Standard Notion of Credence?

There is one important doubt about my interpretation of credence. Credence is normally understood with regard to the standard case of belief in some candidate for actual truth. Rational credence in P should reflect in how far one's evidence supports *P*. There is a close connection between credence *tout court* and credence given one's evidence. Both in turn are closely related to conditional probability. There are principles which spell out the connection, most prominently Lewis's (Principal Principle): roughly, one's credence in P should equal the objective probability of P given one's total evidence. It is a key requirement for any take on credences in counterfactuals that it be integrated into this overall picture. I note that eventually my proposal and the main other account to yield high credence in (D1), Schulz's arbitrariness account, lead to the same result: we should count the proportion of consequent worlds among the relevant antecedent worlds. I arrive at this result not quite as straightforwardly as the arbitrariness account.

In Schulz's account, the link between the standard picture of credence and the arbitrariness semantics comes about in two steps: firstly, credence in a counterfactual should equal the objective probability that the counterfactual is true. This step is intuitively plausible, just as the Principal Principle is. Secondly, the objective probability of a counterfactual is determined by the proportion of consequent worlds among the relevant antecedent worlds. This move in turn is explained by the semantics: the probability that the arbitrarily selected relevant antecedent world is a consequent world equals the proportion of consequent worlds among the relevant antecedent worlds. It is an advantage of the arbitrariness account that it can preserve this close connection to the chance–credence link and that it does so well in motivating the interpretation of probability by the proportion of consequent among the closest antecedent worlds.

While the high credence in (D1) as attained by counting consequent worlds among the relevant antecedent worlds is a straightforward consequence of the arbitrariness account (which is tailored to yield this result), I have to present the analogous move as a constructive proposal for how to deal with the request of telling one's credence in (D1). This constructive proposal selects the most salient measurement scale, which is given by the grading perspective. The grading perspective leads to a peculiar interpretation of credences in counterfactuals. However, this interpretation is motivated by the parallel to the reinterpretation of other epistemic expressions, in particular epistemic modals. These expressions permit a non-epistemic reading in counterfactual contexts. They are used to make explicit the grading perspective: some, most, all relevant antecedent worlds are consequent worlds. As long as the non-epistemic reading of such modals is granted, it can motivate a perfectly analogous transition for credences.

Summarizing, I have presented a relevant alternative to radically revisionary semantics. There is independent evidence that, just as descriptions, conditionals display homogeneity and non–maximality. These features can be used to explain the puzzles about lottery conditionals.

2.4 Problems with Similarity

In the sections to come, I shall develop several problems with Lewis's similarity ordering. The first of these problems concerns so-called Morgenbesser cases, the other problems concern the notorious future similarity objection. Morgenbesser cases allow me to apply results from the discussion of lottery cases to the discussion of similarity as they feature probabilistic outcomes such as coin tosses.

2.4.1. Morgenbesser Case

Morgenbesser's Coin is a notorious counterexample to the way Lewis supplements the standard semantics of counterfactuals by a similarity ordering. After taking issue with a recent attempt to dismiss the intuition, I discuss the two outstanding attempts at a solution in broadly Lewisian terms. Paul Noordhoff argues that facts which probabilistically depend on the antecedent should not count towards similarity. Jonathan Schaffer argues that facts which causally depend on whether the antecedent obtains or not should not count. I show that their discussion ends in a stalemate, thereby also fending off criticisms by Wong. None of them succeeds at refuting the other. Moreover, I present variants of the original Morgenbesser case which evade both solutions.

I repeat the main ingredients of Lewis's analysis:

A counterfactual P>>Q is non-vacuously true iff some P&Q–world is more similar to the actual one than any P&not–Q–world.

For simplicity, I will talk as if there were a set of closest P–worlds. For the deterministic case, Lewis presents a default similarity ordering of worlds:

'(1) It is of first importance to avoid big, widespread, diverse violations of law [big miracles].
(2) It is of second importance to maximize the spatio-temporal region throughout which perfect match of particular fact prevails.
(3) It is of third importance to avoid even small, localized simple violations of law [small miracles].
(4) It is of little or no importance to secure approximate similarity of particular fact, even in matters that concern us greatly.'[93]

Under indeterminism, it is of first importance to avoid *amazing* patterns of particular matters of fact (quasi–miracles).

Enters Morgenbesser (I slightly vary Jonathan Schaffer's presentation):

Morgenbesser's Coin
A coin is tossed. At indeterministic w_0, while the coin is in midair, Lucky bets heads. The coin lands tails, so Lucky loses. The following counterfactual seems intuitively true at w_0:

(E1) If Lucky had bet tails, he would have won.

Lewis analysis entails that the relevant counterfactual is false: Lucky merely *might* have won. To see this, compare the (Lucky–bets–tails&coin–lands–tails)–world w_T, with the (Lucky–bets–tails&coin–lands–heads)–world w_T. w_T and w_L will come out equidistant from w_0. Each costs perfect match with actuality from Lucky's bet on, and each buys an aspect of imperfect match – w_T preserves the outcome of the flip (tails), while w_H preserves the outcome of the bet (unlucky).[94]

As Lewis's argued in discussing the Nixon counterfactual (A33), it would take a widespread cover–up action to erase all traces the antecedent event has left. At least in the light cone of Lucky's bet (in a relativistic world), there is no *perfect* match in particular matters of fact to be had. And since the outcome of the coin toss, albeit not influenced by the bet, lies within this region, the decision has to draw on *imperfect* match. But then a world w_T which

[93] Lewis, *Counterfactual Dependence*, pp. 47–48.
[94] Jonathan Schaffer, 'Counterfactuals, causal independence and conceptual circularity', *Analysis*, 64 (2004), 299–309 (p. 300).

preserves the outcome of the flip (tails) does not fare better than a world w_L which preserves the outcome of the bet (Lucky loses). In sum, Lewis's criteria of similarity cannot explain our intuition that Lucky would have won.

Is the Example Coherent?

Recently doubts have been voiced as to whether the Morgenbesser argument is coherent. Ian Phillips has argued that it presupposes indeterminism while the intuition depends on 'closet determinism': without determinism, one has no reason to accept Morgenbesser counterfactuals, says Phillips.[95] In indeterministic worlds, nothing ensures that Lucky would have won.

Phillips' objection is that the Morgenbesser case is incoherent. The incoherence is due to combining the implicit presupposition of 'closet determinism' and the assumption of indeterminism. The latter assumption is required for Morgenbesser cases to pose a challenge for Lewis's view. For under determinism it could be easily explained why we hold onto the outcome of the coin toss (tails). It is predetermined by the facts and laws at w_0, which are preserved at w_T but not w_H. To be sure, there must be some changes in facts and laws compared to w_0 to implement the antecedent. But intuitively they do not interfere with the way the coin toss comes about.

This argument is not convincing, though. Arif Ahmed has retorted that the Morgenbesser intuition just depends on our tendency to hold fixed *factors which do not causally depend on the antecedent* in counterfactual reasoning.[96] This explanation is intuitive without closet determinism. Morgenbesser intuitions are deeply entangled with the role of counterfactuals in everyday life. As psychological evidence shows, one core role of genuine counterfactuals is to evaluate past actions one could have taken such as to trigger regret or relief ('If Lucky had only bet tails, he would have won').[97] It is part and parcel of this practice to mark the agent's own contribution to the way things have gone. To do so, one must draw a line between facts to which the actions considered would have made a difference and other facts.

In evaluating Lucky's betting activity, we distinguish things to which his alternative options would have made a difference from things to which they would not have made a difference. Lucky's betting behaviour had no influence on the coin toss. But keeping fixed the coin toss, his betting behaviour had an influence on the outcome of the bet. So he might regret not to have bet tails. Yet he cannot regret (at least strictly speaking) but only, say, lament that the coin fell tails. Counterfactuals are a tool of tracking this difference. This claim can be integrated into a more general view endorsed by many philosophers: in evaluating a counterfactual, we are interested in a scenario which 'makes the antecedent true without gratuitous departure from actuality'.[98] The Morgenbesser intuition draws on our intuitive ways of telling apart gratuitous from non-gratuitous departures. One challenge remains: 'Ahmed must justify a closeness metric which delivers this result.'[99] The Lewisian default ordering is the outstanding candidate for such a metric.

Phillips himself would reject Ahmed's intuitive rationale of counterfactual reasoning. He heralds a suppositional analysis of counterfactuals. Roughly, a counterfactual P>>Q is true or assertable iff the conditional probability of Q given P was sufficiently high at a suitable point

[95] Ian Phillips, 'Morgenbesser cases and closet determinism', *Analysis*, 67 (2007), 42–49.
[96] Arif Ahmed, 'Out of the closet', *Analysis*, 71 (2011), 77–85; Ian Phillips, 'Stuck in the closet: a reply to Ahmed', *Analysis*, 71 (2011), 86–91 (p. 87).
[97] James Olson and Neil Roese, 'A critical overview', in *What might have been. The social psychology of counterfactual thinking*, ed. by James Olson and Neil Roese (Mahwah: Lawrence Earlbaum Associates, 1995), pp. 1–57.
[98] Lewis, *Counterfactual Dependence*, p. 41.
[99] Phillips, 'Stuck in the Closet', p. 88.

in time.[100] Yet I think that even given Phillips' own account, Morgenbesser intuitions can be upheld.

To Phillips, the problem for Morgenbesser intuitions arises from their combining determinism and indeterminism. Determinism is needed to support Morgenbesser inuitions, indeterminism is needed for them to spell trouble for Lewis. However, the inconsistency does not arise if Morgenbesser intuitions are forwarded as a challenge to the suppositional account. Grant that they depend on assuming determinism. We do not need to asume indeterminism for them to pose a challenge to the suppositional semantics. The intuition is that, since the coin fell tails, Lucky would have won had he bet tails. Phillips's suppositional semantics cannot accommodate this as the prior possibility of Lucky winning given he bets tails is not high enough to accept that he would have won. One might still object that Morgenbesser intuitions depend on determinism. But the suppositional account in turn should also apply to a deterministic world, especially if our everyday use of counterfactuals has been developed within a climate of closet determinism (as Phillips must assume).

Having dissolved Phillips's criticism, I shall now discuss several proposals to mend Lewis's criteria of similarity such as to accommodate Morgenbesser intuitions. In confining my discussion to amendments of Lewis's standard semantics, I shall not take into consideration solutions not based on a closeness semantics.[101]

Paul Noordhoff has espoused the following solution: disregard match in facts which probabilistically depends on the antecedent. His solution can be inscribed as an amendment into Lewis's metrics:

(1)…
(2´) It is of second importance to maximize the spatio-temporal region throughout which perfect match of particular facts prevails *as far as these facts are probabilistically independent of the antecedent*.
(3)…
(4´) It is of little or no importance to secure approximate similarity of particular facts *as far as these facts are probabilistically independent of the antecedent*.

Noordhoff's notion of probabilistic independence is the following: a fact F is probabilistically independent of the antecedent iff, for any time t, its probability is the same in the closest world where the antecedent obtains and the closest world where it does not. This criterion is not circular provided we can figure out what the probabilities of F are in the antecedent worlds at stake without already knowing whether F is probabilistically independent.

The outcome of the toss is probabilistically independent of Lucky's betting, says Noordhoff, and thus not probabilistically independent of the antecedent. Here is my interpretation: the closest world where Lucky does not bet *tails* is the actual one. Until the coin has fallen, P(tails) = P(heads) = 0.5; afterwards, P(tails) is = 1 and P(heads) = 0.[102] The same goes for the closest world where Lucky bets *tails*. But, Noordhoff says, the same does not go for the outcome of the bet (Lucky wins). The outcome is not probabilistically independent of the antecedent, i.e. whether Lucky bets or not. Thus w_T beats w_H.

[100] Phillips, 'Morgenbesser', p. 43.
[101] E.g. Stephen Barker, 'Counterfactuals, probabilistic counterfactuals and causation', *Mind*, 108 (1999), 427–69; Eric Hiddleston, 'A causal theory of counterfactuals', *Noûs*, 39 (2005), 632–657. Yet see below my criticism of Barker's intuitions.
[102] Paul Noordhoff, 'Morgenbesser's coin, counterfactuals and independence', *Analysis*, 65 (2005), 261–263 (p. 262), cf. Paul Noordhoff, 'Prospects for a counterfactual theory of causation,' in *Cause and chance: causation in an indeterministic World*, ed. by Paul Dowe and Paul Noordhoff. (London: Routledge, 2004), pp. 188–201 (p. 193).

I shall consider two criticisms of Noordhoff's approach. The first is due to Chiwook Won. Won has argued that approaches to counterfactuals in terms of probabilistic independence cannot deal with certain variants of Morgenbesser Cases. I summarize Won's Morgenbesser cases, Paul Noordhoff's probabilistic independence solution, and Won's criticism. Then I show why the criticism fails. I close this part of my discussion with a real problem for Noordhoff.

Morgenbesser cases are core examples which a semantics for counterfactuals must deal with. Just as I have done at the beginning, Won distinguishes approaches in terms of probabilistic (Noordhoff) and approaches in terms of causal independence (Schaffer). The idea of the former is to hold onto facts which are probabilistically independent of the antecedent, the idea of the latter is to hold onto facts which are causally independent. Won's main claim is that approaches in terms of probabilistic independence fail. He focuses on Noordhoff's exemplary account.

Here is Won's example:

'Susan offers Lucky a bet on an indeterministic coin toss. Lucky bets heads and Susan tosses the coin. But the coin lands tails. Now consider a counterfactual:
[E2] If Lucky had tossed the coin, it would still have landed tails.
Intuitively, this is false[…]
[E3] If Lucky had bet tails, he would have won.
Intuitively, this is true.'[103]

Won uses a somewhat simplified version of Noordhoff's criterion of probabilistic independence:

'B is probabilistically independent of A just in case:
[E4] If A were to occur, the chance of B's occurring would be x.
[E5]If A were not to occur, the chance of B's occurring would be y.
[…] x = y.'[104]

The criterion underpins (E3), says Won, but only provided we subscribe to (strong) centering: the actual world where Lucky bets heads is closer than any other world where Lucky does not bet tails. In the closest worlds where Lucky does not bet tails (A does not occur), the actual one, the prior probability of the coin landing tails (the chance of B occurring) is 0.5. And the same goes for the closest worlds where Lucky bets tails (A does occur). The coin landing tails is probabilistically independent of Lucky betting tails. Thus, it is held fixed and (E3) comes true. But alas, says Won, centering cannot be upheld in Noordhoff's account. For otherwise (E2) would come true, counterintuitively. Given centering, the closest worlds where Lucky does not toss the coin are worlds where Susan tosses the coin. The probability of the coin falling tails is the same as in any relevant world where Lucky tosses it (0.5). The coin falling tails is probabilistically independent of the antecedent, and thus should be held onto. We have a dilemma: without centering, no (E3); with centering, (E2).[105]

The argument fails for two independent reasons. Firstly, Noordhoff does not need centering to get (E3).[106] Assume the closest world where Lucky does not bet tails is not the actual one but one, say, where Lucky does not bet at all. This assumption has no impact on the probabilistic independence of the tails outcome. At that world, too, the tails outcome is probabilistically independent of the antecedent: the fair coin is tossed and the probability of the

[103] Chinook Won, 'Morgenbesser's Coin, Counterfactuals, and Causal vs. Probabilistic Independence', *Erkenntnis*, 71 (2009), 345–354 (p. 346).
[104] Won, 'Morgenbesser's Coin', p. 349.
[105] Won, 'Morgenbesser's Coin', p. 351.
[106] And he seems committed to rejecting centering, judging from his reply to Schaffer (cf. Schaffer, 'Counterfactuals', p. 307, Noordhoff, 'Morgenbesser's Coin', p. 261)).

coin landing tails is 0.5.[107] According to Noordhoff, we uphold that the coin falls tails in the closest worlds where Lucky bets tails, irrespectively of whether centering obtains. And that entails his winning in a world where he bets tails. Thus, we get (E3).

Secondly, Won is wrong that centering would commit Noordhoff to (E2). Won just attends to the probability of the coin falling tails *before* the coin has fallen (= 0.5 at relevant worlds where the coin is tossed). But we must also attend to the probability *after* the coin has fallen (1 or 0). Single–case probabilities may vary in time. In assessing probabilities at different worlds according to Noordhoff's criterion, we must fix a set of points in time common to the different worlds where probabilities are evaluated. To achieve probabilistic independence à la Noordhoff, the probability in the different worlds must be the same for *any* time of evaluation. As for (E2), clearly in all relevant worlds, the probability of the fair coin falling tails is 0.5 before the coin has fallen. But *after* the coin has fallen, this does not hold. In the closest worlds where Lucky does not toss the coin, i.e. the actual world where Susan does instead (by virtue of centering), the probability of tails becomes 1. But the same does not have to go for any closest world where Lucky tosses the coin. For any such world, surely the probability of tails becomes 1 or 0. But since we cannot presuppose that the closest worlds where Lucky tosses the coin are worlds where the coin falls tails, there is no reason to deem a (Lucky tosses)–world where the probability of tails would be 1 (the coin fell tails) closer than a (Lucky tosses)–world where the probability of tails would be 0 (the coin did not fall tails).

Thus, there is no reason for Noordhoff to accept

(E6) If Lucky had tossed the coin, after the coin has fallen, the probability of tails would be 1.

But (E6) would have to be true for the tails outcome to be probabilistically independent (à la Noordhoff) of the antecedent (Lucky tosses the coin). Condition P3 of probabilistic independence is not satisfied. There is no reason to hold onto the tails outcome, i.e. the consequent of (E2). Thus, Won is wrong that centering commits Noordhoff to (E2).

The point can also be used to defend Noordhoff against a criticism of Schaffer. Schaffer notes in passing that the outcome of the bet (Lucky wins), too, is probabilistically independent of Lucky betting tails or not betting tails.[108] In the closest world where Lucky bets tails, before the coin has fallen, P(winning) = 0.5. The same goes for the closest world where Lucky bets heads.

Noordhoff tries to rebut Schaffer's remark: there is not only the alternative of betting heads but also the alternative *not to bet at all*.[109] Noordhoff seems to have in mind the following: Provided there is a relevant point in time t when the latter alternative has non–zero probability, the outcome of the bet probabilistically depends on whether Lucky bets tails or not. In a world where he bets tails, from his betting onwards the chance of winning is 0.5, in a world where he does not, it is *smaller* than 0.5. Noordhoff's reply is mistaken. In adopting Lewis's metrics for his test of probabilistic independence, he adopts centering: no world is closer to the actual world w_0 than w_0 itself. According to Noordhoff's criterion, we must consider the world closest to w_0 where Lucky does not bet tails. And since Lucky actually does not bet tails *but heads*, the closest world to w_0 is a world where Lucky bets *heads*, not a world where he does not bet at all. Thus the alternative of not betting at all is irrelevant. If there is a probability that Lucky does not bet at all, it is the same in the closest world where he bets tails and w_0.

Nevertheless Schaffer is mistaken, too (but for a different reason): the outcome of the bet (Lucky wins) does probabilistically depend on the antecedent (Lucky bets tails). Schaffer grants that the outcome of the coin toss (tails) is probabilistically independent of the antecedent.

---

[107] In a moment, we will see that this assumption of Won's is too simplified. We must also attend to what happens after the coin has fallen. But the assumption works *ad hominem* Won.

[108] Schaffer, 'Counterfactuals', p. 307.

[109] Noordhoff, 'Morgenbesser's Coin', p. 261.

For any time *after* this outcome is fixed, in the closest world where Lucky bets tails (w_T), P(Lucky wins) = 1. In the closest world where he does not bet tails, the actual world in which he bets heads (w_0), P(Lucky wins) = 0.

While I think that Won's and Schaffer's objections fail, I close with a more grievous objection to Noordhoff's account: consider

(E7) If Lucky had bet tails and the coin had been tossed, the coin would have fallen tails.

Arguably this counterfactual is true if (E3) is.[110] In Noordhoff's account, for us to hold onto the result of the coin toss (tails), it must be probabilistically independent of the antecedent: the probability of tails must be the same in the closest worlds where Lucky bets tails and the coin is tossed and in the closest worlds where it is not the case that Lucky bets tails and the coin is tossed, i.e. the actual one. After the toss, the probability is the same (= 1) just if we hold onto the result tails in the closest worlds where Lucky bets tails and the coin is tossed. But we hold onto the result tails just if it is probabilistically independent of the antecedent. Thus, the criterion is viciously circular.

What if centering is given up? In that case, things get more complicated. Consider the closest worlds where the antecedent does not obtain: if they are worlds where Lucky does not bet tails and the coin is *not* tossed, the tails result is not probabilistically independent and (E7) comes out false. And if the closest worlds are worlds where Lucky bets tails but the coin is not tossed, again (E7) comes out false. If the closest worlds are worlds where Lucky does not bet tails but the coin is tossed (the actual one), the vicious circularity arises. If worlds of all three kinds are equally close, we cannot claim the tails result to be probabilistically independent. Again (E7) is wrong. Noordhoff could restrict the time of evaluation to points in time before the result of the toss is fixed. But then his theory would fall prey to Schaffer's objection.

To bring out a further problem, I vary an example of Stephen Barker's:[111]

Morgenbesser's Coin II:
Mandrake the Magician has placed magnets such as to influence the coin toss. When Lucky bets *heads*, there is a probability of 5% that Mandrake manipulates the outcome of the coin toss such that Lucky loses.[112] Actually Lucky bet heads but Mandrake did not interfere.

There was a point in time after Lucky's bet when there was a chance of Mandrake interfering. We still accept that Lucky would have won if he had bet *tails*. But the outcome of the toss is no longer probabilistically independent of whether he bet tails or not (i.e. heads at w_0). For in the latter but not in the former case there was a chance of Mandrake interfering. Thus, at some point in time, at w_0 the chance of Lucky winning was 0.5 * 0.95, while at w_T it was 0.5. In sum, while Noordhoff evades Schaffer's objection, he cannot evade the other problems.

Schaffer presents a different solution to Morgenbesser problems: in maximizing perfect match in facts, disregard match which causally depends on whether the antecedent obtains or not. The purported advantage of this solution is that it dissolves a great number of problem cases, including the Morgenbesser case. Lewis metrics is amended:

---

[110] Cf. Lee Walters, 'Morgenbesser's Coin and Counterfactuals with True Components', *Proceedings of the Aristotelian Society*, 99 (2009), 365–379 (p. 370).
[111] Barker, 'Counterfactuals', p. 431.
[112] Barker puts this example to a different use: in his version, when Lucky bets tails, there is a 5% chance of Mandrake interfering such as to change the outcome of the toss to heads. Barker thinks that if Lucky had bet tails, the probability of his winning would have been 95%. I do not share this intuition. Presumably the mere chance of Mandrake interfering is sufficient for us not to hold onto the outcome of the coin toss, even in cases where Mandrake does not interfere (cf. Kment's discussion of *Nixon's Game* below). So the probability of Lucky winning would have been 0,95 * 0,5.

'(2c) It is of the second importance to maximize the region of perfect match, *from those regions causally independent of whether or not the antecedent obtains.*
...
(4c) It is of the fourth importance to maximize the spatiotemporal region of approximate match, *from those regions causally independent of whether or not the antecedent obtains*.'[113]

The outcome of the bet causally depends on whether Lucky bets tails or not. The outcome of the coin toss does not. Thus, w_T comes out closer than w_H. Morgenbesser's counterfactual is true.

Schaffer's solution has its own problems, though. Noordhoff disagrees with Schaffer's analysis. I present a variant of his counterexample (which varies Tichy's hat–example):

Fred's Hat:
Fred has a reliable disposition to wear his hat. Sometimes weather conditions interfere with this disposition. There is a chance–device in his brain that makes him attend to the weather in 50% of the cases. Whenever he attends to the weather and the weather is fine, he takes off his hat. Actually he wears his hat and it is raining. We reject

(E8) If it had not been raining, Fred would have worn his hat.[114]

The rainy weather does not cause Fred to take off his hat. Moreover, consider the closest counterfactual situation where it is not raining *and* he still wears his hat: in this situation Fred wearing his hat does not seem to depend on the weather conditions either. He just does not attend to the weather. Moreover, this situation preserves more match in particular matters of fact than a situation where it is not raining and Fred does not wear his hat. According to Schaffer's criterion, we should hold onto the fact that Fred wears his hat. Thus, we should accept (E8).

Noordhoff surely is right that Schaffer's notion of causal dependence needs interpretation. I propose to use the account of Boris Kment, which is very close in spirit to Schaffer but evades Noordhoff's criticism. The upshot of Kment's account is that match in particular facts should only count towards similarity of worlds as far as these facts have the same explanation.[115] And as far as the relevant aspects of explanation in cases like Morgenbesser's Coin boil down to causal explanation, we may take Schaffer and Kment as aiming at the same criterion: a fact is causally independent of whether the antecedent obtains or not precisely if it has the same explanation in the closest worlds where the antecedent obtains and in the actual world.[116]

In how far does Kment help us to evade (Fred's Hat)? Kment more extensively than Schaffer deals with the question how to confine explanatory history. He presents an argument for a very broad conception of relevant causal–explanatory factors:[117]

Nixon's Game
Assume Nixon's missile system is indeterministic. There is a chance that the signal fizzles out and there is a chance that a nuclear holocaust occurs. We reject

(E9) If Nixon had pressed the button, the signal would have fizzled out.

[113] Schaffer, 'Counterfactuals', p. 305.
[114] Noordhoff, 'Morgenbesser's Coin', p. 262.
[115] Cf. Boris Kment, 'Counterfactuals and Explanation', *Mind*, 115 (2006), 261–310 (p. 296).
[116] I disregard Kment's preoccupation with counterlegals, which is not relevant to the case.
[117] Kment, p. 299.

There is a close parallel between (Nixon's Game) and (Fred's Hat). For (E9) to be false, worlds where Nixon presses and a nuclear holocaust occurs must be at least as close as worlds where he presses and the signal fizzles out. But in the closest worlds where the signal fizzles out, many actual matters of fact obtain unaltered which would be affected by the holocaust. At some point in time, there is an unrealized chance that Nixon's pressing affects these facts.[118] Kment suggests that this unrealized chance amounts to a difference in the explanatory history of the facts which would be affected by a holocaust, compared to their *actual* explanatory history (without Nixon's pressing, there is (almost) no chance of a holocaust). The same goes for (Fred's Hat): for (E8) to be false, the closest worlds where it is not raining must not all be worlds where he wears his hat. So the closest worlds where it is not raining must split into worlds where Fred does not attend to the weather and wears his hat and worlds where Fred attends to the weather and does not wear his hat. In the former worlds, at some point in time there is an unrealized chance that the weather conditions catch Fred's attention and he does not wear his hat. Given Kment's broad reading of explanatory history, this unrealized chance amounts to a difference in the causal–explanatory history of Fred wearing his hat. Thus, Fred wearing his hat does not count towards similarity.

In sum, the discussion between Schaffer and Noordhoff ends in a stalemate. None successfully rebuts the other. Yet Schaffer's approach can be modified along the lines of Kment such as to evade Noordhoff's counterexample. I see no parallel modification for Noordhoff's account.

This is not the end of the story yet. As I have presented a counterexample to Noordhoff's approach, I will present two new counterexamples, which spell trouble for the Schaffer–Kment approach.[119] It seems that some non-producing factors must not count towards relevant causal–explanatory history. I vary an example owing to Schaffer and Kment:[120]

The King's Coin:
The king tosses a coin. The evening before, the king's enemy has placed a bomb under the king's throne. The detonating mechanism is refined: there is a box. In the box, there is a clock which activates the bomb some time before the coin toss unless it is stopped. The box is causally isolated save for its connection to the bomb. The clock within the box is built such that two independent signals within the box, x and y, are each sufficient to stop it. Each signal can be activated or deactivated by a minuscule chance event. Signal x occurs but y does not. Later the box is cleanly disposed of.[121] We accept, Kment and Schaffer say,

(E10) If y instead of x had occurred, the outcome of the coin toss would have been the same.

---

[118] Cf. Kment, p. 300.

[119] Further criticism has been forwarded by Walters, 'Morgenbesser's Coin'. To evade Walters's criticism, one might simply restrict the clause 'from those regions...' in Schaffer's amended similarity metrics to genuine contrary–to–fact antecedents (or to the contrary–to–fact part of such antecedents). As a consequence, facts which causally depend on the antecedent as far as it actually obtains *do* count towards similarity.

[120] Cf. Kment, p. 300.

[121] The box does not have to disappear without a trace. But differences within the box must not make a difference to subsequent history over and above activating or not activating the bomb. In devising the causally isolated box, I follow Wasserman (Ryan Wasserman, 'The future similarity objection revisited', *Synthese*, 150 (2006), 57–67 (p. 59)). I need the box to achieve perfect future match of facts in accordance with Lewis's criterion (2). We might doubt the nomic possibility of a causally isolated box and of cleanly destroying it. But firstly, the possibility of a causally or energetically isolated box is used in the philosophy of physics (cf. Laurence Sklar, 'Causation in statistical mechanics', in *The Oxford handbook of causation*, ed. by Helen Beebee, Christopher Hitchcock, Peter Menzies (Oxford: Oxford University Press, 2009), pp. 661–672 (p. 669)). Secondly, Wasserman himself concedes that his example is nomologically impossible; but he insists that this does not disqualify it as a counterexample to Lewis (Wasserman, p. 65). Schaffer rejects doubts that the example is too far–fetched to trigger reliable intuitions (Schaffer, 'Counterfactuals', p. 302). So he must accept the (King's Coin) scenario.

Kment concludes that y instead of x occurring must not count towards a difference in the explanatory history of the coin toss. The problem is the following: relevant explanatory history would have to be tailored such as to reconcile the opposing demands imposed by (E9) and (E10). Yet it is doubtful that Kment can evade the problem by tailoring explanatory history. For (E10) to come true, the explanatory history of the coin toss must be the same no matter whether x or y occurs. Now consider

(E11) If x had not occurred, the outcome of the coin toss would have been the same.

(E11) seems intuitively wrong. However, Kment insists that y replacing x does not make a difference to the causal–explanatory history of the coin toss. Thus, one candidate outdoes all other worlds where x does not occur in terms of match in facts and needs nothing more than two minuscule chance events: y replaces x in the history of the coin toss.[122] We gain perfect match with the complete future of the actual world, including the outcome of the coin toss. So according to both Lewis's metrics and the Schaffer–Kment amendment, (E11) should come out true. The question is: why do we not imagine signal y to take over if x fails to occur?

Consider Eric Hiddleston's variant of the Mandrake case, originally directed against Lewis's metrics:[123]

Morgenbesser's Coin III:
Again Mandrake has placed magnets such as to influence the coin toss. When Lucky bets, there is a probability of 5% that Mandrake manipulates the outcome of the coin toss such that Lucky loses. Actually Lucky does *not* bet. The outcome is tails. We accept

(E12) If Lucky had bet tails and Mandrake had not interfered, Lucky would have won.

The problem this example poses to the Schaffer–Kment template is captured by Hiddleston's gloss: Lucky's not betting is an actual cause of the coin landing tails. (Lucky's not betting is a cause of Mandrake's not activating the magnet, and that is a cause of the coin landing tails.)[124]

Kment's broad conception of explanatory history underpins Hiddleston's diagnosis: Lucky's not betting *actually* forms part of the causal–explanatory history of Mandrake's not intervening and thus of the coin landing tails. Since this part of the causal–explanatory history of the coin landing tails would have to be different in the counterfactual situation (Lucky's *betting* would replace it), Schaffer and Kment provide no reason to hold onto the result *tails* and thus should reject (E12).

In conclusion: Morgenbesser's Coin can be given a twist that eludes all hitherto known Lewisian attempts at turning the game around. Does this show the principled limits of such accounts? I don't think so. For instance, one might try to fix the Schaffer–Kment account by using a context-sensitive *contrastivist* notion of causation.[125] Here I can only give a hint: causal dependence or relevance is determined with respect to a context-sensitive causal contrast. The causal contrast is sensitive to the minimal context created by the antecedent. The difference between (E10) and (E11) is explained by the difference between the causal contrasts invoked. As for (E10), signal x occurring is contrasted to signal y occurring instead. Intuitively, this is causally irrelevant to whether the coin toss is the same. Thus, we do hold onto the outcome of the coin toss. As for (E11), signal x occurring is contrasted to no signal occurring at all. Intuitively, this is causally relevant to whether the coin toss is the same. Thus, we do not hold

[122] The example can be varied such as to require only *one* chance event: Assume that among the different minuscule chance events that could undo x there is one which also triggers y. I think we still reject (E11).
[123] Cf. Hiddleston, 'Counterfactuals', p. 637.
[124] Adapted from Hiddleston, 'Counterfactuals', p. 637.
[125] Jonathan Schaffer, 'Contrastive Causation', *The Philosophical Review*, 114 (2005), 297–328.

onto the outcome of the coin toss. What concerns (E12), the causal contrast might be restricted to the contrary–to–fact part of the antecedent.[126] Lucky betting tails may be contrasted to Lucky not betting at all. Again this does not make a difference to the outcome of the toss. Of course, a lot more would have to be said how the relevant causal contrast is confined.

Pending such a more in-depth assessment, I still deem the Schaffer–Kment approach the most promising way of dealing with problems like Morgenbesser cases and the problems with convergence worlds to be discussed in the next section.

[126] This restriction might be explained as follows within the context of the example: the contrary–to–fact possibility of Mandrake interfering is explicitly raised. But then it is explicitly denied in the antecedent. Thus, neither the actual world nor the closest antecedent world is supposed to be one where Mandrake interferes. This might serve as a signal that no contrast between Mandrake interfering and Mandrake not interfering is intended.

2.4.2. World Convergence Made Easy: The Future Similarity Objection

In the introductory part (1.), I presented Lewis's default criteria of similarity, which I repeat here:

'(1) It is of first importance to avoid big, widespread, diverse violations of law [big miracles].
(2) It is of second importance to maximize the spatio–temporal region throughout which perfect match of particular fact prevails.
(3) It is of third importance to avoid even small, localized simple violations of law [small miracles].
(4) It is of little or no importance to secure approximate similarity of particular fact, even in matters that concern us greatly.'[127]

As we have seen, Lewis in 'Counterfactual Dependence' uses these default criteria to avoid Fine's original future similarity objection. I shall now discuss several problems of this solution. I start with mentioning some puzzles and then discuss two exemplary ones in depth: Bennett and Elga worlds.

Lewis distinguishes big miracles that must be avoided at all costs and small ones which are cheap: As long as the miracles remain small, maximizing perfect fit of particular matters of fact weighs more. 'A big miracle consists of many little miracles together, preferably not all alike.'(p. 56) But how many?[128] How varied? Miracles are individuated like events (p. 56). But can't there be single events which take more of a miracle, perhaps a big one? Hence Bennett's distressed question: 'How much of a bump (or a click) is a fair trade for a twelve–hour shortening of the ramp [the smooth development towards the antecedent from the actual world modified by a small miracle]? As soon as the question is asked, one sees its absurdity. It has nothing to do with our actual uses of subjunctive conditionals.'[129] But how can we avoid it given Lewis's criteria? Bennett defends Lewis at this point:

'(2) No coherent account can be given of the nomological structure of a world that exactly matches α [the actual world] up to some time when it forks away through the occurrence of a small miracle.
To regard 2 as [...]counting much against it [assuming miracles] would be unduly optimistic about the conceptual aspects of the human condition.'[130]

However, the general questions have quickly been condensed into more focused puzzles. I mention two of them by way of examples. The following is valid in Lewis's logics:

P>>R, not(P>> not–Q), hence (P&Q)>>R.

We can develop the following counterexample:[131] Let A be: The kitchen works, B: The rooms are cold; C: The stove is lighted, G: The gas is on. A, C, D, G are actually false. Rooms can be heated independently by electricity or a combination of gas and stove. Furthermore, the kitchen works as well with gas as independently with electricity.
(i)A>> not–C
(ii) not(A>> not(notB))
(iii) from (i): A>> not(C&G)

[127] Lewis, 'Counterfactual Dependence', pp. 47-48.
[128] Assume I can reach an outcome by one small miracle, or two small miracles which together may add to a bigger one, or three… while each time gaining some amount of perfect match (the more small miracles required, the later they may take place).
[129] Bennett, *Conditionals*, p. 326.
[130] Bennett, *Conditionals*, p. 226.
[131] Bennett, *Conditionals*, pp. 333–334.

Hence (A&not–B)>> not(C&G). But when we think over this conclusion, it seems absurd.

In words:
(i) If the kitchen had worked, the stove would not have been lighted.
(ii) It is not the case that, if the kitchen had worked, it would not have been the case that the rooms would not have been cold. That is reconcilable with the rooms having been cold in some closest and not cold in other closest worlds.
(iii) from (i): if the kitchen had worked, it would not have been the case that the stove was lighted and the gas was on.

Hence if the kitchen had worked and the rooms had not been cold, it would not have been the case that the stove was lighted and the gas was on.

Another counterexample:

'My coat was not stolen from the restaurant where I left it. There were two chances for theft, two times when relevant indeterminacies or small miracles could have done the trick. [...] and the candidate for the later theft is a rogue who always sells his stuff to a pawnbroker named Fence. If the closest A–world involves the latest admissible fork, it follows from the above story that *if my coat had been stolen from the restaurant, it would now be in Fence's shop*. That is not acceptable.'[132]

Instead of discussing these counterexamples, I shall discuss two different ones, which have to do with the possibility of achieving perfect future match in facts by just a small miracle. The first is due to Adam Elga, the second is due to Jonathan Bennett. I shall assess exemplary ways of meeting these challenges and then compare them.

[132] Bennett, *Conditionals*, p. 220.

2.4.2.1 Elga Worlds

In this chapter, I discuss Elga's counterexample to Lewis's default criteria of similarity for possible worlds: a largely counterentropic world may achieve perfect future match in facts at the cost of just a small miracle. I summarize Elga's argument. Then I discuss several attempts at dissolving Elga's challenge. I most extensively discuss a proposal by Jeffrey Dunn to write the preservation of special science laws into the similarity criteria. I argue that the Schaffer–Kment proposal known from the chapter on Morgenbesser cases is superior in meeting Elga's challenge. I add another facet of the challenge, the problem of amazingness. I argue that the most thorough way of solving that independent problem in terms of high–likelihood properties better accords with the Schaffer–Kment approach than Dunn's.

I sketch Elga's counterexample to Lewis's criteria of similarity. Lewis's criteria were to ensure the asymmetry of counterfactual dependence: while it takes only a small miracle for an antecedent world to diverge in a very short time from the actual world towards the antecedent, it takes a big, widespread miracle for an antecedent world to perfectly reconverge such as to perfectly match the actual world in particular matters of fact.

Elga considers

> 'At 8:00, Gretta cracked open an egg onto a hot frying pan. According to the analysis, are the following counterfactuals true?
> [E13] If Gretta hadn't cracked the egg, then at 8:05 there wouldn't have been a cooked egg on the pan.
> [E14] If Gretta hadn't cracked the egg, then at 7:55 she wouldn't have taken an egg out of her refrigerator.'[133]

Consider two competitors for closest non-cracking worlds: Lewis's favourite w_2 perfectly matches the actual world w_1 until shortly before 8:00. At that point, a small miracle occurs such as to prevent the egg from being cracked. Then w_2 develops according to the laws such as to never again perfectly match the actual world. w_3, in contrast, differs from the actual world before 8:00 such that the egg is not cracked (and not taken out of the refrigerator at 7:55) but some time after 8:00 converges to the actual world by dint of a small miracle. Lewis insists that the w_3 – strategy is not feasible. For any normal event leaves many and varied traces. It would need many and varied unlawful events, a big miracle, to suppress these traces.[134] So the asymmetry is this: certain divergence worlds are better candidates for closeness than any convergence worlds.

Elga sets out to show that pace Lewis there are candidate worlds for w_3 that need nothing but a small miracle.[135] In deterministic statistical mechanics, a physical state can be completely described by the positions and momenta of particles. Starting from the egg in the pan at 8:05, Elga contrasts a normal *future* development to the thermodynamically atypical future development which reverses the actual forward-directed development of the egg between 8:00 and 8:05; the egg uncooks and jumps back into the shell. The reversed development is extremely unstable. Its closest neighbours in phase space are states which differ from it just in a small local group of molecules. It needs only a tiny change at 8:05, a small miracle, to proceed to one of these developments. But this tiny difference very quickly spreads such as to give rise to a completely different, thermodynamically normal development. A tiny change leads to a completely different result. Now an analogous situation can be achieved by running the symmetrical laws *backwards*: it needs only a tiny variation immediately before 8:05 to switch from the normal *past* of the egg (cracking and cooking) to a completely different, thermodynamically reversed past development. Elga envisages a process of 'reversed rotting': a possible future development of the egg is reversed and projected into the past. A coniform

[133] Adam Elga, 'Statistical Mechanics and the Asymmetry of Counterfactual Dependence', *Philosophy of Science*, 68 (2001), S313–S324 (p. S314).
[134] Lewis, 'Counterfactual Dependence', p. 47.
[135] Elga, 'Statistical Mechanics', p. S318.

'infected region' comprises this complete development from the distant past up to the point of convergence at 8:05. In the distant past, the infected region was huge. But due to its instability, it rapidly shrank and gave way to normal developments up to the small miracle immediately before 8:05.[136] The egg has never been taken out of the refrigerator, never been cracked. Nevertheless, the thermodynamically reversed development eventually comes so close to the actual development as to achieve perfect match by a small miracle administered immediately before 8:05. Since this match is perfect, it comprises all the alleged traces of the cracking.

As it seems, we cannot rule out that the Elga world is closer than all competing non-cracking worlds. It perfectly abides by the actual fundamental laws of nature except for a small miracle, and it counters the perfect pre–antecedent match Lewis's candidate worlds exhibit by perfect post–antecedent match. If w_3 is closer than its competitors, Lewis's criteria have the wrong counterfactuals come out true, for instance the intuitively false

(E14) If Gretta hadn't cracked the egg, then at 7:55 she wouldn't have taken an egg out of her refrigerator.

So Lewis's similarity metrics grossly misses our common counterfactual verdicts.

Elga undone? Dunn's proposal

I shall discuss in detail a proposal by Jeffrey Dunn how to mend Lewis's criteria of similarity such as to demote Elga worlds from competing for closeness. While Lewis has in mind fundamental laws, Dunn also takes into account the laws of the special sciences, among them 'lawlike relations that are not entailed by the fundamental laws'.[137] He gives them fourth importance, demoting Lewis's fourth criterion to fifth importance:

(4´) It is of the fourth importance to avoid violation of the special science laws.(p. 84)

The Elga world is disqualified as a candidate for closeness due to its violating the laws of thermodynamics. Briefly, Dunn invokes the second law of classical thermodynamics: heat cannot spontaneously flow from a hotter location to a cooler location.[138] Yet that is what Elga's reverse process would amount to; the backwards rotting egg would have to absorb heat from its relatively cool surroundings in order to end in the pan at cooking temperature. Lewis's candidate worlds abide by the second law of thermodynamics while the Elga world violates it. So the former are closer according to Dunn's criteria.

In my critical discussion of Dunn's approach, I will proceed as follows: I will point out several intrinsic problems of Dunn's amendment of Lewis. I will argue that Dunn's strategy is not the best way of meeting Elga's objection. Elga's counterexample is only one instance of the notorious future similarity objection; other cases completely evade Dunn's strategy. Jonathan Schaffer and Boris Kment provide a strategy which is superior to Dunn's; it applies to all counterexamples presently on offer. I present this strategy and its merits. Dunn ignores it, although it arguably is required even to evade intrinsic problems of his own approach. I close with concerns about the Schaffer–Kment strategy.

Intrinsic problems of Dunn's approach

[136] Elga, 'Statistical Mechanics', p. S323.
[137] Jeffrey Dunn, 'Fried Eggs, Thermodynamics, and the Special Sciences', *The British Journal for the Philosophy of Science*, 62 (2011), 71–98 (p. 73).
[138] For a more detailed modern formulation cf. Dunn, p. 82.

I begin with some methodological remarks. Facing doubts as to whether his default criteria of similarity fit our snap judgements, Lewis maintains that an account of similarity of worlds should be ultimately judged exclusively by intuitively compelling test counterfactuals.[139]

Even if we accept that an account is *ultimately* to be judged *only* by sustaining intuitive counterfactuals, the very idea of a general similarity metrics as contrasted to a casuistry requires us to look for unifying traits counterfactual verdicts have in common. These traits can be used to explain our verdicts by general criteria these verdicts implicitly draw on. And they guide us in conjuring up test counterfactuals. In identifying such traits, feelings of simplicity and naturalness are likely to play a heuristic role.[140] Nevertheless I will try to provide test counterfactuals as far as possible. There are other things which should play a role in evaluating an account, though. For instance, we will see that Dunn's account clashes with Lewis's basic assumption that the similarity relation is a total preorder (transitive and complete) on which his logics for counterfactuals rests. Even if that assumption itself should face the tribunal of our best intuitive verdicts about particular counterfactuals, its violation weighs heavily against an account of similarity.

I concede a certain prima facie plausibility to Dunn's approach; it pays due respect to *irreducible* laws of the special sciences.[141] However, there are difficulties. The Elga world is described in terms of statistical mechanics. Dunn confesses his uncertainty how the latter relates to thermodynamics:

'An extremely tentative view about the relation between the two is that statistical mechanics is an attempt to explain how we get the special science laws of thermodynamics, given certain fundamental physical laws. It is important to note that I am attempting to construe classical thermodynamics as a special science, not statistical mechanics.'(p. 82)

Dunn's treatment raises a question: If certain laws of the special sciences can be *reduced* to more fundamental laws and facts such as those of statistical mechanics as used by Elga, what should the place of these laws in the similarity ordering be?

One may deny that this is a problem, arguing as follows: assume a special science law is reducible to the fundamental laws in the sense of being entailed by them. By dint of the entailment, for the special science law to be violated, the fundamental laws must be violated as well but not vice versa. So in case of reducibility, Dunn's criterion yields the following result: a world where the fundamental laws are violated but the special science laws are not is closer, other things being equal, than a world where the special science laws are also violated.

To elaborate the problem, I begin with an example of Kment, concerning the counterfactual dependence of particular facts on fundamental laws:

'[E15] If (Law of Gravitation) had not been a law, then events would still have at least approximately conformed to it.

No one I asked believed that this counterfactual was true[...]

[139] Lewis, 'Counterfactual Dependence', p. 43. I am grateful to anonymous referees for insisting on this point.

[140] For instance, they may guide us in disregarding 'gruesome' similarities (Lewis, 'Counterfactual Dependence', p. 42). In a similar vein, Brian Weatherson, 'What Good Are Counterexamples', *Philosophical Studies*, 115 (2003), 1–31, (p. 11), espouses a reflective equilibrium between linguistic intuitions and features like simplicity and naturalness, albeit without giving the former ultimate priority.

[141] Dunn hints at an independent motivation (Dunn, p. 81 ann. 9). Whether special science laws should form part of the default ordering of worlds is a matter of further debate: as noted by Dunn, which laws we tend to preserve is very context–sensitive; psychologists will rather tend to hold onto the laws of psychology than those of chemistry (cf. p. 95). Thus instead of building the special science laws into the default similarity ordering, one might rather claim them to override Lewis's default criteria *in the context of the respective sciences*.

[E16] If the master law [comprising all fundamental laws] had not been a law, the history of the world would still have been very similar to what it was actually like.'[142]

Kment suggests that the relationship between laws and particular matters of fact which makes us reject these conditionals is *explanation*. In the closest counterfactual situation where the actual *explanans* does not obtain, we do not hold onto the *explanandum* either. Without committing myself to this view, I maintain that there is a parallel counterfactual dependence of special science laws on laws they can be reduced to.

Consider an outstanding candidate for reducing the laws of thermodynamics: the Albert–Loewer recipe, which comprises:

'(i) the Newtonian dynamical law: $F = ma$; (ii) the Past Hypothesis: the initial conditions are low entropy; and (iii) the Statistical Postulate: there is a probability distribution uniform on the standard measure over those regions of phase space compatible with our empirical information.'[143]

Assume the laws of thermodynamics can be reduced à la Albert–Loewer. Then the following should be rejected:

(E17) If the Past Hypothesis had not been true, still the laws of thermodynamics would not have been violated.

Yet Dunn seems committed do (E17).

I anticipate a reply: Dunn distinguishes between violating the laws of thermodynamics and these laws not obtaining (i.e. being laws) at all (pp. 94–95 footnote 33). In the latter case, they are not violated. Lewis insists that a fundamental law which has an exception does not obtain in the first place.[144] We might expect any situation where the Past Hypothesis does not obtain to be a situation where the laws of thermodynamics do not obtain at all (and thus are not violated).

I use the Elga world to conjure up a counterexample. By Dunn's lights, the Elga violates the laws of thermodynamics. To be violated, these laws must obtain in principle. In Elga's vision, the infected region grows the further back we go in time. So holding onto the Albert–Loewer recipe, we may construe the Elga world as follows: the Past Hypothesis does not hold; in the infected region, initial conditions are not low entropy. However, in the regions surrounding the infected region, initial conditions are low entropy. And that is sufficient to ensure that the laws of thermodynamics obtain in principle. Of course, there is no reason to deem the Elga world the closest world where the Past Hypothesis does not hold. But some world like it is a good candidate. We cannot accept (E17) as long as we cannot decide between two candidates for the closest world where the Past Hypothesis does not obtain: a world where the laws of thermodynamics principally obtain but are violated as contrasted to a world where the laws of thermodynamics do not obtain.

I anticipate a second reply: Dunn is critical of the Albert–Loewer recipe. He has a strategic motive for his criticism: if the Past Hypothesis is a fundamental law, the Elga world can be rejected for violating it. We do not need Dunn's amendment. To counter this threat,

[142] Kment, pp. 280–281.

[143] Jonathan Schaffer, 'Deterministic Chance', *British Journal for the Philosophy of Science*, 58 (2007) 113 –140 (p. 122), cf. Dunn, p. 83. I surmise that things would be the same if thermodynamics were ultimately founded on other theories, say the GRW version of quantum mechanics (cf. Schaffer, p. 122 ann., Jill North, 'What is the Problem about the Time–Asymmetry of Thermodynamics? – A Reply to Price', *The British Journal for the Philosophy of Science*, 53 (2002), 121–136). Dunn notes that in case of a reduction, there might be no true counterfactuals except ones merely specifying *probable* consequences (Dunn, p. 84). But it should not come as a surprise if fundamental physics were to reveal many of our folk counterfactuals as mere approximations.

[144] Lewis, 'Counterfactual Dependence', p. 45.

Dunn voices doubts that the Past Hypothesis is a fundamental law. Fundamental laws à la Lewis are confined to perfectly natural properties; *entropy* is no such property. And fundamental laws are usually regarded as regularities; the Past Hypothesis is no regularity (cf. pp. 83–84). Be that as it may, I use the Albert–Loewer recipe only as an outstanding model of reduction. So if we do not principally eschew reduction of special science laws, there should be other reductive efforts which would serve the task.

There are two further problems. Both are mentioned but not thoroughly solved by Dunn.

We cannot exclude the Elga world as a candidate for being closest as long as it might exhibit more match in particular facts than the divergence world à la Lewis. This can happen when the world is finite and stretches further into the future than into the past, all relative to the time of convergence. As a remedy, Dunn gerrymanders a reading of the second criterion: 'we do not quantitatively compare a region of past match with a region of future match.'(p. 86)

There are two ways of cashing out Dunn's reading; the first is: if, other things being equal, world w_A exhibits more perfect match with actuality in pre–antecedent facts but w_B exhibits more perfect match with actuality in post–antecedent facts, both are *equally* similar to actuality. The unfortunate consequence is this: for some w_A and w_B, there will be a world w_C which fares even worse than w_A in perfect post–antecedent match but equals w_A in pre–antecedent match (Assume a different small miracle leads to greater regions of post–antecedent mismatch than the small miracle by which w_A departs from actuality). As a consequence, w_C is as similar to actuality as w_B but less similar than w_A. This is irreconcilable with Lewis's view that the relations of overall similarity among worlds form a total preorder, which is transitive and complete.[145] Since w_A and w_B are equally similar and so are w_B and w_C, by transitivity w_A and w_C must be equally similar. But w_C is less similar than w_A.

The second way of putting Dunn's interpretation is that w_A and w_B are *incommensurable* with regard to overall similarity.[146] This is not reconcilable with the completeness of the similarity ordering: if w_A and w_B at all qualify for overall similarity, either one is more similar or both are equally similar. Even if an account must ultimately be judged by its ability to deal with paradigm counterfactuals, the basic formal properties of the similarity ordering are crucial to Lewis's standard analysis of counterfactuals. Dunn's proposal conflicts with these properties.

I come to what could be the most grievous difficulty as it is easily fleshed out in terms of individual paradigm counterfactuals. Particular matters of fact might interfere with laws of the special sciences:

'Grant that biology is a special science, and imagine that there was some critical event that occurred in the past, say a crucial step in the move toward DNA, in spacetime region R, that led biology on its current course. Let's assume that had this particular critical event not occurred, then biology would have been very different. Now, consider the counterfactual:

[E18] If lightning had struck in region R, then the laws of biology might have been very different.

[E2] strikes us as true.'(p. 92)

The problem (E18) poses is that in the standard Lewisian construal, it entails

(E19) It is not the case that if lightning had struck in region R, then the laws of biology would have been the same.

This contradicts Dunn's (4′) as far as (4′) prescribes to hold onto the laws of biology. Dunn follows Lewis's suggestion that 'might' can also be read as 'it would be that: different laws are possible.'(p. 93) To judge this proposal, we have to distinguish two cases. In one case,

[145] Lewis, *Counterfactuals*, p. 14.
[146] cf. Morreau, 'Trouble with Similarity'.

say where lightning directly strikes the critical event, the fundamental laws and the facts in the scenario entail different biological laws. Then, Dunn concedes, we accept (p. 94 ann.):

(E20) If lightning had struck the critical event, then the biological laws would have been different.

Then we must also accept that the laws might have been different, 'might' understood in the standard way.

There is another, more problematic case: the fundamental laws and the facts modified by lightning do not ensure which biological laws will come to obtain. Worlds with our biological laws and worlds with different biological laws are equally close. One might worry how such a situation can be reconciled with determinism. In reply, first, the antecedent is vague. It could allow for different default resolutions. Second, even if the antecedent were perfectly precisified, nothing precludes that two different small miracles lead to antecedent worlds that are equally close, one with our biological laws, one with alien ones.[147] In that situation, doctoring the 'might'–conditional won't help. For we pace Dunn reject that the laws *would* have been the same. This clashes with Dunn's explicit commitment to:

(E21) If lightning had struck in region R, then the laws of biology would have been just as they actually are (p. 92).

Perhaps there is a way out: in light of Dunn's comments in his footnote 33, it is not clear that his account really commits him to (E21).[148] Footnote 33 presents a situation where the actual biological properties fail to be instantiated; as a consequence, the actual biological laws do not obtain (cf. pp. 94–95). Dunn insists that his criterion (4´) does not rule out such a world from being closest. Assume that for any lightning situation where the actual facts and the lightning together with the fundamental laws do not entail the actual biological laws, either the laws of biology obtain or they do not obtain at all. Then Dunn is not committed to (E21).

This is literally right as far as (4´) demands that the laws of the special sciences should *not be violated*, not that they should obtain – but only given a specific reading of the criterion: a law which does not obtain because the properties to which it applies are not instantiated is not thereby violated. In order to sustain this reading, Dunn must distinguish situations where the laws are not violated because they do not obtain at all and situations where they are violated. Otherwise laws, at least those of the special sciences, *could* not be violated at all; all possible situations whatsoever would fare equal with respect to Dunn's (4´). The Elga world could not be ruled out.

Consequently we can further develop the second lightning case I have discussed into two subcases: in one counterfactual situation where lightning occurs, the laws and the facts together only ensure that one of two alternatives will come to pass: either the actual laws of biology obtain or they do not obtain at all because our biological properties are not instantiated. For that subcase, Dunn gets the right result: if the lightning situation had obtained, the actual laws would not have been violated but might not have obtained at all.

Yet there is another subcase: in a different counterfactual situation where lightning occurs, the fundamental laws and the actual facts modified by lightning do not ensure that our biological laws *will not be violated;* for in that situation, our biological laws perfectly hold or they will be violated (Perhaps there is also the third alternative of the laws not obtaining at all). Dunn must falsely maintain that if this situation had come to obtain, the laws of biology would not have been violated.

[147] Dunn presents a structurally analogous case (p. 95).
[148] In considering this reading, I follow the advice of an anonymous referee.

To assess the possibility of this second subcase, I take a closer look at what it could mean to violate a law. By Lewis's lights, fundamental deterministic laws do not allow for the distinction between a law being violated and not obtaining at all. What concerns laws of the special sciences, things are more intricate. These laws permit exceptions. To fit Dunn's distinction of not obtaining at all and being violated, a violation must steer between an exception permitted by the laws and the laws not obtaining at all. For instance, Dunn must prevent Elga from retorting that in w_3, the laws of thermodynamics are not violated because they do not obtain in the first place; so Dunn must insist that in the Elga world, the actual thermodynamic properties are instantiated in spite of the infected region violating the laws of thermodynamics.

If he can do so concerning the Elga world, he cannot rule out that the lightning situation can be further developed along the following lines: let there be several regions R_1,..R_n, each of which is sufficient to bring about our biological properties and the concomitant laws; but assume that had lightning struck region R_1, that region might have been infected by biological systems behaving deviantly such as to violate the actual laws of biology;[149] more precisely, the fundamental laws and the facts modified by lightning would not have ensured that R_1 would not have been infected (In one closest lightning situation R_1 would have been infected, in another, it would not have been). We should reject but Dunn must accept

(E22) If lightning had occurred in region R_1, the actual laws of biology would have gone unviolated.

In sum, it does not help Dunn to introduce the case where laws are not violated because the respective properties are not instantiated at all. Note that my argument does not commit me to accepting the distinction of laws being violated or not obtaining at all; I just point out what Dunn is committed to.

I see two different solutions to the problems outlined about Dunn's approach. The first is amenable to Dunn's overall proposal that the laws of the special sciences be given due weight in the similarity ordering. But what is their due weight? Granting them fourth importance spells trouble, as we have seen. What about *pro*moting them to second order, thereby degrading match in particular facts to third order? Dunn does not discuss this suggestion. The example from biology can be used to rule it out. Assume we reject (E22). Yet surely we could tailor particular facts in the counterfactual situation such as to get our biological laws unviolated in spite of the lightning, say by removing the decisive region (the region that is decisive *in the counterfactual situation*) from the zone of lightning to a more quiet place such as to arrive at our DNA. This requires us to change particular facts aplenty but does not have to violate any (actual) laws. Our rejection of the above counterfactual testifies against our holding onto the laws of the special sciences at any cost in particular matters of fact.

If the laws of the special sciences sometimes counterfactually depend on particular matters of fact, there seems only one way left to accommodate them: Lewis's second criterion must not be demoted but differentiated. Laws of the special sciences not entailed by the fundamental laws are traded against match in facts. But how to weigh them? Drawing on a relative naturalness order of facts envisaged by Lewis and John Hawthorne, Dunn depicts a hierarchy of laws of the special sciences (p. 89). This naturalness order could give rise to a hierarchy of weights imposed on both facts and laws of the special sciences. Not all facts have the same weight. By default, more natural ones count more than less natural ones, more general laws (say those of chemistry) count more than less general ones (say those of biology), or, as Dunn suggests, laws count less the more exceptions they allow, and so on (p. 95). And laws of

[149] One may doubt that such deviant biological systems are microphysically possible. But there will be other examples; one is my scenario of the Elga world conflicting with the Past Hypothesis such as to violate the laws of thermodynamics.

the special sciences may be traded against match in particular facts. This seems a promising way of differentiating Lewis's treatment. The second criterion could be amended thus:

(2´) It is of second importance to maximize the weighted sum of matching particular facts and laws of the special sciences.

However, we might have to give up Lewis's cherished idea of a handy system of priorities. It will prove to be extremely complicated to spell out the weighted sums of facts and special laws on which the ordering of worlds depends.

And there are two devastating problems:

(a) It is doubtful that the naturalness ordering allows us to deal with the contingency of biological laws as depicted in the lightning counterexample. The criterion does not give the decisive region R_1 more weight than other regions; R_1 just happens to be the right place at the right time to give rise to the laws of biology. Match in laws of the special sciences is integrated into the third criterion such as to be traded against match in facts; so there is no reason why to hold onto the actual region R_1 as decisive for the laws of biology rather than to hold onto these laws themselves. The alternative solution I will present seems superior in dealing with this issue. It pays due respect to the decisive role of R_1.

(b) We cannot easily dismiss the Elga world. Perhaps due to their importance in the system of sciences, the laws of thermodynamics get enough weight to outdo any advantages in particular matters of fact the Elga world may have, for instance in virtue of convergence at a very early stage of the world; but this is by no means sure.

Intimidated by this outlook, we might prefer a different way of saving Dunn's proposal from the pitfalls I have outlined. I shall consider several alternatives.

Alternative solutions

As for the first alternative, it can be derived from a closer look at the Albert–Loewer recipe. From an isolated viewpoint of statistical mechanics, our world with its characteristic thermodynamic asymmetry seems very amazing.[150] Many philosophers and scientists feel the need for an explanation. A prominent explanation is the Albert–Loewer recipe. To repeat, the recipe comprises:

'(i) the Newtonian dynamical law: $F = ma$; (ii) the Past Hypothesis: the initial conditions are low entropy; and (iii) the Statistical Postulate: there is a probability distribution uniform on the standard measure over those regions of phase space compatible with our empirical information.'

Assume the Past Hypothesis describes a huge fact about the early universe. Under determinism, the antithermodynamic processes at the Elga world violate the Past Hypothesis on a large scale. After all, Elga's paradigm process is coniform. It ends with a tiny divergence but spreads the further backwards we move in time. So it requires a huge change in the initial conditions of the universe.

Thus, there is an easy amendment of Lewis's default metrics, which demotes Elga worlds from being closest. Lewis denies that any particular physical fact whatsoever should form part of the ideal physical theory. His idea of such a theory is enshrined in his best system analysis. A scientific system is best iff it strikes the best balance of simplicity, fit and strength. Such a system will largely consist of laws. But it might also contain certain facts, provided these facts contribute enough to its strength:

[150] Huw Price, 'Boltzmann's Time Bomb', *The British Journal for the Philosophy of Science*, 53 (2002), 83–119.

'The ideal system need not consist entirely of regularities; particular facts may gain entry if they contribute enough to collective simplicity and strength. (For instance, certain particular facts about the Big Bang might be strong candidates.) But only the regularities of the system are to count as laws.'[151]

In light of these considerations, the following amendment of Lewis's default metrics is suggestive:

(1´) It is of the first importance to avoid a big, widespread, diverse departure from the ideal physical theory.
(2´) It is of the second importance to maximize the spatio–temporal region throughout which perfect match of particular fact prevails.
(3´) It is of the third importance to avoid even a small, localized simple departure from the ideal physical theory.
(4´) It is of little or no importance to secure approximate similarity of particular fact, even in matters that concern us greatly.

Instead of laws, the default metrics resorts to the actually true physics, including both regularities and, perhaps, particular facts of special importance, for instance facts about the Big Bang. This minimal amendment preserves the complete spirit of Lewis's metrics.[152] It deviates from the original metrics just in case the Past Hypothesis does not qualify as a law but as an especially significant fact to be included in the ideal physical theory. To be sure, the amendment won't work if the Past Hypothesis or some comparable explanation of the thermodynamic asymmetry does not figure in the ideal physical theory. But it would be surprising if physical theory fell completely silent about one of the most striking structural features of our world.

There is an alternative for how to use the Past Hypothesis. If the Past Hypothesis is granted the status of a fundamental *law*, it introduces a fundamental asymmetry. The Past Hypothesis breaks the symmetry between the past and the future. In fact, any explanation where a fundamental nomic necessity underlies the thermodynamic asymmetry is likely to conflict with Elga's template. The Elga world would be demoted from closeness. Elga would only have taken into account part of the actual fundamental laws of nature. However, there are many reservations about this solution. Shouldn't the semantics of counterfactuals be neutral to particular developments in physics? Moreover, there are reasons why the Past Hypothesis does not qualify as a Lewisian law: (i) the Past Hypothesis is no regularity, (ii) entropy is no perfectly natural but a high–level organizational property (Dunn, pp. 83–84). Thus, the outlook of a solution which takes care at once of Elga's and Bennett's problems is doubtful.

Dunn and the Schaffer–Kment remedy
Problems with other convergence worlds

I have pinpointed intrinsic problems of Dunn's approach. There is an argument why Dunn's overall strategy is not recommendable to save Lewis from convergence objections. Elga worlds are only one sort of convergence worlds threatening Lewis's semantics. There are other such worlds as witnessed by the many examples varying the so-called future–similarity objection.[153] Some of these examples are thermodynamically perfectly inconspicuous. For instance, there is a variation of the original Nixon case: a beetle is placed in a causally isolated box, which has

[151] David Lewis, 'New Work for a Theory of Universals', *Australasian Journal of Philosophy*, 61 (1983), 343–377 (p. 367).
[152] One might feel concerned that even a large–scale albeit not *varied* violation of the initial low entropy condition does not count as a big violation. Yet I do not see what could prevent tailoring the vague boundary of big and small deviations such as to fit my needs.
[153] Initiated by Fine's famous Nixon–example, overview of the literature in Schaffer, 'Counterfactuals', Kment, 'Counterfactuals and Explanation'.

only one connection to the rest of the world: there is a wire transmitting a signal to a doomsday machine. The signal can be activated by the beetle. But the beetle does not activate the signal. Shortly afterwards the whole box is cleanly disposed of. We accept

(E23) If the beetle had activated the signal, doomsday would have occurred.

But it would have taken only a small miracle to interrupt the signal.[154]

I will now summarize the most eligible recipe on offer that allows both to save Dunn's account from the intrinsic problems discussed above and to dispel convergence worlds of whatever kind, including the Elga world. As we have seen in discussing Morgenbesser cases, Schaffer demands that one should only maximize match in facts '*from those regions causally independent of whether or not the antecedent obtains'*.[155] Kment replaces Schaffer's criterion by a sameness–of–explanation clause: match in facts should count as far as their explanatory history is the same.[156]

When we consider the minimal variation of our world that goes together with implementing an antecedent situation, it is a matter of course that some things will also vary: those which depend on the antecedent situation or the situation replaced by the antecedent situation. So match in these things should not count. In the actual world, the causal history of the post–convergence facts includes cracking the egg, in the Elga world, it does not. Since match thus achieved does not count, the Elga world is disqualified by the criterion 'match in facts'. Thus, the Schaffer–Kment approach helps Dunn to avoid one problem: Elga's world may prevail what concerns match in post–antecedent facts compared to divergence worlds with a short past, but this prevalence is due to match in the convergence region; and since the antecedent obtaining or not figures in the causal–explanatory history of the convergence region, match in this region does not count. The Schaffer–Kment remedy can also be used to evade another problem of Dunn's: we may extend the irrelevance of facts downstream from the antecedent to *laws* as far as the obtaining or non-obtaining of the antecedent figures in the latters' explanatory history.[157] In the lightning situation where it is open whether the actual biological laws will be violated or not, their preservation does not contribute to closeness as the antecedent figures in the explanatory history of their being violated or not.

Yet Dunn has no reason to be comforted: the Schaffer–Kment argument supplants Dunn's. It demotes Elga's world without any appeal to laws of the special sciences. And it has important advantages in terms of theoretical economy; it also removes other convergence worlds against which Dunn's argument is of no avail, for instance smoothly converging worlds. Since preventing Elga worlds is the only motivation Dunn provides for his proposal, Dunn is preempted.[158]

One may feel concerned that the Schaffer–Kment approach does not fit into the dialectical context of Dunn's:[159] Dunn aims at a *Lewisian* response to Elga's counterexample.

---

[154] Cf. Wasserman, p. 59.

[155] Schaffer, 'Counterfactuals', p. 305.

[156] Explanatory history not restricted to causal explanation, but more broadly confined, in order to account for counterlegals. I mention Schaffer and Kment as the most recent versions of the account, though Kment refers to a more remote ancestry (Kment, p. 273).

[157] Kment on similarity to world w:

'1. It is of the first importance to ensure sameness of laws.
2. It is of the second importance to avoid big alien violations of the laws of *w, provided the conformity to the relevant laws has the same explanation* as in *w*.'(Kment, p. 296, m.e.).

[158] There is a vague hint at an independent motivation (Dunn, p. 81 ann. 9).

[159] This concern has been voiced by an anonymous referee.

Lewis wants to use counterfactuals to analyse the notion of causation. So to avoid circularity, a Lewisian response might be bound to avoid causal notions as used by Schaffer and Kment.

Even supposing the Schaffer–Kment approach diverges from Lewisian orthodoxy while Dunn's approach preserves it, it seems still worthwhile to consider the former as a competitor for Dunn's overall aim of providing a convincing similarity metrics which meets Elga's objection. Faced with the problems of Dunn's account, an appropriate reply to Dunn may just put up an alternative which is independently plausible and allows both to remedy certain flaws of Dunn's approach and to meet Elga's objection, even at the price of running counter to Lewis's further metaphysical ambitions. Then the price of orthodoxy can be better judged. Moreover, since Dunn and Lewis insist that a similarity metrics should be judged mainly by getting intuitive counterfactual verdicts right, we may feel encouraged to tackle Lewis's analysis of counterfactuals as an autonomous topic – to be treated independently of the further metaphysical use in analysing causality it may then be put to. And if we disregard the analysis of causality, the Schaffer–Kment approach seems no less orthodox than Dunn's; the former includes explicitly explanatory relationships, the latter an additional type of laws over and above the fundamental level where the other criteria are situated.

Anyway orthodoxy is unlikely to be an all–or–nothing matter. The Schaffer–Kment approach is Lewisian in sharing Lewis's truth–condition for counterfactuals and most of Lewis's four–part lexical similarity ordering. Moreover, far from announcing his ideas as running counter to Lewis's aims, Schaffer takes great pains at reconciling the use of causal terms with Lewis's metaphysical ambitions:

'Might one adopt *both* a causal independence account of counterfactuals, *and* a counterfactual account of causation? Is the resulting circularity *problematic*? *Ontologically* speaking, I see nothing problematic here. The truth about both counterfactuals and causality still *supervenes* on the arrangement of events. Or at least, nothing here contradicts that. The causal and counterfactual facts can still, for instance, be regarded as 'co-supervenient' upon a Humean base. If there were a problem, it could be a *conceptual* problem. One would lose *linear definability* – no ordered chain of definitions could wind from the Humean base up through the conceptual superstructure. But perhaps linear definability was never in the offing. *Because concepts do not have definitions.'*[160]

Schaffer intimates that preserving Lewis's conception of supervenience matters more than the non-circular definability of causality by counterfactuals: the truth about both causality and counterfactuals supervenes on the mosaic of events; but there is no non-circular definition of concepts like causality.

In sum, in advocating the Schaffer–Kment approach, I do not trade a Lewisian for an un-Lewisian approach, but at worst a more against a less orthodox Lewisian one. Preserving a counterfactual analysis of causality should go as one asset among others into the trade–off rather than be put up as the shibboleth of Lewisian and non-Lewisian approaches. And weighing the question of causality against the advantages of the Schaffer–Kment approach, I think the latter better qualifies as an amended Lewisian view than Dunn's.

The Problem of amazingness

So far the Schaffer–Kment approach seems to be the best strategy against Elga's argument. Yet the following reasoning calls for a more differentiated view: while a similarity metrics is ultimately to be judged by getting intuitively plausible counterfactuals right, it might be heuristically important for devising this metrics to give a more fine-tuned diagnosis what guides our intuitions in eschewing the Elga world. The reason is that the intuitive examples often

[160] Schaffer, 'Counterfactuals', pp. 307–308.

display some common feature that is relevant to finding both general criteria of similarity and paradigm counterfactuals that could be crucial to testing them.

Besides subsuming Elga's example under the future similarity objection, a further classification is tempting and has not been told apart from the future similarity objection for a long time. As Elga grants, the thermodynamically reversed world is *amazing*. Lewis himself discusses amazing convergence for indeterministic worlds.[161] However, in the aftermath of Lewis's and Elga's discussion the issue of amazingness and the future similarity objection have somewhat grown apart: unremarkable convergence worlds like Wasserman's beetle and the worlds to be discussed below have been developed; and other counterexamples to Lewis's standard analysis show that there is a problem with *amazing* worlds as candidates for closeness, which is independent of the convergence problem. Consider

'(D39) If I had dropped the plate, it would have fallen to the floor.
[...] there is a small chance that the consequent fails to obtain, given the antecedent. Thus, the following is tempting:
(D40) If I had dropped the plate, it might have flown off sideways'.[162]

(D40) seems acceptable if we take into account quantum physics.

But in Lewis's original theory, (D39) and (D40) are contradictory. As a consequence, if 'might'–counterfactuals like (D40) are true, most everyday counterfactuals like (D39) seem false. If the world where the plate flies sideways can be somehow prescinded from worlds like ours, the problem disappears. This example testifies to an independent problem of amazing candidates for closeness, which should not be simply classified together with the future similarity objection ((D40)–antecedent–worlds do not converge to ours). Yet both difficulties are connected; as Elga shows, future match can be easily achieved provided circumstances are allowed to be amazing. The amazingness problem common to Elga's case and the above example of the plate flying sideways is that our intuitive verdicts against counterfactuals supported by amazing candidates for closeness, such as (E14) ('...Gretta wouldn't have taken an egg out of the refrigerator') as supported by the Elga world and (D40) ('...the plate might have flown off sideways'), do not seem backed by Lewis's original criteria. The reason for eschewing the plate scenario arguably lies in the latter's being amazing; and this reason in principle also applies to the egg–scenario.

When Elga published his counterexample, the future similarity problem and the problem of amazingness had not yet been as clearly separated. Elga's counterexample exemplifies both problems, the future similarity problem and the problem of amazing candidates for closeness. For this reason, the challenge posed by Elga may be interpreted as anticipating both problems further developed by subsequent literature (although Elga had in mind only the convergence problem).

So to do justice to the full dialectical impact of Elga's challenge, it seems important to also discuss the Elga world as an instance of an *amazing* world, independently of its already being covered by a strategy against future similarity. For there will be examples of the amazingness problem where the strategy against future similarity does not work (the plate scenario). The systematic question arising from this discussion is how to supplement the most eligible solution to the future similarity problem by an answer to the problem of amazingness. I think that in this respect, too, the superiority of the Schaffer–Kment approach to Dunn's becomes obvious. Moreover, just in case these approaches do not work, it might be interesting that there is an independent strategy against Elga's counterexample, which arises from the latter's entanglement with amazingness.

[161] Lewis, 'Counterfactual Dependence', p. 63.
[162] Williams, 'Chances', p. 386.

Lewis treats amazing convergence as a problem of *indeterministic* worlds. He introduces the non–standard reading of 'might' exploited by Dunn and discards amazing worlds as exhibiting *quasi–miracles*. Yet firstly, as witnessed by the Elga world, the problem of amazingness is not confined to indeterministic worlds. The assumption of determinism does not rule out (D40). Perhaps our world is microphysically configured such as to allow for a world where a small miracle leads to the plate being dropped and flying off sideways. Secondly, Lewis had a hard time spelling out what constitutes a quasi–miracle.[163] The problem is not the mere *improbability* of some course of events:

'What makes a quasi–miracle is not improbability per se but rather the *remarkable* way in which the chance outcomes seem to conspire to produce a pattern. If the monkey at the typewriter produces a 950–pages dissertation on the varieties of anti–realism, that is at least quasi–miraculous; the chance keystrokes happen to simulate the traces that would have been left by quite a different process. If the monkey instead types 950 pages of stumbled letters, that is not at all quasi-miraculous. But given suitable assumptions on what sort of chance device the monkey is, the one text is exactly as improbable as the other.'[164]

If we follow Lewis's suggestion, the question becomes how to spell out remarkability here. The most thorough present proposal to spell out typicality vs. remarkability in terms of objective random properties is

'[...]to look, not at the probability of a particular outcome arising, but at the probabilities of a suitable set of properties which that outcome instantiates. When considering the outcome of flipping a fair coin, 'all heads' is a low-likelihood property (in the infinite case, it is probability 0 that the outcome has this property). 'Having as many heads as tails, in the long run' is a high-likelihood property (In the infinite case, it is probability 1 that the outcome has this property).'[165]

Being *remarkable* is defined in contrast to being *random*: remarkable situations fail to instantiate a suitable range of high-likelihood properties.[166] For instance, any particular sequence of infinitely many coin tosses has probability zero. The intuitively remarkable sequence *all heads* is as probable as a particular sequence of as many heads as tails. Yet only the latter instantiates the high–likelihood property 'as many heads as tails'.

Drawing on his definition of remarkability by a lack of high–likelihood properties, Robert Williams modifies Lewis's criteria: it is of first importance to avoid atypicality of the world as a whole (by the lights of the probabilistic laws of nature), of second importance to maximize the spatio–temporal region of perfect match, of third importance to avoid localized atypicalities, *all judged from the evaluation world* (usually the *actual* one. Williams' proposal gives a flavour of what we might aim at even for deterministic worlds. The Elga world as an instance of the amazingness problem can be dismissed due to its partial failure to instantiate high–likelihood properties.[167]

---

[163] Lewis, 'Counterfactual Dependence', p. 50.
[164] Lewis, 'Counterfactual Dependence', p. 60, my emphasis.
[165] Williams, 'Chances', 409, cf. Adam Elga, 'Infinitesimal Chances and the Laws of Nature', *Australasian Journal of Philosophy*, 82 (2004), 67–76 (p. 71).
[166] Williams, 'Chances', pp. 409–410.
[167] Williams, 'Chances', p. 418. Here is why I doubt that the proposal works as it stands. As Ryan Wasserman notes, pace Lewis and Williams we do not accept

(E24) If there had been a trillion coin tosses and I had bet against them falling all heads before, I would have won (cf. Wasserman, p. 62).

There seems to be a parallel to the lottery paradox in epistemology. Perhaps a solution to the latter could be worked into Williams' criteria such as to take care of this counterexample.
A low entropy boundary condition as it might be invoked in the foundation of thermodynamics also spells trouble. It provides a reason to deem our world atypical as a whole (cf. Price, p. 111) Yet just in case the initial condition

While any of the independently motivated accounts, Williams' answer to the amazingness problem and the Schaffer–Kment answer to the future similarity objection, rule out the Elga world as a competitor for closeness, both approaches might be necessary to take care of the different counterexamples that have been subsequently developed to bring out the pure voice of the two independent problems. So Williams' amendment of Lewis may supplement the Schaffer–Kment approach, each meeting one of the two problems enshrined in Elga's example. In principle, Dunn's and Williams' proposals seem reconcilable as well. Yet they sit awkwardly with each other. While each is sufficient to dismiss the Elga world, the unremarkable examples of the future similarity objection remain untouched. So the explanatory overkill is not allayed by complementarity in other cases.

---

of low entropy is both amazing, contingent, and unavoidable, we might accommodate it by subordinating the other default criteria of similarity to the following requirement:

It is of first importance that the overall structure of our world (the evaluation world) be preserved.

The low entropy boundary condition and fundamental laws may amount to overall structural features to be preserved.

2.4.2.2 Bennett Worlds

Having discussed amazing convergence worlds, I shall present a further problem for Lewis's standard analysis of counterfactuals. It arises from a counterexample of Jonathan Bennett's. As I will argue below, Bennett's example is a precursor of the more elaborated counterexample of Adam Elga's. Nevertheless it deserves more than historical attention. The argument used in its motivation is very different from Elga's. Moreover, as we will see, it poses a threat which is independent of Elga's challenge. Elga's challenge requires that his imagined world competes with Lewis's own candidates for closeness to the actual world. In contrast, the *mere possibility* of Bennett worlds menaces Lewis's semantics. Closeness does not matter. Curiously, Bennett himself denies that Lewis has to bother. So does Lewis, but for very different reasons. And both, I contend, are wrong: Bennett worlds spell trouble for a core tenet of Lewis': the asymmetry of post–determination. The asymmetry of post–determination in turn is required for Lewis's default ordering of similarity to support the right counterfactuals. I begin with outlining the purported asymmetry of post–determination and its role in Lewis's argument.

The asymmetry of post–determination

I confine my attention to determinism: two worlds which abide by the same laws either always or never perfectly match in facts. I repeat Lewis's similarity ordering of worlds for the sake of convenience:

'(1) It is of first importance to avoid big, widespread, diverse violations of law [big miracles].
(2) It is of second importance to maximize the spatio–temporal region throughout which perfect match of particular fact prevails.
(3) It is of third importance to avoid even small, localized simple violations of law [small miracles].
(4) It is of little or no importance to secure approximate similarity of particular fact, even in matters that concern us greatly.'[168]

Lewis purposively builds no asymmetry of time into his metrics. But such an asymmetry arises from the contingent features of worlds *like ours*, says Lewis: given the actual past, a contrary–to–fact antecedent P can be brought about in a very short time span by a small miracle. The small miracle is an event e_m which is unlawful, judging from the actual laws, but not from the laws of the closest antecedent world. Before e_m, the actual world w_0 and the closest P–world w_P perfectly match. However, there is no comparable small miracle which could *undo* P and achieve perfect match afterwards. The reason is that in worlds *like ours*, any event leaves a plethora of *traces*. These traces amount to *post–determinants*.

Consider:

(A33) If Nixon had pressed the button, there would have been a nuclear holocaust.

To Lewis, the antecedent event takes just a small miracle. The miracle brings about a divergence from some suitable point in actual history such as to make Nixon press very shortly afterwards. In Lewis's example, some additional neurons fire spontaneously in Nixon's brain. Maximizing perfect match in particular facts weighs more heavily than a small miracle. Thus, the closest world w_N where Nixon presses the button will be a world which is perfectly like the actual world w_0 up to a small miracle. This miracle leads to the antecedent event.

Yet Lewis denies that the same recipe can be applied after the antecedent has occurred. The reason is that

[168] Lewis, 'Counterfactual Dependence', pp. 47–48.

'[...]Nixon's deed has left its mark on the world[...] There are his fingerprints on the button. Nixon is still trembling[...] His gin bottle is depleted. The click on the button has been preserved on tape. Light waves flew out of the window, bearing the image of Nixon's finger on the button, are still on their way into outer space. The wire is ever so slightly warmed where the signal current passed through it. And so on, and on, and on.'(p. 45)

It would take *a big miracle* to undo all the traces Nixon's deed has left. A big miracle would consist of many small unlawful events (relative to the actual laws) of different kinds. Take a world where Nixon presses but which perfectly reconverges to the actual one. The perfect future match thus gained cannot outweigh a big miracle. In contrast, take a world where the lawful consequence of pressing, a nuclear holocaust, ensues. This world is more similar than a reconvergence world with a big miracle. The result is the intuitive asymmetry of counterfactual dependence: future events counterfactually depend on past ones, but (usually) not the other way round.[169]

Why does it take *unlawful* events to undo all these many traces? Why can't most of them be disposed of in a lawful way? Here the asymmetry of *post–determination* becomes crucial. Many of the traces envisaged by Lewis are not mere traces. A trace, I presume, is whatever can be used as evidence for its cause. Lewis's traces additionally give rise to partial post–determinants. Under determinism, any complete cross–section of the world and the laws entail any other. Over and above this entailment relation, the past is vastly overdetermined by the future, says Lewis. For any normal event e, there are many independent post–determinants at any point in time after e has occurred. A post–determinant is a set of facts which together with the laws of nature *entails* e. Many independent minimal cross–sections of the world amount to post–determinants over and above the complete cross–sections.[170] But usually there are relatively few minimal *pre*–determinants *before* e has occurred which together with the laws entail e. Lewis's favourite example are parts of a circular wave. These parts are taken to post–determine a point-sized source as their origin (p. 50). The *symmetric* fundamental natural laws would also allow for waves contracting inwards and so *pre*determining a point-sized target. But this *de facto* does not happen. A note in passing: the wave and the related cases I am going to present are of course just toy examples. In order to really entail an earlier event, a realistic minimal post–determinant would normally have to be overwhelmingly complex. But I am confident that nothing hinges on the details of the toy examples.

Lewis relies on the asymmetry of post–determination to ensure that reconvergence of antecedent worlds to the actual one usually needs a big miracle. Any event e which occurs actually but not in some contrary–to–fact antecedent world leaves many post–determinants. The (antecedent +) reconvergence world would have to perfectly match the future of the actual world. Consider facts which actually post–determine e as far as they occur after convergence. The reconvergence world must display these facts. But *at the reconvergence world*, they must not post–determine e. e has not happened. In contrast, the actual post–determinants plus the actual laws *entail* e. Thus, a break in the actual laws (i.e. different laws) is necessary for e not to have happened in the reconvergence world. So if there are sufficiently many and varied independent minimal post–determinants, reconvergence will need many and varied small miracles. These varied miracles add up to a big one.

The asymmetry of post–determination is of crucial importance to the success of Lewis's default similarity metrics. Without this asymmetry, Lewis could not tell why it takes a big, widespread *violation* of the actual laws to undo the many traces of a normal antecedent event. He could not exclude that almost all traces are lawfully effaced. For example, he could not exclude that the closest world where Nixon presses the button is a reconvergence world: almost

[169] I incur no commitment as to how facts, events, and ways of talking about them are to be understood.

[170] Respectively the intersection of any complete cross–section of the world at a time with the light cone of e in a relativistic world. Independence of cross–sections requires that they do not share any crucial part or ancestry where they could all be undone by *one* small miracle occurring after e.

perfect future match would be achieved by lawfully effacing almost all the many traces of Nixon pressing the button. Under determinism (as understood by Lewis), admittedly this match could not be perfect. Perfect reconvergence always needs a miracle, but without vast overdetermination of past facts by partial future cross–sections, that miracle might be a small one. Without the asymmetry of post–determination, Lewis's default similarity metrics would not deliver the asymmetry of counterfactual dependence. The asymmetry of post–determination, I am going to show, is put into doubt by the very possibility of Bennett worlds. Lewis feels the imminent threat, but he mistakenly thinks there is a simple remedy: distinguishing Bennett worlds from worlds *like ours*.

Bennett worlds

Here is Lewis's recipe of a Bennett world:[171]

'Begin with our base world w_0, the deterministic world something like our own. Proceed to w_1, the world which starts just like w_0, diverges from it by a small miracle, and thereafter evolves in accordance with the laws of w_0. Now extrapolate the latter part of w_1, backwards in accordance with the laws of w_0 to obtain what I shall call a Bennett world. This Bennett world is free of miracles, relative to w_0. That is, it conforms perfectly to the laws of w_0; and it seems safe to say that these laws are the laws of the Bennett world also. From a certain time onward, the Bennett world and world w_1 match perfectly, which is to say that w_1 converges to the Bennett world. Further, this convergence is brought about by a small miracle, the very same small miracle whereby w_1 diverges from w_0[...] Thus the Bennett world is a world to which convergence is easy, since w_1 converges to it by only a small miracle.'(p. 56)

The thrust of a Bennett world is the following: take a candidate world w_1 which diverges from a base world w_0, let it be the actual one, by just a small miracle. Take the future of that world. Revert that future so that it becomes the past, which is compatible with the symmetric laws. What we get is a world on which w_1 by just a small miracle.

Bennett worlds would be a counterexample to the claim that there can be no lawful convergence of worlds. Yet Bennett himself denies that Lewis has to bother:

'All he claims is that it takes a large miracle to produce a convergence between the actual world (or one like it) and *a plausible candidate for the title of closest P–world*, where P is the antecedent of any counterfactual we are trying to evaluate; and I have no argument to show that any of *those* convergences could be produced by a small miracle.'[172]

The Bennett world w_B does not converge to the *actual* world w_0 but to the *antecedent* world w_1. w_B cannot compete with w_1 for closeness to the actual world w_0. Unlike w_1, it displays a huge mismatch in pre–antecedent facts. And it displays mismatch in post–antecedent facts. So it seems that Lewis does not have to bother. The question becomes: why does he nevertheless take pains at discarding Bennett's example?

Lewis does not precisely tell what difficulty the Bennett world would pose. He contents himself with arguing that Bennett worlds are not worlds 'like ours' (p. 58). To begin with, unlike our world, Bennett worlds are *deceptive*:

'A Bennett world is deceptive. After the time of its convergence with w_1, it contains exactly the same apparent traces of its past that w_1 does; and the traces to be found in w_1 are such as to record a past exactly like that of the base world w_0. So the Bennett world is full of traces that seem to record a past like that of w_0.'(pp. 57–58)

In what ways is the Bennett world deceptive? Bennett helps us to understand what Lewis has

[171] Jonathan Bennett, 'Counterfactuals and Temporal Direction', *The Philosophical Review*, 93 (1984), 57–91.
[172] Bennett, *Conditionals*, p. 64.

in mind. Assume w_B converges to w_1:

'At least for a while after T the history books, geological and archaeological records, etc. of w_0 are exactly like those at w_1, the world at which through a small miracle Nixon presses the button (If not, then we can say that if Nixon had pressed the button the history books and other records would instantly have been changed; which is absurd) [...] their long-term pasts are grossly different; so w_B must be chock–full of deceptive records in a way that w_0 is not.'[173]

The Bennett world is deceptive: its pre–antecedent section largely diverges from the common past of w_0 and w_1. Nevertheless its post–convergence section from some point onwards contains all the post–antecedent facts which the actual world w_0 shares with w_1 (i.e. facts which are not abolished by the latter's divergence from w_0). Indeed this does seem very unlike our world. Our world, we think, contains reliable information which points to *our* past. In contrast, w_B contains information which merely *purports* to point to our past.

Lewis presents an even stronger claim: some of the information about the actual world w_0 as preserved in w_1 amounts to actual post–determinants of events at w_0. But as far as w_B does not encompass the actually post-determined events (due to its past divergence from w_0), it cannot contain the latter's post–determinants either:

'To be sure, any complete cross section of the Bennett world, taken in full detail, is a truthful record of its past; because the Bennett world is lawful, and its laws are ex hypothesi deterministic (in both directions), and any complete cross section of such a world is lawfully sufficient for any other. But in a world like w_0, one that manifests the ordinary de facto asymmetries, we also have plenty of very incomplete cross sections that postdetermine incomplete cross sections at earlier times. It is these incomplete postdeterminants that are missing from the Bennett world. Not throughout its history; but the postdetermination across the time of convergence with w_1 is deficient.'(pp. 57–58)

So Lewis has given good reasons to deem the Bennett world very peculiar. It looks like a giant conspiracy, some Potemkin village deluding its wretched inhabitants into thinking they are in w_0. This is not a world like ours.

Bennett worlds vs. post–determination

We might feel comforted. Bennett himself grants that w_B is no menace to Lewis's metrics. And Lewis has shown that it is a world not like ours. Ought we to feel comforted? I doubt it. Lewis has given no reason to deem w_B *impossible*. In fact, he apparently regards it as perfectly possible. We can extrapolate w_1, the closest world where Nixon presses the button, forwards from Nixon pressing the button in accordance with the laws common to w_1 and w_0. There is only one difference in these laws: at w_1, they allow for the miracle to bring about the pressing event. Now take the cross–section of w_1 at the time when Nixon presses the button. Surely we can also extrapolate backwards from that cross–section in accordance with the laws of w_0.[174] The small miracle lies in the past of the pressing event. So taken in isolation, the cross–section of w_1 when Nixon presses the button seems perfectly in tune with the laws of w_0. The laws are assumed to be symmetric. So if there is any problem with extrapolating backwards, there should also be a problem with extrapolating forwards in according with the actual laws (to derive the future of Lewis's w_1).

Now if w_B is possible, the asymmetry of post–determination is in bad shape (at least post–determination from those post–antecedent regions where w_0 and w_B match, i.e. do not come apart due to the small miracle, Nixon's pressing and its horrid aftermath). We have seen

[173] Bennett, *Conditionals*, p. 295.

[174] which, I repeat, are the laws of w_1. There is only one difference: what is a small miracle, judging from the laws of w_0, is perfectly lawful at w_1. The mere event of Nixon pressing the button does not violate any law.

that Lewis wants to account for the small miracle by an exception clause (pp. 54-55, see below). For any point in spacetime outside the precise spatiotemporal location of the small miracle, the laws of w_0 and w_B are the same. I propose that this isolation of the small miracle gives rise to the following claim: as far as facts which are partial post–determinants for some pre–divergence event e at w_0 are preserved at w_B, the same goes for their status as partial post–determinants. If they post–determine e at w_0, they also post–determine e at w_B.

Consider the following exemplary pre–miracle event: at some time t briefly before the miracle, Nixon was calm. But assume that, after pressing the button, he was very excited. Call his actual calm state of mind at t S_C, his extremely excited state of mind S_B. Assume S_B would have to have occurred at w_B at the same time at which S_C occurred at the actual world and the Bennett world. If there is overdetermination from the future, there will be many actual post–determinants for S_C. Several of them survive the dramatic events the small miracle initiates at w_1.[175] In his argument that w_B is not like our world, Lewis must invoke these post–determinants. For otherwise he could not explain why w_B is deceitful, lures us into thinking that, say, S_C has obtained while S_B has (and the same for other events before the time of the miracle).

As an example of such an undisturbed post–determinant, take the electromagnetic waves emitted by S_C on their way into outer space. Or take an encephalogram of Nixon which is transmitted to some safe place (Brezhnev's bunker, say). The Bennett world w_B, perfectly matching w_N, the worlds closest to actuality at which Nixon presses the button, after the latter converges on it, will contain precisely those images (post–determinants) of Nixon's actual state S_C which are shared by w_0 and w_N. Yet at w_B Nixon was in a quite different state S_B. Since there is no relevant difference in laws between w_B and the actual world, w_B must break the lawful post–determination relationship between Nixon's actual (w_0) state S_C and the later images of S_C. At w_B, S_C does not occur (Nixon is not calm) but the later images of S_C do (the encephalogram showing the calm state). Thus, w_B must obey different laws. But it is assumed to perfectly abide by the very same laws of w_0 which ensure the lawful post–determination relationship. So w_B *cannot* break the relationship.

One has to go, either the Bennett world or lawful post–determination by the images of Nixon's state of mind. And the same argument applies to any actual event e_{Db} which occurs at the actual world w_0 but not at the Bennett world w_B: the possibility of w_B shows that no fact which is preserved in the closest button-pressing world w_N (and thus w_B) actually amounts to a post–determinant of e_{Db}. This is, I surmise, what bothers Lewis about w_B and has him insist that it is not a world *like ours*. But Lewis does nothing to dissolve our dilemma: one has to go, Bennett worlds or the asymmetry of post–determination. It is not the closeness of w_B to w_0 but the very possibility of w_B that is irreconcilable with the asymmetry of post–determination. And if w_B is possible (as anyone involved in the debate seems prepared to grant), it is clear that partial post–determination has to go. To repeat: for any actual event which is different in the actual world and the Bennett world, this event cannot be post-determined by some partial cross–section of w_0 *as far as* that cross–section is preserved at w_N (and w_B).

We may deny that this blow is fatal to Lewis's argument from post–determination. After all, many actual partial post–determinants remain: facts at w_0 which would not have to be preserved in w_N and w_B because the diverging events initiated by the small miracle interfere. For instance, Nixon's actual calm state S_C may be recorded at the actual world w_0 by his official biographer who enjoys privileged access to the leader of the free world. But alas, at the

[175] Assume that, our world being relativistic, precisely in the light cone of the small miracle, there is no perfect match between w_0 and w_1. Thus, all post–determinants of pre–antecedent events at w_0 which do not lie within the light cone are preserved at w_1. As far as they lie after the time of convergence with w_1, they must be preserved at w_B, too. Moreover, even in the region without perfect match, a great many particular matters of fact will be preserved. Even a very forceful antecedent event like Nixon's pressing the nuclear button will leave many actual facts unaltered, waiting for a post–apocalyptic historian to recollect the days before the nuclear holocaust. One should expect all these facts which are preserved at w_1 and w_B together with facts outside the miracle's light cone to provide a sufficient basis for overwhelmingly many post–determinants at w_0.

holocaust worlds w_N and w_B, he is among the first victims before he can record Nixon's state of mind. The extremely faithful and reliable record may serve as a post–determinant of Nixon's state of mind at the actual world without disturbing noise from w_B.

Not so. For the problem spreads. Candidates for counterfactual antecedents are many and varied. Any possible state of affairs can serve as the antecedent of countless non-vacuously true counterfactuals. And as far as such a state of affairs can be reached by a small miracle, the recipe of the Bennett world can be applied. Bennett worlds abound. Thus we have a general recipe against post–determination. I do not maintain that the recipe will work everywhere, but its range of application is wide. Any candidate d for post-determining a normal event e can be ruled out, provided there is a contrary–to–fact antecedent proposition P which fulfils two conditions: (i) there is a Bennett world which converges to the closest P–world; due to the Bennett world's past mismatch with w_0, e does not occur at the Bennett world. (ii) e and d are preserved at the closest P–world; due to the latter's perfect future match with the Bennett world, d does occur at the Bennett world. There are reasons why these conditions are easily fulfilled. Any event can be subject to a counterfactual claim. And for this counterfactual claim, a Bennett world can be conjured up which converges to the closest antecedent world. So the history of the world between e and its purported post–determinant d abounds with opportunities for Bennett worlds which contain d but not e.

Summarizing, Bennett worlds shed grievous doubts on Lewis's original argument for the asymmetry of miracles. This does not mean that there is no asymmetry of post–determination.[176] There might be problems with other premises of Lewis', for instance the very idea of small and big miracles. These premises drive his acceptance of Bennett worlds. My point is just that Bennett worlds pose a more dire threat to Lewis's argument than Bennett and Lewis suppose. Bennett worlds target the very heart of Lewis's reasoning. The explicit aim of *Counterfactuals and Time's Arrow* is to account for 'the mysterious asymmetry between fixed past and open future' in terms of the asymmetry of counterfactual dependence (p. 38). And the argument from the asymmetry of post–determination is crucial to derive this asymmetry of counterfactual dependence (given the perfectly symmetric similarity metrics). Bennett worlds threaten this core argument of Lewis.

Are Bennett worlds possible, after all?
Problems with miracles

One big question deserves closer scrutiny: is Lewis really committed to the possibility of Bennett worlds? The laws of the closest counterfactual world w_1 are the laws of the actual world w_0. There is only one difference: one event which from the perspective of the actual laws is unlawful, a small miracle, is perfectly lawful from the perspective of the laws of w_1. Lewis does not specify what the laws of w_1 look like. But in one place he discusses what could be the

[176] There are independent reasons to doubt the asymmetry of post–determination, however. Jeffrey Dunn questions the wave–example:

'A set of propositions about a small portion of the wave, however, is not sufficient for its emission from a point. To get sufficiency, we must add the further information about what is happening outside this region. Perhaps, for example, the small part of the wave is not a part of a spherical wave at all, but merely a part of space that is identical to what this part of the wave would be like, were there a wave.'(Dunn, p. 78)

The explanation of a partial cross–section of the world is sensitive to the rest of this cross–section, just as the explanation of a portion of a spherical wave is sensitive to information about the rest of the wave. Dunn even intimates that for any normal event, there is only one minimal post–determinant at a time: the complete cross–section of the world (respectively the complete cross–section within the light cone). But his argument does only show that we should be wary of too simple post–determinants. I surmise that Dunn has already in mind Elga's counterexample to the asymmetry of post–determination. I will come to that example in due course.

blueprint of these laws. Lewis compares the laws of three worlds. One is a world with our laws, let it be the actual world w_0. The second is described as follows:

'the best way to write down its laws would be to write down the laws of the first world, then to mutilate them by sticking in clauses to permit various exceptions in an unprincipled fashion. Yet almost everything that ever happens in the second world conforms perfectly to the laws of the first.'[177]

The third world is a world with elegant, simple, powerful laws like ours, 'except for a change of sign here, a switch from inverse square to inverse cube there, and a few other such minor changes.'(p. 55) The laws of the third world in themselves look more similar to those of the first world than the laws of the second. Nevertheless, the first and the second world are more similar to each other, says Lewis. For the elegant, simple laws of the third world lead to huge changes in particular matters of fact. Hence a world where the laws are subject to exception clauses can be closer to ours although these laws look gerrymandered. So Lewis's picture must be that the laws of w_1 are the laws of w_0 save for an exception clause. The exception clause does not surface except where the small miracle e_m occurs. It must not only allow for the miracle but (together with the facts of w_1) determine it. The rest of w_1 perfectly obeys the laws of w_0. This is why Lewis says 'it seems safe to say that these laws are the laws of the Bennett world also.' If w_1 apart from e_m perfectly abides by the laws of w_0, the symmetry of the laws of nature requires that one can extrapolate backwards from any post–e_m cross–section of w_1 in accordance with the laws of w_0.

Lewis's treatment of the small miracle e_m has sparked criticism. e_m is miraculous from the perspective of w_0. And it must be amazing from the perspective of w_1, too. Both worlds share the same history up to e_m. Hence even a super–scientist at w_1, knowing all of history before e_m and possessing ideal capacities of theory–formation, could not have the slightest suspicion that e_m would occur. These vague worries have recently been more thoroughly spelled out by Stephen Barker. When we write a suitable exception clause such as to allow for e_m into the law,

'[...]we are assigning physical significance to bare particularity itself. That's objectionable. In physics we do not attribute nomic relevance to bare particularity: physical objects have the status of bundles of generic properties.'[178]

It won't do to assume that some *alien* quantity surfaces just once in the whole universe, says Barker. A lawful relationship would require certain systematic variations of the quantity in question under slightly varying conditions.[179] It seems weird to claim that such variations of the conditions which determine e_m are never instantiated in w_1. And lossy laws which can be violated are not reconcilable with determinism.[180]

One may use these considerations to question the whole construal of w_B (w_B perfectly conforms to w_1 in post–antecedent facts while abiding by the actual laws). Assume the laws of w_1 cannot be the laws of w_0 plus an exception clause. There must be a more substantial difference in laws. This difference in laws is likely to be manifested in the future of w_1. Since the Bennett world w_B abides by the actual laws and w_1 by substantially different laws, the former cannot simply roll back the future of the latter. The ubiquity of Bennett worlds can be denied. However, the premise that the laws of w_1 substantially diverge from the actual laws is fatal to Lewis's conception of miracles. For the substantially different laws of w_1, which is supposed to be the closest antecedent world, would give rise to big, widespread miracles, judging from the actual laws. Lewis's whole metrics under determinism is built around the

[177] Lewis, 'Counterfactual Dependence', p. 55.
[178] Stephen Barker, 'Can Counterfactuals Really Be about Possible Worlds?', *Noûs*, 45 (2011), 557–576, (p. 567).
[179] Barker, 'Counterfactuals', p. 568.
[180] Barker, 'Counterfactuals', p. 569.

distinction between small and big miracles. Thus, he has no resources to rule out the possibility of Bennett worlds.

The parallel to Elga's counterexample

We might still share Lewis's amazement: the Bennett world abounds of traces which point to events which did not occur. How could it manage to manufacture all these traces? As we have seen, Adam Elga has presented a closely related counterexample to Lewis's analysis.[181] It may reveal the deeper physical rationale of Bennett worlds. Assume the laws of deterministic statistical mechanics are the only ones that count. Any process which exhibits the normal thermodynamic asymmetry, i.e. an increase in entropy, can be approximated up to a small miracle by a completely different, thermodynamically abnormal process. Take any normal antecedent: there will be an antecedent world which converges to the actual one by a small miracle.

We should expect a Bennett world to display a trace-making process of the sort envisaged by Elga. Such a process is the only candidate which explains all the misleading traces of a past that is not past of the Bennett world. Nevertheless, the Elga world is not a Bennett world. Firstly, the Bennett world w_B converges to w_1, Elga's world converges to the actual world w_0. Moreover, the Bennett world *perfectly* abides by the laws of w_0 while Elga's world does not. There is a small miracle. Nevertheless, the comparison of both worlds is fruitful. It also allows us to show why Bennett worlds are interesting in their own right: while Elga aims at presenting a world which competes with divergence worlds for closeness to the actual world, Bennett worlds do *not compete* for closeness. Their mere *possibility* poses difficulties to Lewis's asymmetry of post–determination.

I look at strategies to defend Lewis against Elga. It is interesting to see whether they work against Bennett worlds, too. Most of these counterstrategies concern the claim that Elga worlds compete with divergence worlds for *closeness* to the actual world.

The most prominent response in the literature is Schaffer's, which I already rehearsed in the chapter on Morgenbesser cases: match in facts should not count in the closeness ordering as far as facts are causally dependent on whether the antecedent obtains or not. This demotes the Elga world from being a closest antecedent world, provided its advantages in terms of perfect future match are confined to the region which is causally dependent on whether the antecedent obtains or not. But Schaffer's amendment does not save the asymmetry of post–determination from Bennett worlds. For it is not their closeness but their very possibility that conflicts with the asymmetry of post–determination.

There is another line of response. Bennett reports Lewis's own reaction to Elga's counterexample:

'The worlds that converge onto worlds like ours are worlds with counter-entropic funny–business. I think the remedy – which doesn't undercut what I'm trying to do – is to say that such funny–business, though not miraculous, makes for dissimilarity in the same way miracles do.'[182]

Assume the Bennett world exhibits an anti-thermodynamic process of the sort developed by Elga. Then this passage also enshrines Lewis's last word on Bennett's counterexample. Still Lewis does not question the very possibility of Bennett or Elga worlds. Rather his strategy is to assimilate the role of counter-entropic funny–business to that of miracles. To save Lewis's from counterentropic worlds as competitors for closeness, no amount of match in particular facts must be able to compensate for such funny–business. For if the world stretches further into the future than into the past, the future match of the Elga world may largely outdo the past match

[181] Elga, 'Statistical Mechanics'.
[182] Bennett, *Conditionals*, p. 296.

of Lewis's candidate worlds. The role of funny–business in the similarity metrics must be comparable to the role of a big miracle. Unfortunately, Lewis did never provide a general criterion for funny–business. And even if he had succeeded in demoting counterentropic worlds from closeness, this would not have saved the asymmetry of post–determination. For the very *possibility* of Bennett worlds is sufficient to refute it.

The parallel to Elga's thermodynamically reversed processes underpins the microphysical possibility of Bennett worlds. But it also shows what is puzzling about them. From an isolated viewpoint of statistical mechanics, our world with its characteristic thermodynamic asymmetry seems very amazing. There is a proposal which might indeed take care of both Elga's and Bennett's challenges. If the Past Hypothesis is granted the status of a fundamental *law*, it introduces a fundamental asymmetry. The Past Hypothesis breaks the symmetry between the past and the future. In fact, any explanation where a fundamental nomic necessity underlies the thermodynamic asymmetry is likely to conflict with Bennett's and Elga's template. The Elga world would be demoted from closeness. As for Bennett worlds, surely a great portion of them could be ruled out as unlawful. They are *metaphysically impossible* because they cannot both share our fundamental laws *and* perfectly converge to some closest antecedent world like w_1. While some uncertainties would remain, this might be sufficient to uphold the asymmetry of post–determination. *If there is a necessary fundamental asymmetry*, the possibility of any particular Bennett world is doubtful. Lewis's suspicion that counter-entropic funny–business makes for dissimilarity could be confirmed without altering his metrics. Funny–business makes for dissimilarity because it violates fundamental laws on a large scale. However, as we have seen, there are many reservations about this solution.

Is there smooth convergence to the actual world?

I shall discuss a more daring claim: for any counterfactual antecedent world, there is a Bennett world that smoothly converges on it. Moreover, there are amazing counterentropic worlds like Elga's, which converge on the actual world by a small miracle. Yet couldn't it be that there are also worlds that *smoothly* converge on the actual world without any counterentropic funny–business. To make this idea more vivid, I develop an argument of my own why convergence lurks everywhere. All the counterexamples considered so far demand a peculiar arrangement of facts, in the actual world or in the counterfactual worlds under discussion. The counterfactual Elga world contains an amazing antithermodynamic region; so it can be ruled out by stigmatizing thermodynamically awkward situations. Other examples discussed explicitly display particular configurations in the actual world like the beetle in the box; so they are not *pervasive*. In contrast, the worlds I imagine are intrinsically perfectly inconspicuous. And the recipe guiding their production may be applied anywhere; they are candidates for closeness whatever the antecedent situation considered in a counterfactual is. This casts doubt on any normal counterfactual, construed à la Lewis. Moreover, smooth convergence as opposed to Elga's amazing convergence provides a good opportunity for a principled discussion of Lewis's overall anti–convergence arguments.

As an example of two set–ups which are qualitatively identical in all relevant respects save one, I contrive two containers of gas. Assume for simplicity that the gas is energetically completely isolated from the environment.[183] At t_0, the two sets of gas molecules in the two containers are qualitatively identical, there being only one difference. For each molecule in one container *save a small local group*, there is a perfect physical counterpart in the other (same relative location, velocity and so on). The tiny difference the diverging molecules make will soon spread until at t_1 the two containers are temporarily very unalike. Now consider the following case: at t_2, a *very* long time has passed for the containers to develop in accordance

[183] Sklar, p. 669.

with the laws of thermodynamics. Both containers are intrinsically perfectly inconspicuous, the gas is evenly distributed and so on. Yet at t_2, both are qualitatively almost identical, save some small local group of molecules. Only some tiny deviation from the actual laws would be necessary to bring about perfect correspondence. The velocities of the individual molecules are not precisely reversed as in the Elga world; instead, an extremely smooth and inconspicuous transition over a very long time span leads from divergence increasing to divergence decreasing. Starting from a tiny divergence, we eventually arrive at a comparably tiny divergence. Such a situation will be astronomically improbable compared to the overwhelming majority of alternative, more strongly divergent configurations of molecules which realize the general container scenario. Still it is possible in the following sense: while under determinism, the detailed initial conditions constrain the lawful development of the containers to uniqueness, my *general* description of the example is reconcilable with many such developments, one of them being (near) convergence. If we consider sufficiently many pairs of containers which conform to my general description, some of them will qualitatively converge to each other up to a small miracle. The very idea of smooth convergence is to use time to smoothen a development that would otherwise have been amazing and in need of a big miracle. Time is traded for amazingness.

Now instead of two actual containers, imagine an actual container and its counterfactual variation. The small divergence is due to a small miracle. The container being energetically isolated, the rest of the world makes no difference. We seem bound to reject

(E25) If there had been a tiny group of divergent molecules in the container, after sufficient time the latter would have been completely as it will actually be.

Yet judging from Lewis's criteria, we cannot simply reject this conditional; for just as in the example of the two actual containers, there is the possibility of the counterfactual container smoothly converging to the actual one.
There are two main arguments against smooth convergence.
Convergence worlds violate *ubiquity of traces*.
Convergence worlds are microphysically impossible for other reasons than the ubiquity of traces.

As for the first Lewis rules out convergence by a small miracle. Either convergence worlds display an enduring lack of *perfect* match in post–antecedent facts, or perfect match is achieved at the price of a big, widespread miracle (p. 46). Once it has somewhat spread, divergence cannot disappear without a big miracle:

'Because there are many different sorts of traces to be removed, and because the traces spread out rapidly, the cover–up job divides into very many parts[…] Different sorts of unlawful processes are needed to remove different sorts of traces.'(p. 47)

The one argument of Lewis's that would substantiate this claim is *overdetermination from the future*: the facts at our world are such that for any normal event e, there is a plurality of different partial cross–sections of the future which together with the laws entail e. This lawful entailment can only be contravened by breaking the laws. For any of these cross–sections, it takes at least a small miracle to undo it given the event e. The small miracles needed to undo many such cross–sections add to a big miracle. When there is sufficient over–determination from the future, there is just no way of achieving convergence by a small miracle (p. 57). However, as we have seen, Bennett and Elga worlds refute the asymmetry of post–determination.

In sum, there may well be completely innocent, non-amazing, smooth convergence worlds. If such there are, it is highly doubtful that one of the particular remedies discussed so far is of any avail against them.

2.5. Typicality

We have already seen that the difference between amazing and high–probability events may play a role in saving Lewis's similarity metrics from new versions of the future similarity objection. There is a close relationship to the role of typicality in judging counterfactuals. I shall discuss this role with regard to a debate in the philosophy of thought–experiments, which is closely entangled with a highly relevant use of counterfactuals.

Thought experiments are difficult to understand. In his *The Philosophy of Philosophy* Timothy Williamson has come up with a new puzzle and a proposal how to dissolve it.[184] Any thought experimental description can be realized in a deviant way. Williamson suggests that the problem can be solved if thought experimental reasoning is analysed by counterfactuals. I defend the counterfactual account against two lines of criticism and three alternative proposals forwarded by Anna–Sara Malmgren, Jonathan Ichikawa, Benjamin Jarvis, Thomas Grundmann and Joachim Horvath. I present an interpretation of the pertinent counterfactuals.

The problem of deviant realizations
GC1
At 8:28, somebody looks at a clock to see what time it is. The clock is broken; it stopped exactly twenty–four hours previously. The subject believes, on the basis of the clock's reading, that it is 8:28.

What has been called *the Gettier intuition* is '[…]loosely put: the judgement that [the subject satisfying GC1] has a justified true belief without knowledge [NKJTB] […]'.[185] As a consequence, knowledge cannot simply be justified true belief. The case description directly prompts the reaction that the subject has NKJTB. But neither does the description explicitly say so, nor is there any straightforward entailment relation between the explicit description and the prevalent verdict. The step from the former to the latter is far from trivial. For some situations where the description is fulfilled, it goes wrong. This can be illustrated by the following completion of the story:[186]

GC2
At 8:28, somebody looks at a clock to see what time it is. The clock is broken; it stopped exactly twenty–four hours previously. The subject believes, on the basis of the clock's reading, that it is 8:28. *The subject knew in advance that the clock had stopped exactly twenty–four hours previously*.

The situation described in GC2 is perfectly compatible with the original Gettier description GC1. Yet in that situation, it is not the case that the subject has NKJTB. The subject *does* know the time. GC2 is just one example of a whole class of completions where the usual Gettier verdict does not apply. Another class are completions where the subject lacks justification.

Solution: The Counterfactual Account

[184] Timothy Williamson. *The Philosophy of Philosophy* (Oxford: Blackwell, 2007).
[185] Anna Sara Malmgren, 'Rationalism and the Content of Intuitive Judgements', *Mind*, 120 (2011), 263–327 (p. 264).
[186] Timothy Williamson, 'Replies to Ichikawa, Martin and Weinberg', *Philosophical Studies*, 145 (2009), 465–476 (p. 467).

How are we to analyse our reasoning from the case description to the refutation of the JTB–theory in light of deviant realizations? As a first stab, the following formalization is appealing.[187] Start with the JTB–analysis:

(F1) Necessarily, someone knows some proposition P if and only if she has justified true belief in *p*.

(F1) is refuted by:

(F2) Possibly, someone stands to some proposition P in the relation described by GC1.
(F3) Necessarily, if someone stands to some proposition P in the relation described by GC1, she has a justified true belief in P without knowing P.
(F4) It is possible that someone has justified true belief in P without knowing P.

The argument falls prey to the problem of deviant realizations. Among the situations over which the strict conditional (F3) ranges, there is a situation as described by GC2. Since in a GC2–situation, the subject does not have NKJTB, (F3) is false. Its failure gives rise to a strategy of dealing with deviant realizations: find a repair which explains and justifies our confident endorsement of the Gettier verdict notwithstanding deviant completions.

Williamson proposes to replace the strict conditional (F3) by a counterfactual:

(F3*) If a thinker were related to a proposition P as described by GC1, she would have justified true belief in P without knowing P.[188]

The main advantage of (F3*) compared to (F3) is that it is not falsified by the mere possibility of a deviant realization. According to the mainstream semantics of counterfactuals, (F3*) just requires that the subject has NKJTB in all closest GC1–situations.

Some clarifications about the precise aim of the formalization are in order. Firstly, it applies to thought experiments in general, at least those refuting some necessity claim. Secondly, the common aim of all participants in the debate is to represent '*our actual route',*[189] a formalization that establishes knowledge of (F4) and is psychologically plausible; it is sufficiently close to how a competent thought experimenter *normally* comes to accept (F4). I summarize the common aim as giving the *normal route*. It is crucial to keep this in mind for the following reason: GC1 might *actually* be realized in a deviant way. There are doubts as to whether the rational route to (F4) is the same whether that epistemic possibility is salient or not. To neutralize these doubts, I shall consider the situation of a normally competent pre–Williamsonian thought experimenter sincerely testing the JTB–theory, say Gettier himself. My claim is not that, after Williamson, we read GC1 differently. I just want to remain neutral whether we read it differently in a context where deviant realizations are salient.

I shall argue that some of the competing formalizations are rather *fallback* positions than representing the normal route. If we look for a fallback position, I offer the following replacement of (F2) and (F3*):

(F2–F3°) there is an extended version v of GC1 such that,
(i) it is possible that someone stands to a proposition P as described by v,

[187] Williamson, *The Philosophy*, p. 183.
[188] Williamson, *The Philosophy*, p. 195.
[189] Malmgren, p. 283, emphasis mine. Malmgren's own explication is a bit misleading: 'more precisely, a rational route from our actual intuitive judgement [to the refutation of the JTB–analysis], one that is plausibly available to us.' Not any route from our actual intuitive judgement that is plausibly available to us is the actual route we normally take.

(ii) if a thinker were related to a proposition P as described by v, she would have justified true belief without knowledge that P.

Assuming that any particular deviant realization can be stipulated away, (F2–F3°) would provide a deviance–proof formalization. It can be supported by the implicit generality of intuitive reasoning: in considering a story like GC1, we grasp a general pattern.[190] So one might be expected to have a whole range of fitting cases within one's purview. Such a grasp may provide a justificatory basis for (F2–F3°). While (F2–F3°) is as plausible, considered as a *fallback* position, as the rival accounts to be discussed, I deem it psychologically implausible as an account of *our normal route*. It is not what we have in mind when dealing with the GC1 story.

Criticism in literature

Williamson admits that (F3*) is false if a deviant realization turns out to be actual (or closest). But, he says, we are disposed to come up with an amended story when GC1 unexpectedly fails.[191] For instance, we might react to GC2 being actual by writing into the antecedent of the counterfactual 'GC1 but not GC2'. However, this concession has not satisfied his critics. There are two main criticisms of the counterfactual account:

(i) It is unconvincing as a psychological account of our normal route.
(ii) It is epistemologically problematic: if this is how we actually reason, we cannot know that the JTB–analysis is false.

I shall discuss these criticisms in due order.

Psychological concerns

The psychological criticism has been forwarded by Malmgren. She construes the task of the formalization as identifying *the intuitive Gettier judgement*. The intuitive judgement states the lesson to be drawn from the case. Malmgren argues that (F3*) cannot be the intuitive verdict. For the problem with deviant realizations reappears. If someone *actually* satisfies GC1 in a deviant way, say by satisfying GC2, (F3*) is false. In that case, we are disposed to retract (F3*). But we are neither disposed to change GC1, nor do we retract the intuitive Gettier judgement based on GC1. And that shows that (F3*) cannot 'conform to our semantic intuitions about deviance'.[192] (F3*) does not capture how the description of the case was *meant*.[193]

I shall argue that there is no such clear psychological evidence against the counterfactual account. Firstly, I shall provide linguistic evidence that the intuitive judgement is sensitive to deviant realizations. Secondly, I shall question the purported role of a unique intuitive judgement in the formalization. Thirdly, I shall point to alternative explanations of our resilience to deviant realizations.

Malmgren claims that we uphold GC1 and the intuitive judgement when a deviant realization turns out to be actual. This view is not shared by all of Williamson's critics: as Jonathan Ichikawa admits, 'It is fairly natural to respond to this sort of challenge with a further

[190] cf. Malmgren, pp. 290–297.
[191] Williamson, *The Philosophy*, p. 204.
[192] Malmgren, p. 278.
[193] '[...]the envisaged (actual) realization of the case [...] is clearly deviant. It requires that we read the case description in a way we know it was not meant to be read.'(Malmgren, p. 279)

spelling–out of the case to be considered.'[194] In order to decide how we would respond, I use a linguistic test. The vulnerability of (F3*) to deviant realizations is confirmed by the following imagined dialogue, taking place in an epistemology class where GC1 is brought up as a Gettier story (I paraphrase the formal expressions to get the linguistic intuitions):

Dialogue10
John: 'Imagine someone who looks at 8.28 at a clock which broke exactly 24h earlier. (F6) She has justified true belief but no knowledge that it is 8.28.'
Mary: 'Not necessarily: the clock at the wall has actually stopped 24h earlier. Betty is looking at the clock to see what time it is, but I have just told her that the clock has stopped 24h earlier.'
John: (F3*): #'But if a thinker came to believe what time it is as described in my scenario, she would have justified true belief without knowing.'

It seems infelicitous to utter (F3*) once the actuality of a deviant realization has been raised to salience.[195] This shows that (F3*) is vulnerable to deviant realizations. If Malmgren is right, one should expect the intuitive Gettier judgement to behave differently. And indeed Malmgren's own formalization of the Gettier argument, which she designs to be invulnerable to deviant realizations, shows quite a different behaviour:

(F5) It is possible that someone stands to P as in the Gettier case (as described [by GC1]) and that she has a justified true belief that P but does not know that P.[196]

(F5) is perfectly fine within a variant of the epistemology class dialogue:

Dialogue11
John: 'Imagine someone who looks at 8.28 at a clock which broke exactly 24h earlier. (F6) She has justified true belief but no knowledge that it is 8.28.'
Mary: 'Not necessarily: the clock at the wall has actually stopped 24h earlier. Betty is looking at the clock to see what time it is, but I have just told her that the clock has stopped 24h earlier.'
John: (F5) 'But it is possible that someone in my scenario has justified true belief but does not know what time it is.'

So the test seems well-calibrated. Its outcome covaries with the sensitivity of the candidate formalizations to deviant realizations. Now if there is some independent informal way of expressing the intuitive judgement, we might use the test to check whether it behaves rather like (F3*) or (F5). I shall use 'the intuitive judgement […]loosely put: the judgement that [the subject satisfying GC1] has a justified true belief without knowledge[…]'.[197] The Gettier verdict, loosely put, is

(F6) A thinker who is related to proposition P as described by GC1 has justified true belief without knowing that *p*.

In an epistemology class, one can put the Gettier verdict in this loose way. The audience will understand what is meant. Consider how it fares in a variant of the epistemology class dialogue:

[194] Jonathan Ichikawa, 'Knowing the Intuition and Knowing the Counterfactual', *Philosophical Studies*, 145 (2009), 435–43 (p. 439).
[195] It seems that raising the mere metaphysical possibility of a deviant realization makes for infelicity. But this just shows that raising the metaphysical possibility of a deviant realization often will suffice to also raise its epistemic possibility.
[196] Malmgren, p. 281.
[197] Malmgren, p. 264.

Dialogue12
John: 'Imagine someone who looks at 8.28 at a clock which broke exactly 24h earlier. (F6) She has justified true belief but no knowledge that it is 8.28.'
Mary: 'Not necessarily: the clock at the wall has actually stopped 24h earlier. Betty is looking at the clock to see what time it is, but I have just told her that the clock has stopped 24h earlier'
John: (F6) #'But someone in my scenario has justified true belief without knowing what time it is.'

Unlike John's use of (F5), his use of (F6) seems infelicitous. So my test indicates that (F3*) is closer to (F6), the intuitive judgement, loosely put, than (F5). We may save (F6) if we read John as stipulating what his scenario is like, but this reading is not a matter of curse, and it does not serve the purpose of the thought–experiment. Whether someone has NKJTB should not be stipulated but follow naturally from considering the scenario.

There are many uncertainties about this result. Perhaps (F6) differs in its felicity conditions from more concise ways of putting the intuition. Moreover, the test shows only that one would *not utter* (F6) after deviant realizations have been raised to salience, not that one would *retract* (F6). But Malmgren owes an explanation why her own formalization (F5) behaves so differently from (F6), the loose way she herself puts the intuitive judgement. In my view, the most promising explanation is that, in a normal epistemological context, deviant realizations are unintended because they do not come to mind. Yet when they are raised to salience, they cannot simply be dismissed as beside the point. We feel a pressure to react, however confident we remain about the overall pattern of Gettier–style examples, as witnessed by the following epistemology class dialogue:

Dialogue13
John: 'Imagine someone who looks at 8.28 at a clock which broke exactly 24h earlier. (F6) She has justified true belief but no knowledge that it is 8.28.'
Mary: 'Not necessarily: the clock at the wall has actually stopped 24h earlier. Betty is looking at the clock to see what time it is, but I have just told her that the clock has stopped 24h earlier.'
John: #'Oh come on. That is not how I intended the story'.

Mary's reaction should seem somewhat inappropriate and John's reply perfectly in order if Malmgren were right. But what Mary says seems ok. And it would be much better for John to say: 'I see. But a subject who is not told that the clock has stopped 24h earlier has justified true belief without knowing that it is 8.28.' In sum, Malmgren's claim that actually deviant realizations do not bother us is contestable.

There are also doubts about Malmgren's demanding use of 'intuition'. Her argument targets the claim that (F3*) *is* 'the intuitive judgement'. The counterfactual account does not entail this claim.[198] So the argument that (F3*) behaves differently than the intuitive judgement does not prove the counterfactual account wrong. The requirement that the formalization is to identify some clear-cut intuitive judgement is not a matter of course. It comes with a certain bias in favour of Malmgren's own account: to her, thought experimental reasoning proceeds via rational *a priori* intuition. As a consequence, there must be some clear-cut judgement that can be intuited *a priori*. Unless one accepts the rationalist account, there is little reason to insist that the formalization must identify such an outstanding intuitive judgement. Herman Cappelen notes 'that it is exceedingly hard (I [Cappelen] argue impossible) to find a particular judgment (or set of judgments) in any of the alleged paradigmatic cases that there is agreement on

[198] Williamson himself states this claim, but in a rather detached way, indicating his reluctance about intuition talk: 'What is sometimes called 'the Gettier intuition' has been expressed by a counterfactual conditional in English…'(Williamson, *The Philosophy*, p. 195).

classifying as intuitive.'[199] Even if we grant that there is something like the Gettier intuition, it does not follow that one precise step of the reasoning *(F1), (F3*); thus (F4)* has to be identified with the intuitive judgement. Perhaps the best candidate for *the* intuitive judgement is the informal

(F6) A thinker who is related to proposition P as described by GC1 has justified true belief without knowing that P.

But just as there is no guarantee that the formalization must be vulnerable to deviant realizations in the way (F6) is, as witnessed by the above dialogues, nothing guarantees that there is a unique intuitive counterpart to (F6) at the deeper level of the formalization. I will even consider an argument of Ichikawa and Jarvis's which sheds doubt on the very possibility of identifying such a counterpart.

Thus, the purported psychological evidence should not be put thus: we retain some precisely confined intuitive judgement regardless of deviant realizations. It should be put more cautiously: there is a feeling that the GC1–experiment would not be doomed just because a deviant realization turned out to be actual. Does this evidence suffice to refute the counterfactual account? I doubt it. There are several alternative explanations of the felt resilience. One explanation would be that we focus on the dialectics against the JTB–analysis. We feel that the overall case against the JTB–analysis is not damaged by deviant realizations. It would be a desperate move on the part of the traditional epistemologist to point to deviant realizations. Another explanation is that we find the task of providing an amendment trivial. Any particular deviant realization can be stipulated away when it turns out to be closest. We do not bother too much about actually providing an amendment. The story obviously instantiates a pattern that will sooner or later prove successful. The latter point can be supported by Malmgren's own view that the concrete case described comes with a general grip on relevantly similar cases. This could explain one's confidence that, if one particular description fails, there will be a fallback position.

In sum, there is no compelling psychological evidence against the counterfactual account.

Epistemological concerns

I come to the epistemological concern voiced by Ichikawa: '…how do we know that the counterfactual is true? I believe that Williamson's account renders it much too difficult to know the Gettier intuition.'[200] In Williamson's analysis, (F3*) must be known in order to know (F4) (NKJTB is possible). To Ichikawa, in order for (F3*) to be known, one must rule out the relevant alternative that a deviant realization like GC2 is actual. Since that alternative cannot be ruled out, it is very difficult to know (F4) via (F3*). But we do not find (F4) so difficult to know. So our normal route from GC1 to (F4) must be different.

My reply to Ichikawa follows the thrust of Williamson's overall theory of modal knowledge: modal knowledge is reducible to the well-established capacity of everyday counterfactual reasoning. Here is my core claim: according to our normal standards of knowing a counterfactual, (F3*) can be known, though one's justification is defeasible by a deviant realization. Although we know (F3*) only if GC1 is not realized in a deviant way, normally we do not have to *rule out* that GC1 is deviantly realized in order to know (F3*). I remain neutral as to whether (F3*) can be known once deviant realizations have been *raised to salience*. Put yourself into a state of innocence before deviant realizations have been raised to salience. You

[199] Herman Cappelen, *Philosophy without Intuitions* (Oxford: Oxford University Press, 2012), p. 55.
[200] Ichikawa, 'Knowing the Truth', p. 440.

do not pay special attention to them. (F3*) seems intuitively true, just as the Gettier verdict does. So nothing prevents you from using (F3*) to obtain the Gettier verdict.

But why do we normally credit ourselves with knowing (F3*)? I shall consider two explanations without deciding between them. According to the first hypothesis, we take it that the actual world or the GC1–worlds closest to it do not realize GC1 in a deviant way. It is not that we explicitly believe or form some background hypothesis that there are no deviant realizations of GC1. In a similar vein, we do not explicitly believe that the zebra we admire in the zoo is not a cleverly disguised mule. One may compare our neglecting deviant realizations to the role of folk physics as Williamson describes it: folk physics is strictly false. So it is problematic to use it as a *premise* in reasoning about the actual world and counterfactual suppositions. Still it may form part of a reliable *method* of forming beliefs in certain areas. It 'may be stored in the form of some analogue mechanism, perhaps embodied in a connectionist network, which the subject cannot articulate in propositional form.'[201] Analogously, neglecting deviant realizations may just be part of some defeasible default heuristics of evaluating counterfactuals. If there is such a default heuristics, it is not confined to philosophical thought experiments. It also applies to dialogues like the following:

Dialogue14
John: 'Is *atropa belladonna* very poisonous?'
Mary: 'Indeed. (F7) If someone were to eat 10 berries of *atropa belladonna*, she would die.'

We accept (F7) although there might be someone who has just consumed 10 berries of belladonna and the antidote. So (F3*) is in no way special. It stands and falls with our heuristics of evaluating many everyday counterfactuals. The default standards of justification are determined with respect to that heuristics.

There is an alternative explanation why we judge (F3*) true. Perhaps (F3*) is read as something like a *habitual*. It closely corresponds to

(F8) People have justified true belief without knowing that it is 8.28 when they are in GC1.

For comparison, take

(F9) Glasses shatter when dropped.

The habitual (F9) can be true although many glasses actually have landed on soft carpets without shattering. I observe a certain tendency to accept the analogous counterfactual:

Dialogue15
The Savage: 'are glasses damageable?'
Mary: (F10) 'If a glass were dropped, it would shatter.'

(F10) seems perfectly all right as uttered by Mary. Both (F9) and (F10) are sensitive to deviant realizations:

Dialogue16
John: 'Some glasses are packed in cotton wool.'
Mary: (F9) '#But glasses shatter when dropped.'/(F10) #'But if a glass were dropped, it would shatter.'

[201] Williamson, *The Philosophy*, p. 146.

Normally, deviant circumstances are not salient. Then we accept (F9) and (F10) although it would be foolish to deny that some glasses have actually been dropped without shattering. It is a matter of further debate how to analyse counterfactuals like (F10).[202] In an moment, I will sketch one proposal. In any case, if (F3*) is to be treated like (F10), and if I am right that (F10) is acceptable by normal standards, again Ichikawa's criticism can be met. (F3*) comes in good company. It meets the normal standards of justifying counterfactuals. To be sure, reading (F3*) as something like a habitual conflicts with centering assumptions (a counterfactual P >> Q is false if P but not Q). Some philosophers find these assumptions non-negotiable.[203] To them, I offer my first alternative explanation.

I shall give at least a sketch how one might spell out the analogy between habituals and counterfactuals. As we have seen in chapter (2.3.2.), there is linguistic evidence that, just as the definite article 'the' is a device of unambiguously referring to individuals, the subordinator 'if' is a device of referring to possibilities. Drawing on this evidence, the (simplified) standard analysis of counterfactuals can be interpreted in terms of definite descriptions:

A counterfactual P >> Q is true iff the closest worlds such that P are worlds such that Q.

Definite descriptions allow for a *non-maximal* reading. Imagine a teacher standing outside the classroom. Noise is coming from inside. The teacher may truly say

(F11) The children are quarrelling

although, strictly speaking, not all of the children are quarrelling. A discourse normally is a cooperative enterprise of exchanging the information most relevant to decision–making. There is a maximal reading of (F11) where all children in a contextually restricted domain (children in the classroom) are quarrelling and a non-maximal reading where some of the children are quarrelling. The latter is selected because it is more decision-relevant: if some children are quarrelling, the teacher should look after them. It makes only a minor difference whether all or some of the children are quarrelling. The *issue* is determined with respect to the decision problem: are any of the children quarrelling? It is answered by (F11), read as stating that some children are quarrelling.

Given the close connection to descriptions, it is tempting to think that counterfactuals also allow for a non-maximal reading. Assume the teachers at a conference discuss their responsibilities. Responsibilities are determined with respect to classes assigned to teachers. One of the teachers may say:

You can't leave the children alone.
(D42) If the children were left alone, they would quarrel.

This may be perfectly assertable although it is based only on the observation that children in class often quarrel when they are left alone, and so there is no special reason to think that they will quarrel in *all* the closest scenarios where they are left alone. It may even be true if all the children in some class actually happen to be left alone but do not quarrel. This can be explained by reading the subjunctor 'If…' in (D42) non-maximally. Given a teacher's responsibility for her classes, the issue that is relevant to the counterfactual is whether *some* (or sufficiently many) particularly close situations where children in class are left alone are situations where they

[202] Lars Bo Gundersen, 'Outline of a New Semantics for Counterfactuals', *Pacific Philosophical Quarterly*, 85 (2004), 1–20; Michael Smith 'Ceteris Paribus Conditionals and Comparative Normalcy', *Journal of Philosophical Logic*, 36 (2007), 97–121.
[203] Cf. Michael Fara, 'Dispositions and Habituals', *Noûs*, 61 (2005), 43–82; Walters, 'Morgenbesser's Coin'.

quarrel. The issue is also relevant to how fine-grained the closeness–ordering is. It seems natural that we do not merely consider the classes that were actually left alone. Assume only the quiet classes were actually left alone, but there are many aggressive classes around. These classes clearly are relevant to the teachers' decision. And they should be relevant to evaluating (D42).

In the same vein, in Dialogue15, what matters is whether glasses should *in general* be handled such as to prevent damage. (F10)(If a glass were dropped, it would shatter) raises a paradigm case of damage due to insufficient care: a glass shatters because it is dropped. (F10) answers the issue whether glasses are damaged in sufficiently many of these paradigm situations. The aim is some general maxim. I surmise an analysis along these lines where the issue is sufficiently general underlies our use of habituals.

In the Gettier experiment, the overarching issue is whether (F1) (necessarily, one knows iff one has JTB). When it comes to evaluating (F3*), the Gettier counterfactual, it is already established that there are GC1–situations. We have to choose between a maximal and a non–maximal reading of (F3*): according to the maximal reading, all closest GC1–scenarios are NKJTB–scenarios, according to the non-maximal reading, some of them are. We realize that in the context created by the overarching issue, what matters is whether some closest GC1–scenarios are NKJTB–scenarios. This is the subordinate issue answered by (F3*). Accordingly, (F3*) is read as the claim that some of the closest GC1–situations are NKJTB–situations. It simply does not matter whether *all* (as contrasted to some) closest GC1–situations are NKJTB–situations. What counts as 'closest', too, is determined with respect to the issue. Even if deviant realizations are actual, normal GC1–scenarios loom so large among the closest antecedent scenarios as to make (F3*) true. This reading allows us to explain our impatience with deviant realizations as noted by Malmgren. We feel that bringing them up does not contribute to answering the issue.

I have argued that (F3*) is justified by our normal standards, and I have offered possible explanations of our justificatory practice. Now Ichikawa may present his epistemological worry as a sceptical challenge to this very practice. In that case, the sceptical challenge should be extended to all counterfactuals which are vulnerable to the epistemic possibility of deviant realizations, among them everyday counterfactuals like (F7) or (F10). Once the general dimension of the problem of deviant realizations is acknowledged, we have just one variant of sceptical alternatives raised to salience among others. There is no special reason why, of all, thought experiments should be exempted from such a general scepticism. It is no convincing objection to the counterfactual account that thought experiments would get enmeshed into a general scepticism about counterfactuals.

I shall now critically review a proposal which comes quite close to the one I have just made. Alexander Geddes opts for a revival of Williamson's counterfactual.[204] His variant is:

(F3N) If someone were to stand to a proposition P as in GC1, then, normally, she would have a justified true belief that P but know that P.

How are we to assess (F3N)? According to Geddes, we have a 'sense of normalcy', which tells us how things normally go. This sense of normalcy tells us that, in a normal counterfactual situation in which Gettier Case is true, the subject of the case has NKJTB. For (F3N) to be true, our sense of normalcy had better be reliable in telling us the truth about counterfactual scenarios like the one considered in (F3N). In my critical assessment, I shall concentrate on the role of this reliable sense of normalcy in Gettier reasoning.

No normalcy clause

[204] Alexander Geddes, 'Judgements about Thought Experiments', forthcoming in *Mind*.

I do not think that the role of normalcy in Gettier reasoning is adequately captured by (F3N). In this section, I shall outline my doubts: an explicit normalcy clause misses our actual route.

The common aim of all participants in the debate is to render our, i.e. Gettier's and his followers' 'actual route'.[205] The actual route should be psychologically credible. It is a matter of further debate whether it amounts to a sound logical argument. The other participants in the debate seem to pursue the same aim.

What is the actual route? Some guidance can be obtained from Gettier's original presentation. He announced a counterexample to the view that JTB is necessary and sufficient for knowledge. He introduced his cases by 'suppose' and 'imagine' + indicative. This does not obviously square with the formalizations in terms of metaphysical modalities considered so far.[206] Still Gettier's lack of explicit modal fine-tuning might be explained by the immature state of modal metaphysics and epistemology at the time. He might have aimed at metaphysical necessity and possibility without making this explicit.

Now Gettier did not hedge his claim by 'normally'. He simply said:

'I shall now present two cases in which the conditions [justification, truth, belief] are true for some proposition, though it is at the same time false that the person in question knows that proposition.'[207]

Here we cannot simply blame the state of the art in modal theory. Gettier could have been aware of deviant completions of his description, and he could have hedged his claim by 'normally'. To be sure, he might indeed have had in mind a hedged claim. Perhaps he feared to distract his readers by adding 'normally'. Still the most straightforward explanation of why he did not explicitly hedge his claim is that he simply did not reckon with deviant realizations. Nor did his successors until Williamson came.

Geddes has surprisingly little to say about the sense of normalcy given its pivotal role. Here is a more informative passage:

'Now, we typically think of how things generally go in terms of the laws, norms and tendencies that we take to be in force. Keeping with this way of thinking, then, we can say that what is in fact normal for a scenario will be any feature whose absence from an instance of that scenario (in one of the worlds just mentioned) involves an exception to some relevant law, norm or tendency (that is in force across those worlds).'(Geddes forthcoming)

For (F3N) to be true, there has to be some law, norm or tendency ruling out deviant realizations as exceptional. The notion of a tendency would need a lot of elaboration.[208] Hence I focus on laws or norms.

Assume there are laws or norms of epistemic appraisal which also bear on Gettier cases. Such laws or norms guide routine appraisals. They also inform our sensitivity to exceptions. Now Gettier cases do not form a natural epistemic category as characterized by its systematic connection to laws and norms of epistemic appraisal. Moreover, they *are* exceptions. When

[205] Malmgren, p. 283.

[206] Williamson, *The Philosophy*, p. 183, mentions Gettier's use of 'suppose' + indicative, but then he says: 'What is sometimes called 'the Gettier intuition' has been expressed by a counterfactual conditional in English…'(p. 195)

[207] Edmund Gettier, 'Is Justified True Belief Knowledge?', *Analysis*, 23 (1963), 121–123 (pp. 121–122).

[208] Candidates for explicating the notion do not seem promising: Millian clauses for causal laws (John Stuard Mill Mill, *A System of Logic Ratiocinative and Inductive* (London: Parker, 1843), p. 445) do not seem to apply. Dispositions (cf. Gilbert Ryle, *The Concept of Mind* (New York: Barnes and Noble, 1949), pp. 117–118, 131–133) enmesh us in unexpected metaphysical debates. A paraphrase like 'most A's are B's'(T. Stephen Champlin, 'Tendencies', *Proceedings of the Aristotelian Society*, 91 (1990), 119–133 (p. 132)) is problematic: we are not in a position to know whether most Gettier cases are (or would be) cases of NKJTB. The same goes for probabilistic judgements. A related proposal is to account for normalcy by high probability properties (Williams, 'Chances'). But there is no reason to think that normal realizations of Gettier Case can be discerned in this way.

things go their normal way, they do not arise. A routine case of JTB is by a strange twist turned into a situation in which our belief might easily have failed to be true.

We cannot have fully specific rules determining any single case. When we look for generalizations bearing on Gettier Case, candidates like the following spring to mind:

(Rule) *if someone has JTB that P, she knows that P, except if she hits the truth only by accident*.[209]

(Rule) has the right level of specification. It tells us what normally happens and specifies a range of exceptions. It supports counterfactuals like: if someone were to have JTB, she would *normally* have knowledge. A generalization of this sort, if any, is most likely to guide our judgement on Gettier Case. In light of (Rule), GC1 seems exceptional. Implicit reliance on generalizations like (Rule) explains why the JTB–theory of knowledge is tempting, and why we find Gettier cases surprising. (Rule) does not tell us anything about exceptions *among* Gettier cases. Given that Gettier cases are unusual, there is no reason to assume that we use more specific generalizations which take Gettier cases as the norm relative to which exceptions can be discerned. The only motive for this assumption could be that we dismiss realizations like GC2 as deviant. But as we shall see, there is an alternative explanation for our attitude towards them.

Deviant realizations ignored

Notwithstanding my reservations about an explicit normalcy clause tracking objective normalcy, I agree that a subjective 'sense' of normalcy plays a role in Gettier reasoning. However, I suggest not to construe it as a reliable capacity to track some objective feature of reality, but as a blindfold streamlining our fast and frugal practices of reasoning. We do not attend to abnormal realizations and hedge our verdict accordingly. Our 'sense of normalcy' simply makes us ignore them. One looks at the case and immediately classifies it as NKJTB, perhaps guided by a generalization as illustrated in the last section.

To summarize my discussion of extant criticisms, so far there is no reason to give up the counterfactual account. To better appreciate the costs of giving it up, I shall critically examine further alternatives to the counterfactual approach in the literature. My result will be that they are more problematic than the original account they are to mend.

Extant alternatives to the counterfactual approach
A possibility claim

If both the possibility and the actuality of deviant completions spell trouble for a formalization of Gettier reasoning, we might settle for a weaker claim which is not vulnerable to deviant realizations. Instead of claiming that *all* possible GC1–situations or some particularly relevant respectively *close* GC1–situations are NKJTB–situations where the subject has justified belief but does not know, it seems sufficient to claim that there *could* be a GC1–situation where the subject has NKJTB. The resulting candidate as proposed by Malmgren is

(F5) It is possible that someone stands to P as in the Gettier case (as described [by GC1]) and that she has a justified true belief that P but does not know that P.

[209] Or: …*might easily have been wrong, is wrong in some nearby possible worlds etc.*.
I use (Rule) only as an example without incurring a commitment to it.

From (F5), one can directly proceed to the denial of (F1), the JTB–analysis. (F5) seems perfectly proportionate to the task: it yields the intended conclusion while deviant scenarios do not bear on its truth.

I have already critically discussed (F5) as part of my discussion of Malmgren's criticism of Williamson, but I have not yet considered it as an alternative in its own right. shall begin my assessment of (F5) with criticism in literature. Then I shall discuss two criticisms of my own: firstly, (F5) is does not track our normal route but rather some fallback position. Secondly, since the Gettier description is conjoined with the NKJTB–result from the outset, (F5) is unpersuasive.

Malmgren's aim is to precisely identify *the intuitive judgement* that leads us from GC1 to the conclusion that the JTB–analysis is false. It seems that acceptance of the formalized argument should be the shibboleth which allows to tell apart those who accept 'the Gettier intuition' and those who don't. One should be considered as accepting the Gettier intuition precisely if one accepts (F5). Drawing on this requirement, Ichikawa and Jarvis provide the following counterargument:

> 'Malmgren's version of the content of the Gettier intuition[...] cannot make sense of the conflict between someone who accepts the Gettier intuition and someone who denies it. Indeed, someone who thinks that standard Gettier cases are knowledge, but who believes in NKJTB on other grounds will accept [5], but reject the content of the Gettier intuition. […] Suppose someone thinks that stopped–clock cases are knowledge, but thinks that, say, fake–barn cases or lottery cases are NKJTB. Then he will reject the content of the Gettier intuition elicited by the story […]; nevertheless, he should accept [5].'[210]

The argument needs some interpretation. (F5) concerns a particular Gettier story. I have plugged in GC1 (the stopped clock case). It is not immediately transparent why the eccentric's belief that fake–barn cases or lottery cases are NKJTB should make him accept (F5). But assume we can realize GC1 as a 'fake–barn' case: there are many fake clocks around. Then someone who believes that, normally stopped–clock cases are knowledge (say because the clock has been reliable sufficiently often before stopping), but believes that fake–barn cases are NKJTB, can accept (F5) because GC1 can be realized as a fake–clock case.

This argument works against Malmgren. She says that (F5) just is *the* (*GC1–related*) Gettier intuition. The eccentric imagined by Ichikawa and Jarvis would deny that a person who non-deviantly satisfies GC1 has NKJTB. This is incompatible with accepting the Gettier intuition if anything is. But the eccentric would accept (F5). So (F5) cannot be the Gettier intuition.

However, I do not think that (F5) stands defeated. (F5) can be put to a more modest use if one does not subscribe to Malmgren's strong commitments. I have voiced doubts that the task of the formalization is to identify a unique intuitive judgement as the Gettier intuition. Once one drops the identification of (F5) with *the* Gettier intuition, Ichikawa and Jarvis's counterargument is of limited avail against (F5): it is too demanding a requirement that an adequate formalization of the argument cannot be hijacked. Presumably any formalization can be accepted by someone who rejects the Gettier intuition but has arbitrarily weird other beliefs. To see this, consider the alternative lines of formalization, beginning with the counterfactual account. Assume Ichikawa and Jarvis's eccentric additionally believes that all *actual* (or closest) GC1–cases are fake–clock cases. Then he will accept the reasoning *(F1), (F3*); thus (F4)*. Now consider a necessity claim (F3). The argument *(F1), (F3); thus (F4)* could be acceptable to Ichikawa and Jarvis's eccentric provided he additionally has the weird belief that *any* realization of GC1 necessarily is a fake–clock case. This belief is very weird indeed, but so are the beliefs imagined by Ichikawa and Jarvis. So there are doubts that the argument

[210] Jonathan Ichikawa and Benjamin Jarvis, *The Rules of Thought* (Oxford: Oxford University Press, 2013) p. 203 and ann..

disqualifies (F5). If it hits (F5), it hits any other formalization, too. I note that this observation sheds further doubts on the project of identifying *the intuitive premise* in the formalization. Any premise hitherto considered can be accepted by someone whom we would not credit with accepting the Gettier intuition.

There is a further compelling criticism of Malmgren's proposal. Consider a misjudgement on the case: the subject neither has justification nor knowledge. This seems false. The most straightforward way of applying Malmgren's proposal is to also put the misjudgement as a possibility claim. But the possibility claim is true! This is evidence that there is a difference between Malmgren's possibility claim and the intuitive judgement.

Coming to my own critical assessment, (F5) arguably does not represent the normal route but a fallback position. It is designed such as to evade the problem of deviant realizations. But that problem is simply disregarded in a normal Gettier reasoning. This impression is confirmed by my epistemology class dialogues. Consider again

Dialogue11
John: 'Imagine someone who looks at 8.28 at a clock which broke exactly 24h earlier. (F6) She has justified true belief but no knowledge that it is 8.28.'
Mary: 'Not necessarily: the clock at the wall has actually stopped 24h earlier. Betty is looking at the clock to see what time it is, but I have just told her that the clock has stopped 24h earlier.'
John: (F5) 'But it is possible that someone in this scenario has justified true belief but does not know what time it is.'

Dialogue11 perfectly illustrates the fallback role of (F5). John's way of putting GC1 at the beginning seems perfectly in order. But it cannot be reaffirmed after deviant realizations have been raised to salience. Then we need something like (F5). This is evidence that we do not read John's *initial* utterance as (F5). If the task were to provide a deviance–proof scenario from the outset, one would expect that either John's way of putting the experiment is a non–starter, or that it can simply be repeated after Mary's interpellation because what John really intends is something like (F5). I note that the alternative rival accounts to the counterfactual-based one face similar difficulties in explaining the infelicity of repeating (F6) in my Dialogue12 (…John: (F6) #'But someone in my scenario has justified true belief without knowing what time it is.'). Just like Malmgren, they provide a deviance-proof candidate for the Gettier intuition. But if John had this candidate in mind, one would expect his repeating (F6) to be perfectly felicitous. We should take (F6) to stand for the deviance–proof intuition. At least John's reaction in Dialogue12 should be perfectly fine (…John: #'Oh, come on. That's not how I intended the story.').

I come to my second criticism: (F5) integrates GC1 and the classification as NKJTB into one big possibility step. To see how problematic this integration is, consider:

GC3
At 8:28, somebody looks at a clock to see what time it is. The clock is broken; it stopped exactly twenty–four hours previously. The subject believes, on the basis of the clock's reading, that it is 8:28. *And the Subject has justified true belief but does not know that it is 8:28*.

GC3 seems inappropriate as a Gettier description. But if the whole thought experiment boils down to evaluating (F5), GC3 should be a perfectly fitting case description. There does not seem to be a relevant difference between (F5) and

(F12) Possibly, GC3.

Why then does GC3 seem inappropriate? Why do we feel that it begs the question? I propose the following explanation: the persuasiveness of thought experiments is very sensitive to the way they are presented.[211] For a counterexample to be convincing, it should first be accepted as a test case. The description should not provoke resistance from the outset. One should be wary of writing anything into a Gettier description that looks like prejudging whether the case is subsumed under *justified true belief* and especially *knowledge*. GC3 clearly fails in this respect.

But (F5) seems problematic, too. It combines the seemingly neutral case description and the general classificatory task in one possibility claim. You are supposed to accept the description as a case of NKJTB from the outset. In processing 'it is possible that … but does not know that *p*', you realize from the outset that 'but' imposes a NKJTB constraint on the description. This is likely to provoke resistance. The psychologically convincing alternative is to separate the description and the classificatory task. First, there is the question whether to accept the test case as a suitable target of epistemic appraisal. It is a suitable target only if it is possible. So the acceptance step is associated with the possibility claim (F1). But its role goes beyond the possibility claim. When accepting the possibility claim, one should also accept the test case as uncontroversial. The description should be carefully designed to be readily acceptable. Once the case has been accepted, you proceed to the classificatory task. Does the test case fall under justified true belief? And does it fall under knowledge? The possibility claim (F1) together with the conditional claim (F3) or (F3*) perfectly instantiates this pattern.

In sum, (F5) is immune to deviant realizations. But this advantage comes dear. The persuasiveness of the original Gettier experiment is endangered.

Restricting the necessity claim

Instead of *replacing* (F3) by a counterfactual (F3*) or a possibility claim, one may prefer to restrict the necessity claim. An explicit non–deviance or a *ceteris paribus* clause seem too uninformative.[212] I consider two more informative ways of mending (F3).

The thought experiment as a fiction

Ichikawa and Jarvis propose to treat the Gettier description as an *everyday fictional story*. The concrete case descriptions used in normal thought experiments resemble fictional narratives. We all credit ourselves with the capacity of evaluating fictions. And we seem to eschew deviant ways of completing fictional stories. Consider GC1 as a minimalistic short story. There are more things true in the story than we are explicitly told (surely the subject in the story breathes air). But no one would deem GC2 true in the story. So one may use the fiction to determine the domain of the strict conditional as follows: consider the proposition *Q* which is true iff all that is true according to the GC1–fiction is true *tout court.* Replace (F2) and (F3) by

(F2´) Possibly, Q.
(F3´) Necessarily, if Q, someone has justified true belief but does not know some proposition P.

(F2´), (F3´); thus (F4) refutes the JTB–analysis.

[211] 'by presenting content in a suitably concrete or abstract way, thought experiments recruit representational schemas that were otherwise inactive, thereby evoking responses that may run counter to those evoked by alternative presentations of relevantly similar content. …' (Tamar Gendler, 'Philosophical Thought Experiment, Intuitions, and Cognitive Equilibrium', *Midwest Studies in Philosophy*, 31 (2007), 68–89 (p. 69)).
[212] cf. Malmgren, pp. 287–288.

In my critical discussion of (what I call) *the fictional account*, I shall stress the general point that fiction does not underlie the same constraints of conceptual coherence and logical consistency as a thought experiment. Then I shall present a concrete case where this leads to a misconstrual of a philosophically interesting thought experiment.[213]

The fictional account presupposes that there is a precise correspondence between the set of fictional truths and the set of possible situations relevant to the thought experiment. This presupposition is not supported by the extant accounts of truth in fiction mentioned by Ichikawa and Jarvis.[214] In Lewis's counterfactual pattern of analysis, roughly P is true in a fiction iff it is true in the closest world where the story is told as known fact (variant: …and the common beliefs of the community of origin are true).[215] This leads us back to the counterfactual account. If the latter has problems with deviant realizations, a fictional account drawing on Lewis's analysis would have problems, too. In Kendall Walton's pluralistic account, the explicit story functions as a *prop*. It invites us to engage in a game of *make believe*, guided by several conventional principles of generation. The constraints imposed on these principles of generation are too weak to ensure that the content of this *make believe* game can be mapped to propositional truths.[216] So Ichikawa and Jarvis cannot use Lewis's or Walton's account but must come up with a picture of their own.

How do Ichikawa and Jarvis ensure that the set of fictional truths corresponds to the right set of possibilities? A key (but not sufficient) requirement is the following: for any coherent thought experimental scenario, the set of truths which correspond to what is true in the description of the scenario, treated as an everyday fiction, must be perfectly coherent. I interpret coherence in a broad sense as *logical consistency and conceptual coherence*. In other words, truth in fiction is subject to the same constraints of preserving conceptual and logical truths as modal reasoning.[217] If we accept the idea of conceptual coherence, this mimimum requirement seems crucial to the success of the fictional account. For assume there is a thought experiment where it is not met. Some description is perfectly coherent when treated as a thought experiment but not as a fictional story. Then we would clearly be bound to settle for the coherent scenario in the philosophical argument. But Ichikawa and Jarvis's procedure would lead us to diagnose an incoherence: the thought experiment fails because there is no coherent proposition according to which what is true in the fiction is true *tout court*. This diagnosis would obviously be misplaced.

Disregarding concerns about its sufficiency, I shall argue that even the minimum requirement of conceptual coherence is not always met. Fictional truth is not strictly bound by conceptual and logical coherence. As Gregory Curry notes, one may write a fiction where it is explicitly told that someone has refuted Gödel's theorem.[218] So far there is no problem for the fictional account. The fiction is inconsistent, and so would be a corresponding thought experiment. The problem arises when the story is *implicitly* incoherent. One may write a fiction where it is not explicit (or entailed) but *only implicitly* true that someone has refuted Gödel's theorem. Assume the author of the story wants to make vivid what a superb genius her

---

[213] There is some debate on the fictional account (cf. Malmgren, pp. 303–306, Ichikawa and Jarvis, *Rules*, p. 210, Williamson, *Replies*, pp. 467–468). But I focus on a hitherto undiscussed and particularly relevant point. There are many other problems: Fictions are incomplete, possible worlds are not. Fictions come with their own explanatory patterns, which are not acceptable in philosophical arguments (cf. David Velleman, 'Narrative Explanation', *The Philosophical Review*, 112 (2003), 1–25 (p. 21)). It is not a matter of course that we have an epistemically firm grip on the implicit fictional truths.

[214] Ichikawa and Jarvis, *Rules*, 265.

[215] David Lewis, 'Truth in Fiction', in *Philosophical Papers I* (Oxford: Oxford University Press, 1983), pp. 261–280.

[216] Kendall Walton, *Mimesis as Make–Believe. On the Foundations of the Representational Arts* (Cambridge/Mass.: Harvard University Press, 1990), p. 42).

[217] Cf. Jonathan Ichikawa and Benjamin Jarvis, 'Thought–Experiment Intuitions and Truth in Fiction', *Philosophical Studies*, 142 (2009), 221–246 (pp. 234, 237)

[218] Gregory Currie, *The Nature of Fiction* (Cambridge: Cambridge University Press, 1990), p. 69.

protagonist Schmidt is. She elaborates the ceremony where Schmidt is awarded the Fields medal for having refuted Gödel's theorem.[219] If the story is suitably told, no hoax, irony, or indication of error, we are quite ready to accept it as true according to the story that Schmidt *has* refuted Gödel's theorem. Narrative plausibility trumps conceptual coherence and logical consistency. But the same would not follow in a literally identical thought experimental scenario (a philosophical thought experiment where metaphysical possibility and *a fortiori* conceptual coherence are to be preserved). Such a scenario underlies constraints of conceptual and logical coherence. One way of securing these constraints in the Smith story would be to interpret the story such that the laureate and the committee *made a mistake*. We are bound to an interpretation like this in considering the story as a thought experiment but not as a fiction.[220] In sum, the purported correspondence between fiction and thought experiments fails. What is perfectly coherent as a thought experiment may become incoherent as an everyday fiction.[221]

Although my argument is sufficient to shed doubt on the fictional account, there is a fallback position: one may claim that the requirement of conceptual coherence is fulfilled for any interesting philosophical thought experiment which is not obviously absurd. For any such thought experiment, the corresponding everyday fiction is coherent. This claim can only be tested by going through a concrete counterexample. My example will be *fission cases*. I shall elicit how the fictional account misconstrues the dialectics of such cases.

The outline of a fission case is the following: imagine a person P0, whose brain is divided into two halves and implanted into two bodies; two normally functioning persons emerge, P1 and P2, who are psychologically continuous with P0. Consider the following story template:

GC4
Someone undergoes fission. Both of the post–fission persons sincerely utter: 'I remember the slightest details about *my* pre–fission life. I remember my early childhood, my grandparents when they were still alive….'

In that story, drawing on psychological continuity, both protagonists P1 and P2 seem to refer to P0 as *'I'*. Martine Nida–Rümelin argues that it is not possible for P1 and P2 to be *both* identical to P0. Only one of them can be P0. So there must be something over and above psychological continuity that constitutes personal identity.[222] Assume Nida–Rümelin is right. Then in the course of a philosophical argument, we should read GC4 as saying that both P1 and P2 *take* themselves to be identical to P0, but not that they *are* identical to P0. However, when we read GC3 as a science fiction story, it does not seem illegitimate to read the story as one where both people *are* P0, provided their claims to identity are made vivid by their personal story. We do not have to go to the philosophy department to check whether that reading is in tune with our notion of personal identity. Still the story does not explicitly say that P1 and P2 are identical to P0. Given these assumptions, we get a counterexample to the correspondence

[219] It is important that 'for…' is read intensionally: the prize committee's opinion that Schmidt has refuted Gödel's theorem is their reason why they award him the medal.
[220] I have been reminded that for instance in an ethical trolley experiment where the issue is whether to save the man who refuted Gödel's Theorem or three other people, we might prefer to accept that Schmidt has refuted Gödel's Theorem. Although this point sheds further light on our tendency to get the lesson right in spite of difficulties with the surface story (and further doubt on Ichikawa and Jarvis's coherence thesis), I cannot pursue it here.
[221] I have encountered doubts about the idea of a 'literally identical' description, read as a thought experiment and as a fictional story. But the intuitive idea is plausible: GC1 can of course be read as part of an epistemological argument. And our conventions of telling fictional stories seem flexible enough to embed GC1 into a speech–act of story–telling (which may be represented by adding an 'according to the fiction'–operator). Anyway doubts about the two readings would rather threaten Ichikawa and Jarvis's account than my counterargument.
[222] Martine Nida–Rümelin, 'The Argument for Subject Body Dualism from Transtemporal Identity Defended', *Philosophy and Phenomenological Research*, 86 (2013), 702–714.

thesis. Read within the constraints imposed on modal reasoning, GC4 describes a perfectly coherent and possible scenario. But our reading of GC4 as a science fiction story is not bound to be coherent.

However, shouldn't the author's intention of presenting a coherent story constrain our interpretation of *the fiction* such as to prevent the incoherent reading? No, that intention is irretrievably bound to the aim of presenting a philosophical argument. The context of that argument is blinded out when the story is treated as an *everyday* fiction. It would be arbitrary to preserve the intention of coherence and to blind out the argument context where that intention arises.

I conclude that the fictional account is flawed. Fiction is not well-regulated enough to be used in a general analysis of Gettier-like thought experiments.

A deviance-proof story

Thomas Grundmann and Joachim Horvath have suggested that the story can be easily completed such as to exclude deviant realizations in a principled way. In my discussion, I shall use their Gettier example in order to make sure to get their proposal right:[223]

GC5
Smith believes that Jones owns a Ford, on the basis of seeing Jones drive a Ford to work and remembering that Jones always drove a Ford in the past. From this, Smith infers that someone in his office owns a Ford. In fact, someone in Smith's office does own a Ford but it is not Jones, it is Brown (Jones sold his car and now drives a rented Ford).

GC5 is subject to deviant realizations. For instance, Smith may have strong evidence that he regularly hallucinates people driving a Ford. Something has to be done to make the story deviance-proof. In Grundmann and Horvath's hands, the final story becomes:

GC6
Smith justifiedly believes that Jones owns a Ford, on the basis of seeing Jones drive a Ford to work and remembering that Jones always drove a Ford in the past. From this belief alone, Smith logically infers, at time *t*, to the justified belief that someone in his office owns a Ford, which provides his only justification for that belief at *t*. In fact, someone in Smith's office does own a Ford, so that Smith's latter belief is true – but it is not Jones, it is Brown, and so Smith's initial belief was false. (Jones sold his car and now drives a rented Ford.) Also, if Smith knows at *t* that someone in his office drives a Ford, then he knows this at *t* only in virtue of the facts described.

GC6 then is inserted into the original strict conditional template:

(F1) Necessarily, someone knows some proposition P if and only if she has justified true belief in P.
(F13) Possibly, someone stands to some proposition P in the relation described by GC6.[224]
(F14) Necessarily, if someone stands to some proposition P in the relation described by GC6, she has a justified true belief in P without knowing P.
(F4) It is possible that someone has justified true belief that P without knowing P.

[223] Thomas Grundmann and Joachim Horvath, 'Thought Experiments and Deviant Realizations', *Philosophical Studies* 170 (2014), 525–533.
[224] I.e. someone occupies *Smith's* role.

Grundmann and Horvath claim that an *expert* epistemologist normally reads GC5 as GC6. How does this square with the distinction between our normal route and a fallback strategy which I introduced in section 2? I interpret Grundmann and Horvath as proposing that *(F13), (F14); thus (F4)* represents the *normal route*, the route of a competent pre-Williamsonian thought experimenter who is to sincerely test the JTB–theory, rather than a fallback strategy. After all, Grundmann and Horvath argue that Malmgren's (F5) is psychologically unconvincing because our *actual* reasoning is more complex. This argument would be of no avail if the aim were a fallback strategy.

In my critical assessment, I shall provide two objections to the claim that GC5 is normally read as GC6. Firstly, analogously to GC3, there are doubts that GC6 fits the requirements of a persuasive thought experiment (as contrasted to GC5). Secondly, there is a linguistic gap in explaining why GC5 is read as GC6. So GC6 represents a fallback position rather than the normal route.

I shall start with some friendly amendments. The conditions which are to ensure that Smith has justified true belief are flawed. To begin with the obvious mistakes: Smith cannot logically deduce the belief that someone in his office owns a Ford *just* from his belief that Jones owns a Ford. He at least needs the additional premise that Jones works in the same office as Smith. Moreover, what precisely does provide his *only* justification to believe that someone in his office drives a Ford? Not just his logical inference. Arguably he additionally must be justified in believing the premises of the inference at *t*. Since it is not obvious where his *only justification* should end, instead of *which provides his only justification* perhaps one had better demand that any justification of Smith's proceeds via the logical inference mentioned in the text.

Coming to my first criticism, as witnessed by GC3, it would be disastrous if the requirement that *Smith does not know* were explicitly written into the description. Grundmann and Horvath are careful to avoid it. Instead of *Smith does not know*, they use

(F15) if Smith knows at *t* that someone in his office drives a Ford, then he knows this at *t* only in virtue of the facts described.

The strategy is clear. In contrast to *Smith does not know*, the conditional claim leaves open whether Smith knows. Just in case he knows, there must be no other factors which contribute to his knowledge (as in the deviant story GC2).

Does this strategy succeed? One condition of its success is that (F15) should be as readily acceptable as GC5. GC5 is perfectly down–to–earth and easy to understand. We are fairly confident that such a situation might occur in everyday life. (F15), in contrast, is not only more difficult to understand. Upon closer inspection, it should definitely give us pause. Given GC6 is a good Gettier story, the consequent of (F15) is impossible. For the sake of argument, I grant that 'only in virtue of' can be clarified. What is required is that nothing over and above the facts described turns one's justified true belief into knowledge. If some set of facts described is sufficient to turn one's belief into knowledge, no additional supplementary description consistent with the original description can interfere such as to prevent that one knows. Assuming GC6 is a good Gettier case, the facts described alone *cannot* turn one's belief into knowledge, as witnessed by the epistemic accident which prevents that Smith knows.

The impossibility of the consequent is problematic. Grundmann and Horvath motivate (F15) as follows: the conditional neither states nor logically entails the only fact about knowledge that ultimately matters for the thought experiment, namely, whether Smith knows that someone in his office owns a Ford. Whether (F15) is in accordance with this motivation depends on what 'logically entails' means. Assume P logically entails Q iff, necessarily, $P \supset Q$. Then (F15) *does* entail that Smith does not know, the consequent being impossible. We may adopt some more restrained notion of logical entailment. Still Grundmann

and Horvath's motivation is insufficient to distinguish (F15) from absurd 'refutations' of the JTB–theory which are structurally similar to (F15):

GC7
There is a proposition P and a subject *S* such that *S* has justified true belief that P. And if *S* knows that P, two plus two does not equal four.

Since GC7 describes a perfectly possible situation (provided the JTB–theory is false), we can infer (G9) (Someone could have NKJTB) from it. The JTB–theory stands refuted. And if (F15) 'neither states nor logically entails' that Smith does not know, the same goes for GC7. A convincing formalization must conform to a stronger condition: it must also avoid any suspicion of *indirectly* stipulating that Smith does not know. As witnessed by our rejecting GC7, a conditional with an obviously impossible consequent intuitively counts as such an indirect stipulation.[225]

For (F15) not to count as an indirect stipulation that Smith does not know, the consequent of (F15) *Smith knows only in virtue of the facts described* should not be too obviously impossible. The consequent of (F15) is impossible. But perhaps it is not *too obviously* impossible. There is a difference to the absurd Gettier case GC7. The impossibility of the consequent *the subject knows only in virtue of the facts mentioned* is not yet transparent to the reader who is still about to sincerely test the JTB–theory. Perhaps this epistemic gap is precisely what is needed for the case to look uncontestable.

Consider one exemplary way for the reader to make up her mind about the requirement imposed by (F15). The situation described must be one where *either* Smith does not know *or* knows just in virtue of the facts described. The first disjunct *Smith does not know* already prejudges the classificatory step against the JTB–analysis. If she is to accept the test case, the reader should not feel from the outset that, by virtue of (F15), all GC6–situations are like that. So whether she accepts the case as uncontentious depends on her attitude towards the second disjunct. Since the latter is impossible, she should at least be doubtful about *Smith does know only in virtue of the facts described*. There is an urgent suspicion that (F15) can only be implemented by Smith not knowing. So instead of accepting GC6 as an uncontentious test case, she will feel the suspicion that somehow the claim that Smith does not know has been smuggled into it. At least a considerate reader should be expected to *ruminate* about (F15) in a way we simply do not find ourselves ruminating about GC5. To be sure, just as GC3, GC6 *is* a counterexample to the JTB–theory. But it is doubtful that it is as immediately persuasive a counterexample as the original descriptions GC1 and GC5. This in turn raises doubts as to whether GC5 is read as GC6 in a normal epistemological context where GC5 prompts immediate persuasion.[226]

---

225 There is an analogy to 'Dutchman' conditionals like:

(F16) If there will be a breakthrow on climate protection in the next two years, I am a Dutchman.

Such a conditional serves to express one's high confidence in the antecedent by adding a consequent which is certainly false. GC7 and (F15) play a somewhat similar role: in using a consequent which is necessarily false, they rule out situations where the antecedent obtains.

[226] I have been suggested that (F15) may be replaced by

(F15') Smith's belief is based only on the facts described.

I cannot exclude that something like (F15') will work. But there are difficulties: there may be facts relevant to whether Smith knows which are not trivially captured by what his belief is *based* on. For instance, Smith may count as knowing because of certain background knowledge of his, which he fails to take into account when basing his beliefs on the facts described. Such uncertainties detract from the persuasiveness of the amended story.

Coming to my second criticism, Grundmann and Horvath claim that philosophers normally read GC5 as GC6. This is far from self-evident. GC6 is concocted by reasoning back from the Gettier intuition in light of deviant realizations. Grundmann and Horvath owe a story how GC5 naturally prompts the GC6–reading. One would expect something like a pragmatic implicature. This expectancy is not met. According to Grundmann and Horvath, the GC6–reading is not a matter of normal linguistic competence. It takes an expert. But it remains completely open what guides the expert reading. The rival accounts considered in the last sections do not face this problem. Malmgren's (F5) needs nothing over and above the literal case description, and the fictional account draws on our undeniable tendency to supplement the story by implicit fictional truths. As long as no comparable mechanism is specified, GC6 rather looks like a fallback strategy than like a normal reading of GC5.[227]

In sum, the criticism of the counterfactual account is far from compelling, and the rival accounts face grievous difficulties. So I propose to reconsider Williamson's elegant proposal.

[227] There may be a deviance–discarding mechanism which supplements the explicit description in our normal reasoning. But I doubt that the original deviance–discarding mechanism is captured by Malmgren's possibility claim (F5), which just leaves the story as it is and adds a NKJTB clause, Ichikawa and Jarvis's fictional account or Grundman and Horvath's *ad hoc* 'expert' reading. If there is such a mechanism, it might well be captured by a refined version of the counterfactual account.

2.6. *Will* and *Were*

Typical cases of counterfactuals are of the had–would–form. In such cases, the difference to indicative conditionals is clearly marked, as witnessed by the following *Adams pair*:

(A11) If Shakespeare did not write Hamlet, someone else did.
(A12) If Shakespeare had not written Hamlet, no one else would have.

When it comes to present-and-future-directed conditionals, the distinction is not so clear any longer, as will become obvious in the section to come. In this section, I shall consider whether future-directed indicative conditionals can be read as counterfactuals. In the next section, I shall consider the distinction between 'will' and 'were'.

Can future indicatives be read as counterfactuals?

Adam Morton argues that some future-directed *indicative* conditionals form Adams pairs. Lara is a bomb expert. Live bombs are marked. We cannot see the marks but Lara can. Most bombs are live. So we can say that

(G1) If Lara touches the bomb, it will explode.

But since Lara is very diligent, she won't touch a marked bomb. So we can say that

(G2) if Lara touches the bomb, it won't explode.

We can say this even if we do not at all believe that Lara touches it, i.e. even if we accept 'it *might not be* (is epistemically impossible) that she touches it'.[228]

Does (G1) express a subjunctive? There is a decisive counterargument. Seth Yalcin has pointed to the inadequacy of combining an indicative 'If *p*' and 'It might be that not *p*' (an epistemic possibility claim).[229] The following seems infelicitous:

(G3) #It is raining and it might be that it is not raining

This holds as well for the conditional:

(G4) #If it is raining and it might be that it is not raining, still the grass is wet.

The above sentences either claim or invite to suppose as part of the informational state relevant to evaluating the epistemic modal that it is raining; so they exclude the epistemic possibility (within the supposition) that actually it is not raining.[230]

In contrast, the subjunctive is in order:

(G5) If it were raining and it might be that it is not raining, the grass would be wet.

In the counterfactual scenarios relevant to assessing this subjunctive, it is raining; but it does not have to be part of the informational state relevant to assessing the epistemic modal that

[228] Cf. Adam Morton, 'Indicative versus Subjunctive in Future conditionals', *Analysis*, 64 (2004), 289–93 (pp. 291-292).
[229] Seth Yalcin, 'Epistemic Modals', *Mind*, 116 (2007), 983–1026, p. 985.
[230] For a thorough account Yalcin, 'Epistemic Modals', pp. 998–999.

it is. Analogously, if Morton were right that 'If Lara touches it, it will explode' expresses the subjunctive mood, the following should be in order:

(G6) # If Lara touches it and it might be that she does not touch it, the bomb will explode.

For there should be a natural subjunctive reading of the indicative conditional. And charity has us choose this reading if available. But the conditional sounds infelicitous. So there is no subjunctive reading.

In contrast, the following seems all right:

(G7) If Lara were to touch it and it might be that she does not touch it, the bomb would explode.

This conditional seems to have a counterfactual reading.[231]

But Morton has not established that future–directed *indicative* conditionals can be used to express them.

Future indicatives and were

Having discussed Adams pairs, I shall now turn to future indicative and 'were'–conditionals. I take issue with two claims developed by Keith DeRose. DeRose concentrates on the role of conditionals in deliberation, but he draws general semantic consequences. Conditionals of deliberation must not depend on backtracking grounds. 'Were'ed–up conditionals coincide with future–directed indicative conditionals; the only difference in their meaning is that they must not depend on backtracking grounds. I use Egan's counterexamples to causal decision theory to contest the first and an example of backtracking reasoning by David Lewis to contest the second claim. I tentatively outline a rivalling account of 'were'ed-up conditionals, which combines features of the standard analysis of counterfactuals with the contextual relevance of the corresponding indicative conditionals.

DeRose addresses two questions, which are of crucial importance to a general theory of conditionals:[232] (i) One main function of conditionals is practical deliberation. We deliberate what the consequences are given we perform some action. But what are the conditionals suitable for expressing such deliberations? (ii) There is a fairly standard view according to which indicative and subjunctive conditionals are distinguished by the latter usually expressing counterfactuals, at least when they are of the form: 'If A had been the case, C would have been the case.' But how are we to understand future–directed 'were'ed-up conditionals ('If A were the case (at some future time t), C would be')?

DeRose answers: (i) Practical deliberation usually proceeds by indicative conditionals; yet in order to be deliberationally useful, conditionals must not depend on *backtracking grounds*. (ii) There are no genuine future–directed counterfactuals expressible by 'were'ed-up conditionals. What looks like a counterfactual, in fact roughly shares the semantics of indicative conditionals, although it slightly diverges in assertability conditions.

I want to criticize both answers in light of some counterexamples. I argue for two claims:

[231] So it might give rise to a future Adams pair. However, although he grants this, Keith DeRose notes that the following pair is inconsistent: 'If I put Eve into situation S_1, she will sin; but if I were to put her into situation S_1, she wouldn't sin'(DeRose, 'Conditionals', p. 9) As a consequence, DeRose denies that a future–directed 'If …were, …' can be a counterfactual. In contrast, I think that combinations with 'might' provide evidence for a counterfactual reading. Yet DeRose's example sheds further doubt on there being future Adams pairs. No Adams without Eve.

[232] DeRose, 'Conditionals'.

1. What disqualifies conditionals for deliberational purposes is not backtracking.
2. 'Were'ed-up conditionals are not just souped-up indicative conditionals.

Conditionals of deliberation may depend on backtracking grounds

DeRose points out a problem of his central hypothesis that the conditionals of deliberation are indicatives. As the well-known counterexamples to evidential decision theory show, some indicative conditionals convey links which are merely evidential but not causal. These conditionals may give rise to ineligible courses of action if used in practical deliberation. To evade this problem, DeRose provides a criterion that allows to tell apart conditionals which may be used in deliberation and conditionals which may not. The former should not depend on backtracking grounds (pp. 28–30). An example:

'[...]if Sophie is deciding between going to seminary or joining the army, and knows that (even after she has heard about the connection between her career choice and the likelihood of her having the condition) her choosing to go to seminary would be very strong evidence that she has a certain genetic condition that, if she has it, will almost certainly also result in her dying before the age of 40 years, she has strong grounds to accept that, very probably

[G8] If I go to seminary, I will die before the age of 40

Yet, as most can sense, this, plus her desire not to die young, provides her with no good reason to choose against the seminary, for she already either has the genetic condition in question or she does not, and her choice of career paths will not affect whether she has the condition.'(p. 22)

To DeRose, (P) is deliberationally useless because it is backtracking:

'Sophie's grounds for (P) [...] involve this backtracking pattern of reasoning. After provisionally making the supposition that she goes to seminary, she then reaches backward in the causal order to conditionally alter her view of what her genetic condition is (from agnostic to supposing that she (probably) has the lethal condition), to explain how that antecedent (likely) would become true, and she then conditionally reasons forward to her untimely death.'(p. 29)

However, some recent paradigm cases presented by Andy Egan should give us pause. Egan heralds them as counterexamples to causal decision theory.

'The Psychopath Button
Paul is debating whether to press the 'kill all psychopaths' button. It would, he thinks, be much better to live in a world with no psychopaths. Unfortunately, Paul is quite confident that only a psychopath would press such a button. Paul very strongly prefers living in a world *with* psychopaths to dying. Should Paul press the button? (Set aside your theoretical commitments and put yourself in Paul's situation. Would *you* press the button? Would you take yourself to be irrational for not doing so?)'[233]

By Egan's lights, sound intuition has it that Paul should not press. As it seems, any reasoning that leads to this result irremediably depends on backtracking grounds. If Paul presses, he must have been a psychopath all along in order to do so; hence he will be killed. This reasoning exactly parallels DeRose's example of Sophie, the difference being that only Paul's conditional plays a role in evaluating *the causal consequences* of the choice to be made. There are some reservations about Egan's examples. But I have not yet seen an argument that successfully

[233] Andy Egan, 'Some Counterexamples to Causal Decision Theory', *The Philosophical Review*, 116 (2007), 93–114 (p. 97).

counters their intuitive pull.[234] The lesson is that practical deliberation sometimes has to embark on backtracking considerations that lead from the chosen action to the causal structure that makes one choose it and that also bears on the causal questions of one's action. Hence assuming that 'were'–conditionals guide deliberation does not give us reasons to deny that they sometimes rest on backtracking.

Independently of Egan cases, there is reason to doubt the backtracking diagnosis. DeRose's main evidence is his version of Gibbard's riverboat example (DeRose, Conditionals, pp. 21–25) as already quoted and discussed in section (2.2.):

'Sly Pete and Mr. Stone are playing poker on a Mississippi riverboat. It is now up to Pete to call or fold. My henchman Zack sees Stone's hand, which is quite good, and signals its content to Pete. My henchman Jack sees both hands, and sees that Pete's hand is rather low, so that Stone's is the winning hand. At this point, the room is cleared. A few minutes later, Zack slips me a note which says 'If Pete called, he won,' and Jack slips me a note which says 'If Pete called, he lost.' I know that these notes both come from my trusted henchmen, but do not know which of them sent which note. I conclude that Pete folded.'

Assume Zack at some point accepts a future-directed conditional:

(G9) If Pete calls, he will win

Jack, in contrast, knows that Pete has the losing hand. So he justifiedly accepts

(G10) If Pete calls, he will not win

As DeRose notes, when the conditionals are reported to Pete, (G10) is useful in deliberation but (G9) is not. DeRose's explanation is that (G9) depends on backtracking reasoning: if Pete plays, that will be because he has the higher card; but then of course he will win. (p. 29)

Judging from the Sophie case, we should expect backtracking to go as follows: Zack arrives at (G9) by 'provisionally making the supposition' that Pete calls and then 'reaching backwards in the causal order' such as to revise his beliefs about Pete's playing dispositions. Sophie *must* use the supposition of her going to seminary as evidence for a certain causal order to arrive at (P). That's why her reasoning *does depend* on backtracking. Nothing like that has to occur in Zack's reasoning. Without reaching backwards in the causal order from the supposition that Pete calls, he can derive (G9) from independently justified assumptions about Pete's using method M: arrange for knowing the cards! Call precisely if you are signalled that you have the higher cards! Hence (G9) does *not depend* on backtracking reasoning.[235]

What disqualifies (G9) for Pete's deliberational purposes is that for (G9) to be acceptable in the first place, Pete must use method M. (G9) would be undermined if he were to use (G4) *instead of M* to reason: 'If I call, I will win. So I should call.' In contrast, (Oc) does not exhibit this pattern of dependence. More generally, a conditional is useless in a deliberational process when it is acceptable only provided the deliberational process does not depend on this very conditional.

[234] Doubts about Egan cases are expressed by Frank Arntzenius, 'No regrets, or: Edith Piaf Revamps Decision Theory', *Erkenntnis*, 68 (2008), 277–297; John Cantwell, 'On an Alleged Counter–Example to Causal Decision Theory', *Synthese*, 173 (2010), 127–152.

[235] In spite of these shortcomings, we might reckon the 'that will be because'–template a shibboleth of useless conditionals. Indeed this template might provide some evidence against a conditional being deliberationally useful. It indicates that the conditional draws on causal facts that are 'sunk'. Yet there are counterexamples which are less demanding than Egan cases: 'Should I go to the exhibition? I should go only because I appreciate the artworks for their own sake. If I go, that will be because of my snobbery and not because of my appreciating the artworks for their own sake. So I should not go.' Note that unlike Egan, one does not have to claim that a certain choice is ultimately preferable but only that these considerations play a legitimate role in deliberating action.

'Were'ed-up conditionals are not just souped-up indicative conditionals

By DeRose's lights, 'were'ed-up conditionals coincide with indicative conditionals, the difference being that they can be used to convey that the antecedent is probably false and that they are unassertable when the corresponding indicative conditionals for their assertability depend on backtracking reasoning (pp. 37–38).

My criticism takes three steps. (i) I outline a counterintuition which I take to show that DeRose's solution is wrong. (ii) I summarize the main evidence assembled by DeRose. (iii) I indicate an alternative way of dealing with this evidence, which combines features of the standard analysis of counterfactuals with the contextual relevance of the corresponding indicative conditionals.

Problems of DeRose's reading

Why is DeRose's approach problematic? I think the immediate intuition how to deal with 'were'ed-up conditionals is to assimilate them to other subjunctive conditionals. As a consequence, the following reasoning of David Lewis's seems to apply. Lewis famously eschews backtracking counterfactuals. Yet he allows for certain exceptions triggered by suitable contextual clues which override the standard non-backtracking solution:

'Jim and Jack quarreled yesterday, and Jack is still hopping mad. We conclude that if Jim asked Jack for help today, Jack would not help him. But wait: Jim is a prideful fellow. He never would ask for help after such a quarrel; *if Jim were to ask Jack for help today, there would have to have been no quarrel yesterday*. In that case Jack would be his usual generous self. So [G11] if Jim asked Jack for help today, Jack would help him after all.'[236]

Nothing seems to preclude modifying Lewis'story as follows: Just replace the last sentence by 'So if Jim were to ask Jack for help later today, Jack would help him after all.' This is infelicitous by DeRose's lights.[237] But it seems perfectly in order.

DeRose's evidence

To appreciate DeRose's evidence, consider the problem of future Adams pairs. The classical Adams pair is this:

(A9) If Oswald didn't kill Kennedy, someone else did
(A10) If Oswald hadn't killed Kennedy, someone else would have (DeRose, Conditionals, p. 2)

The following seems perfectly acceptable:

(G12) If Oswald didn't kill Kennedy, someone else did; but if Oswald hadn't killed Kennedy, no one else would have.

In contrast, the following future-directed pair sounds inconsistent:

[236] Lewis, 'Counterfactual Dependence', p. 33.
[237] Curiously DeRose accepts that 'were'ed–up conditionals might be used in this way provided the backtracking reasoning is explicit (p. 35 ann. 31). But I do not see how this concession can be reconciled with his overall account of their meaning and purpose: 'Were'ing–up is a device of clearly marking out conditionals as based on the right sorts of grounds to be deliberationally useful.'(p. 38)

'[G13] If I put Eve into situation S_1, she will sin; but if I were to put her into situation S_1, she wouldn't sin.'(p. 9)

While DeRose is more cautious (p. 10), one may take the Eve–pair to provide further evidence against future Adams pairs than the one I have given above. No Adams without *Eve*. DeRose's approach neatly explains why *Eve* sounds inconsistent: 'Were'ed-up conditionals just coincide with indicative conditionals in the relevant respects. The independent lesson to draw is that (a) *there is an intimate connection between an indicative and the corresponding 'were'ed-up conditional*; *the indicative is not reconcilable with the contrary 'were'ed-up conditional*.

Yet the riverboat example teaches an opposing lesson. While Zack justifiedly accepts

(G9) If Pete calls, he will win,

according to DeRose, he should reject (p. 32)

(G10) If Pete were to call, he would win.

Gibbard prefers a nearness analysis of (G10) as it is standard for subjunctives of the 'had-would' type.[238] In contrast, DeRose maintains:

'Gibbard's response is to place [G9] and [G10] on opposite sides of the great semantic divide among conditionals. [...]though I agree with Gibbard that [G9] seems right and [G10] wrong for Zack, the difference between the two conditionals seems slight and subtle. They seem to mean approximately the same thing, which, together with the sense that one seems right and the other wrong here produces a bit of a sense of puzzlement about the situation.'(p. 33)

DeRose's account is to dissolve this puzzlement. In contrast to (G9), the 'were'ed-up (G10) is unassertable because it rests on backtracking grounds. Yet if DeRose's evidence so far shows anything, then only that (b) *sometimes there is a divide between the indicative conditional and its 'were'ed-up version; one is assertable while the other is not.* I am not sure about DeRose's intuitions about (G10), but I shall accept them for the sake of argument.

This lesson is enforced by a further argument of DeRose's. Indicative conditionals underlie a paradox (pp. 16–17). The following reasoning seems all right:

(G14) Either the butler or the gardener did it.
(G15) Therefore, if the butler didn't do it, the gardener did.

So does the following:

(G16) The butler did it.
(G14) Therefore, either the butler did it or the gardener did it.[239]

But we cannot reason as follows:

(G16) The butler did it.
(G15) Therefore, if the butler didn't do it, the gardener did.

[238] Gibbard, pp. 228-229.
[239] I am not so sure whether this really seems compelling to the untutored. But let us grant the point.

While DeRose has it that future indicative conditionals underlie the paradox, he reports mixed intuitions as to whether 'were'ed-up conditionals do (pp. 36–37 ann.). His explanation is this: We cannot simply reason

(G17) Either the butler or the gardener will do it.
(G18) Therefore, if the butler were not to do it, the gardener would.

For when we accept (G17), still we cannot be sure that the assertability conditions of (G18) are met. Backtracking reasoning might be involved. There are two concerns about this argument: Firstly, assume we check and rule out first that any of our premises depends on backtracking; then the reasoning should seem convincing. If we are still reluctant, this calls for a different explanation. Secondly, why are the results mixed? If DeRose were right, every competent speaker should feel the same about (G17)–(G18).

How to deal with DeRose's evidence

Given my intuitions about Lewis's Jim–and–Jack example, I need a way of reconciling lesson (a) and (b) that diverges from DeRose's:

(a) the close connection between indicative and the corresponding 'were'ed-up conditionals that rules out *Eve*
(b) the difference between indicative and corresponding 'were'ed-up conditionals that accounts for
–(G9) being assertable but (G10) not
–the paradox of indicative conditionals pertaining to future-directed indicative conditionals but not clearly to their 'were'ed-up version.

Ad (a) A leitmotiv of DeRose is that indicative and 'were'ed-up conditionals are too close to each other to be placed on 'opposite sides of the great semantic divide'. Yet this can be accommodated as follows: The factual/counterfactual–distinction is not as well marked with respect to the future as with respect to the past. This distinction is crucial for the sharp boundary between indicative and subjunctive conditionals as it is manifested in past-directed Adams pairs. We take the past to be fixed. Past-directed indicative conditionals are assessed by (hypothetically) taking the antecedent to be *part of the fixed past*. When they give rise to Adams pairs, the antecedent situations of past-directed subjunctive conditionals are taken to be *ruled out by the fixed past*. In contrast, we are prone to regard the future as not yet fixed. There is a tendency towards considering the antecedent of a future-directed conditional as an option that has not yet been ruled out and is not predetermined to come about either. As a consequence, the demarcation of future indicatives and counterfactuals tends to become obliterated. This is the reason why future indicative conditionals and their 'were'ed-up variants are so close to each other; and why the very same antecedent possibility that is envisaged in the indicative partner of an alleged future Adams pair like *Eve* is counted among the antecedent possibilities relevant to evaluating the contrary 'were'ed-up version. In the very same scenario of Eve being put into situation S_1, she would have both to sin and not sin for the conditionals to be reconcilable. Yet by opting for the 'were'ed-up version we express that we feel hesitant about the antecedent situation coming about in due course, i.e. in the way the indicative conditional conveys; as a consequence, we normally open the range of situations relevant to evaluating the 'were'ed-up version for the standard ways we take a counterfactual antecedent situation to come about.

There are different ways of further elaborating these findings. I want to keep my approach as simple as possible. To start with, although one should wary of going 'Into the

swamp' of indicative conditionals (De Rose, pp. 39-40), I need some minimal common ground between indicative and (a standard view of) subjunctive conditionals:

A conditional 'If A, C'/'If A were the case, C would be' is true/assertable iff C in all salient A-situations.

I hope that my use of this vague condition squares with DeRose's Ramseyan account: '…one is positioned to assert [the indicative conditional] A→C if and only if adding A as a certainty to one's belief set would put one in a position to assert that C.'(p. 15) Nothing I say should preclude the situations salient in indicative conditionals from being those that vindicate one's beliefs about the actual world updated with the certainty A.

I combine this with a simplified standard closeness analysis of 'were'ed-up conditionals. Just let the salient A–situations be those that are closest or most similar to the actual situation. When we ask ourselves how the antecedent A might come about, the A–situations envisaged in the future indicative spring to mind. Drawing on the Ramsey test, I surmise that these situations are those that make our belief system updated with A true. We usually reckon them among the closest A–situations with regard to which the subjunctive is assessed.

I suggest the following constraint on 'were'ing–up:

Eve–constraint
Whenever a future-directed indicative conditional 'If A, C' is assertable, an A–cum–C–situation must be among the closest situations relevant to evaluating its 'were'ed-up versions.

More precisely, whenever the indicative conditional is assertable, the contrary 'were'ed-up version 'If A were the case, C would *not* be the case' is not. Yet in evaluating the 'were'ed-up conditional, we also attend to ways in which we take a standard counterfactual situation to come about; hence the A–cum–C–situations envisaged in the indicative conditional normally are *only one among several* candidates for the closest A–situations. As a consequence, the assertability of the indicative conditional normally is not sufficient for the 'were'ed-up version being assertable as well. Pace Gibbard, I think that DeRose is completely right *not* to simply place indicative conditionals and their 'were'ed-up versions on 'opposite sides of the great semantic divide'. Yet I follow Gibbard's nearness analysis. As far as the great divide exists for future-directed conditionals, it cuts through 'were'ed-up conditionals.

Ad (b) There is an eligible explanation of our rejecting (G10): A standard Lewisian analysis is available which parallels the notorious Nixon example:

(A33) If Nixon had pressed the button, nuclear holocaust would have ensued.

Lewis proposal under determinism is this: By default, we take a small miracle to bring about Nixon pressing the button, say an additional neuron firing in his brain. Under indeterminism, some comparable chance process makes Nixon press the button. We do not resort to Nixon's reasons for pressing the button or the like.

This analysis may be transferred to (G10): We take a small inconspicuous divergence to bring about Pete calling. We do not resort to Pete's reasons for calling or the like. So we do not care about Pete knowing the cards of his opponent and reacting rationally. Since we do not posit a connection between Pete's calling and the distribution of cards, we have no reason to assume that Pete will win in all salient situations.

However, taking into account the constraint that rules out *Eve*, we cannot simply settle for Lewis's criteria. (G9) seems less clearly distinguished from (G10) than

(C2) If Pete called, he won

is distinguished from

(G19) If Pete had called, he would have won.

as uttered by Zack from an ex post perspective yet given the same evidence.

This can be explained by the asymmetry between past-directed and future conditionals. Lewis's criteria are *partially overridden* by the *Eve–constraint*: The antecedent–cum–consequent–situations that are salient in (G9) are among the closest situations considered in evaluating (G10). So we tend to amend Lewis's criteria by this constraint. But this is reconcilable with our adhering to them to a certain extent. Among the closest situations considered are situations where Pete's calling comes about by a small miracle or the like. This accounts for our rejecting (G10).

Concerning Lewis's Jack–and–Jim example, the nearness account explains why we accept

(G20) If Jim were to ask Jack for help, Jack would help him after all.

Since we tend to rule out that Jim will ask Jack, we do not feel inclined to the indicative conditionals

(G21) If Jim asks, Jack will / will not help him.

If we deem them unassertable, the *Eve constraint* is not binding. In this case, the 'were'ed-up conditional is treated according to a standard nearness analysis and converges to the counterfactual

(G11) If Jim asked for help today, Jack would help him after all.

But assume we are pressed about the indicative conditional ('yes, but *if* Jim asks?'). If we deem an indicative conditional assertable, the context has us rather accept 'if Jim asks, Jack will help him'. Then the *Eve constraint* supports the 'were'ed-up version.

A general concern: How can situations that are framed so differently, on the one hand in terms of the Ramsey test, on the other hand in terms of Lewis's small inconspicuous divergence count as equally close? Due to the specific openness we accord to the future, we waver between two quite different options for closeness: The first is to take the antecedent A as a new piece of evidence in light of which we revise our view of the actual situation. So we consider the situations that make true our system of beliefs about the actual world updated with A. Yet by the subjunctive mood, we signal that we do not simply take the actual situation as giving rise to A in due course. Hence we also consider the closest situation which is different from the actual course things will take. *That* closest situation is not reckoned a candidate for updating our beliefs about the actual world. It is distinguished from the actual world by a small inconspicuous divergence (miracle) that brings about A.

The remaining task is to account for the mixed results regarding the paradox of indicative conditionals. To begin with, although the reading just developed is the default reading, there may even be a reading of (G10) in which the situations that are salient in (G9) *completely override* Lewis's criteria. One may attend exclusively to the features which guide (our prediction of) Pete's rational deliberations: Pete knows the cards of his opponent, he knows the rules of the game, he aims at winning and so on. I have suggested that these are the relevant features common to situations that make true our belief system updated with 'Pete calls'. Under these circumstances, the closest situations in which he calls will be situations in which he wins.

As a consequence, in one non–default reading 'were'ed-up conditionals *come close to the corresponding indicative conditionals*, perhaps so close as to coincide with them. Yet we do not settle for this reading unless there are sufficient clues enforcing it.

On this basis we may account for the paradox of indicative conditionals. Those who deem the inference (G17)–(G18) invalid, treat the 'were'ed-up conditional according to a default nearness analysis modified by the *Eve constraint*. For instance, they take into account that while the gardener is innocent, the butler is about to do it but some small miracle interferes. Hence they deny that if the butler were not do to it, the gardener would. Those who tend to accept the inference follow the contextual pull of assimilating it to the corresponding indicative conditional. Where may this pull come from? (G14)/(G17) focus attention on the possibilities of the butler and the gardener doing it. In order for the indicative conditional (G15) to follow from (G14), one must *rule out* any further possibilities as salient.[240] The contextually relevant situation in which the butler does not do it is one in which the gardener does.

The result is a neat picture of 'were'ed-up conditionals:

'Were'ed-up conditionals conform to the standard analysis of counterfactuals, the difference being that Lewis's default criteria are either
– partially overridden by the Eve–constraint (the standard case)
or
– completely overridden by the situations that are salient in the corresponding indicative conditionals (given certain contextual clues).

[240] Otherwise one could not assert that the gardener did it upon adding that the butler did not do it as a certainty to one's belief system.

3. Conclusion

I have introduced (1.) the standard view of counterfactuals:

First, counterfactuals have truth–conditions.
Second, these truth–conditions can be spelled out in terms of possible worlds.
Third, the possible worlds deciding on the truth of falsity of a counterfactual are those that minimally differ from the actual world.

I have presented (2.) challenges to the standard view. I summarize the results: (2.1.) there are no sufficient reasons to preserve inferences that are invalid in the standard semantics. (2.2.) Counterfactuals do not form 'disturbing noise' but teach interesting lessons on Gibbard cases. (2.3.) There are ways of accommodating lottery phenomena by minimally amending the standard semantics. (2.4.) The future similarity objection and (2.5.) the distinction between deviant and normal antecedent scenarios raise formidable difficulties to spelling out the standard account, but there are promising strategies of meeting these difficulties. (2.6.) Uncertainties about the modal status of the future are reflected in future-directed 'were'-conditionals.

Literature

Adams, Ernest, 'Subjunctive and Indicative Conditionals', *Foundations of Language*, 6 (1970), 89–94.

Ahmed, Arif, 'Out of the closet', *Analysis*, 71 (2011), 77–85.

Arntzenius, Frank, 'No regrets, or: Edith Piaf Revamps Decision Theory', *Erkenntnis*, 68 (2008), 277–297.

Barker, Stephen, 'Counterfactuals, Probabilistic Counterfactuals and causation', *Mind*, 108 (1999), 427–69.

——'Can Counterfactuals Really Be about Possible Worlds?', *Noûs*, 45 (2011), 557–576.

Barnett, David, 'Zif is If', *Mind*, 115 (2006), 519–565.

——'The Myth of the Categorical Counterfactual', *Philosophical Studies*, 144 (2009), 281–96.

——'Zif Would Have Been If: A Suppositional View of Counterfactuals', *Noûs*, 44 (2010), 269–304.

Bennett, Jonathan, 'Counterfactuals and Temporal Direction', *The Philosophical Review*, 93 (1984), 57–91.

——*A Philosophical Guide to Conditionals* (Oxford: Oxford University Press, 2003).

Bittner, Maria, 'Topical Referents for Individuals and Possibilities', *SALT*, 11 (2001), 36–55.

Brogaard, Berit, and Joe Salerno, 'Counterfactuals and Context', *Analysis*, 68 (2008), 39–46.

Cantwell, John, 'On an Alleged Counter–Example to Causal Decision Theory', *Synthese*, 173 (2010), 127–152.

Cappelen, Herman, *Philosophy without Intuitions* (Oxford: Oxford University Press, 2012).

Chalmers, David, 'Frege's Puzzle and the Objects of Credence', *Mind*, 120 (2011), 587–635.

Champlin, T. Stephen, 'Tendencies', *Proceedings of the Aristotelian Society*, 91 (1990), 119–133.

Chierchia, Gennaro, and Sally McConnell–Ginet, *Meaning and Grammar. An Introduction to Semantics* (Cambridge/Mass.: MIT Press, 1999).

Chisholm, Roderick, 'The Contrary–to–Fact Conditional', *Mind*, 55 (1946), 289–307.

Currie, Gregory, *The Nature of Fiction* (Cambridge: Cambridge University Press, 1990).

DeRose, Keith, 'Can It Be That It Would Have Been Even Though It Might Not Have Been?', *Philosophical Perspectives*, 33 (1999), 385–413.

——‘The Conditionals of Deliberation’, *Mind*, 119 (2010), 1–42.

Dunn, Jeffrey, ‘Fried Eggs, Thermodynamics, and the Special Sciences’, *The British Journal for the Philosophy of Science*, 62 (2011), 71–98.

Edgington, Dorothy, ‘On Conditionals’, *Mind*, 104 (1995), 235–330.

——‘Truth, Objectivity, Counterfactuals and Gibbard’, *Mind*, 106 (1997), 107–116.

Egan, Andy, ‘Some Counterexamples to Causal Decision Theory’, *The Philosophical Review*, 116 (2007), 93–114.

Elga, Adam, ‘Statistical Mechanics and the Asymmetry of Counterfactual Dependence’, *Philosophy of Science*, 68 (2001), S313–S324.

——‘Infinitesimal Chances and the Laws of Nature’, *Australasian Journal of Philosophy*, 82 (2004), 67–76.

Fara, Michael, ‘Dispositions and Habituals’, *Noûs*, 61 (2005), 43–82.

Fine, Kit, ‘Critical Notice: Counterfactuals’, *Mind*, 84 (1975), 451–58.

Geddes, Alexander, ‘Judgements about Thought Experiments’, forthcoming in *Mind*.

Geirsson, Heimir, ‘Conceivability and Defeasible Modal Justification’, *Philosophical Studies*, 122 (2005), 279–304.

Gendler, Tamar, ‘Philosophical Thought Experiment, Intuitions, and Cognitive Equilibrium’, *Midwest Studies in Philosophy*, 31 (2007), 68–89.

Gettier, Edmund, ‘Is Justified True Belief Knowledge?’, *Analysis*, 23 (1963), 121–123.

Gibbard, Alan, ‘Two recent theories of conditionals’, in *Ifs: Conditionals, Belief, Decision, Chance, and Time*, ed. by William L. Harper, Robert Stalnaker, Glenn Pearce (Dordrecht: Reidel, 1981), 211–47.

Gillies, Anthony, ‘Counterfactual Scorekeeping’, *Linguistics and Philosophy*, 30 (2007), 329–360.

Goodman, Nelson, ‘The Problem of Counterfactual Conditionals’, *The Journal of Philosophy*, 44 (1947), 113–128.

Grundmann, Thomas, and Joachim Horvath, ‘Thought Experiments and Deviant Realizations’, *Philosophical Studies*, 170 (2014), 525–533.

Gundersen, Lars Bo, ‘Outline of a New Semantics for Counterfactuals’, *Pacific Philosophical Quarterly*, 85 (2004), 1–20.

Hajek, Alan, ‘Most Counterfactuals are False’, unpublished Manuscript.

Hiddleston, Eric, ‘A causal theory of counterfactuals’, *Noûs*, 39 (2005), 632–657.

Ichikawa, Jonathan, ‘Knowing the Intuition and Knowing the Counterfactual’, *Philosophical Studies*, 145 (2009), 435–43.

——‘Quantifiers, Knowledge, and Counterfactuals’, *Philosophy and Phenomenological Research*, 82 (2011), 287–313.

Ichikawa, Jonathan, and Benjamin Jarvis, ‘Thought–Experiment Intuitions and Truth in Fiction’, *Philosophical Studies*, 142 (2009), 221–246.

Ichikawa, Jonathan, and Benjamin Jarvis, *The Rules of Thought* (Oxford: Oxford University Press, 2013).

Keynes, John Maynard, *A Treatise on Probability* (London: Macmillan, 1921).

Klecha, Peter, ‘Two Kinds of Sobel–sequences. Precision in Conditionals’, *Proceedings of WCCFL*, 32 (2015), 131–140.

Klinedienst, Nathan, ‘Quantified Conditionals and Conditional Excluded Middle’, *Journal of Semantics*, 28 (2011), 149–170.

Kment, Boris, ‘Counterfactuals and explanation’, *Mind*, 115 (2006), 261–310.

Koslicki, Karin, ‘The Semantics of Mass–Predicates’, *Noûs*, 33 (1999), 46–91.

Kratzer, Angelika, ‘Modality’, in *Semantics*, ed. by Arnim von Stechow and Dieter Wunderlich (Berlin: DeGruyter, 1991), 639-650.

Križ, Manuel, ‘Homogeneity, Non–Maximality, and *all*’, *The Journal of Semantics*, 33 (2016), 493–539.

Križ, Manuel, and Emanuel Chemla, ‘Two Methods to Find Truth–Value Gaps and Their Application to the Projection Problem of Homogeneity’, *Natural Language Semantics*, 23 (2015), 205–248.

Kung, Peter, ‘You Really Do Imagine It: Against Error Theories of Imagination’, forthcoming in *Noûs*

Kutach, Douglas, ‘The Entropy Theory of Counterfactuals’, *Philosophy of Science*, 69 (2002), 82–104.

Kvart, Igal, ‘Counterfactuals’, *Erkenntnis*, 36 (1992), 139–179.

——‘Causal Independence’, *Philosophy of Science*, 61 (1994), 96–114.

Leitgeb, Hannes, ‘A Probabilistic Semantics for Counterfactuals’, *The Review of Symbolic Logic*, 5 (2012), 26–121.

Leslie, Sarah–Jane, ‘Generics. Cognition and Acquisition’, *The Philosophical Review*, 117 (2008), 1–47.

Lewis, David, *Counterfactuals* (Oxford: Blackwell, 1973).

——'Causation', *The Journal of Philosophy*, 70 (1973), 556–567.

——'Score–keeping in a language game', *Journal of Philosophical Logic*, 8 (1979), 339–59.

——'New Work for a Theory of Universals', *Australasian Journal of Philosophy*, 61 (1983), 343–377.

——'Truth in Fiction', in *Philosophical Papers I* (Oxford: Oxford University Press, 1983), 261–280.

——'Counterfactual dependence and time's arrow', in *Philosophical Papers II* (Oxford: Oxford University Press, 1986), 32–66.

——'Humean Supervenience Debugged', *Mind*, 103 (1994), 473–490.

Lewis, Karen, 'Elusive Counterfactuals', *Noûs*, 50 (2016), 286–313.

——'Counterfactual Discourse in Context', *Noûs*, 52 (2018), 481–507.

Lowe, E.J., 'Wright versus Lewis on the Transitivity of Counterfactuals', *Analysis*, 44 (1984), 180–183.

——'Conditionals, context and transitivity', *Analysis*, 50 (1990), 80–87.

——'The Truth about Counterfactuals', *The Philosophical Quarterly*, 45 (1995), 41–59.

Lycan, William G., *Real Conditionals* (Oxford: Clarendon Press, 2005).

Malamud, Sophia, 'The Meaning of Plural Definites: A Decision–Theoretic Approach', *Semantics&Pragmatics*, 5 (2012), 1–58.

Malmgren, Anna Sara, 'Rationalism and the Content of Intuitive Judgements', *Mind*, 120 (2011), 263–327.

Mill, John Stuart, *A System of Logic Ratiocinative and Inductive* (London: Parker, 1843).

Morreau, Michael, 'It Simply Does Not Add Up: Trouble with Overall Similarity', *Journal of Philosophy*, 107 (2010), 469–490.

Morton, Adam, 'Can Edgington Gibbard Counterfactuals?', *Mind*, 106 (1997), 100–105.

——'Indicative versus Subjunctive in Future conditionals', *Analysis*, 64 (2004), 289–93.

Moss, Sarah, 'Subjunctive Credences and Semantic Humility', *Philosophy and Phenomenological Research*, 87 (2013), 251–78.

Nida–Rümelin, Martine, 'The Argument for Subject Body Dualism from Transtemporal Identity Defended', *Philosophy and Phenomenological Research*, 86 (2013), 702–714.

Nolan, Daniel, ‘Impossible Worlds*:* A Modest Approach’, *Notre Dame Journal of Formal Logic*, 38 (1997), 535–572.

Noordhoff, Paul, ‘Prospects for a counterfactual theory of causation,’ in *Cause and chance: causation in an indeterministic World*, ed. by Paul Dowe and Paul Noordhoff (London: Routledge, 2004), 188–201.

——‘Morgenbesser’s coin, counterfactuals and independence’, *Analysis*, 65 (2005), 261–263.

North, Jill, ‘What is the Problem about the Time–Asymmetry of Thermodynamics? – A Reply to Price’, *The British Journal for the Philosophy of Science*, 53 (2002), 121–136.

Northcott, Robert, ‘On Lewis, Schaffer and the Non–Reductive Evaluation of Counterfactuals’, *Theoria*, 75 (2009), 336–343.

Phillips, Ian, ‘Morgenbesser cases and closet determinism’, *Analysis,* 67 (2007), 42–49.

——‘Stuck in the closet: a reply to Ahmed’, *Analysis*, 71 (2011), 86–91.

Price, Huw, ‘Boltzmann’s Time Bomb’, *The British Journal for the Philosophy of Science*, 53 (2002), 83–119.

Olson, James, and Neil Roese, ‘A critical overview’, in *What might have been. The social psychology of counterfactual thinking*, ed. by James Olson and Neil Roese (Mahwah: Lawrence Earlbaum Associates, 1995), 1–57.

Rothschild, Daniel, ‘Do Indicative Conditionals Express Propositions?’, *Noûs*, 47 (2013), 49–68.

Ryle, Gilbert, *The Concept of Mind* (New York: Barnes and Noble, 1949).

Sandqvist, Tor, ‘Circularities in the Analysis of Counterfactuals’, *Studia Logica*, 73 (2003), 281–298.

Schaffer, Jonathan, ‘Counterfactuals, causal independence and conceptual circularity’, *Analysis*, 64 (2004), 299–309.

——‘Contrastive Causation’, *The Philosophical Review*, 114 (2005), 297–328.

——‘Deterministic Chance’, *British Journal for the Philosophy of Science*, 58 (2007), 113 – 140.

Schlenker, Philippe, ‘Conditionals as Definite Descriptions’, *Research on Language and Computation*, 2 (2004), 417–462.

Schulz, Moritz, ‘Counterfactuals and Arbitrariness’, *Mind*, 123 (2014), 1021–1055.

Sklar, Lawrence, ‘Causation in statistical mechanics’, in *The Oxford handbook of causation*, ed. by Helen Beebee, Christopher Hitchcock, Peter Menzies (Oxford: Oxford University Press, 2009), 661–672.

Smith, Michael, ‘Ceteris Paribus Conditionals and Comparative Normalcy’, *Journal of Philosophical Logic*, 36 (2007), 97–121.

Stalnaker, Robert, ‘A Theory of Conditionals’, *Studies in Logical Theory. American Philosophical Quarterly Monograph*, 2 (1968), 98–112.

——‘Indicative Conditionals’, *Philosophia*, 5 (1975), 269–286.

——‘A Defense of Conditional Excluded Middle’, in *Ifs: Conditionals, Belief, Decision, Chance, and Time*, ed. by William L. Harper, Robert Stalnaker, Glenn Pearce (Dordrecht: Reidel, 1981), 87–104.

Starr, William S., ‘What If?’, *Philosophers’ Imprint*, 14 (2014).

Van Inwagen, Peter, Modal Epistemology, *Philosophical Studies*, 92 (1998), 67–84.

Velleman, David, ‘Narrative Explanation’, *The Philosophical Review*, 112 (2003), 1–25.

Von Fintel, Kai, and Anthony Gillies, ‘“Might” Made Right’, in *Epistemic Modality*, ed. by Andy Egan and Brian Weatherson (Oxford: Oxford University Press, 2011), 108–130.

Von Heusinger, Klaus, ‘Choice Functions and the Anaphoric Semantics of Definite NPs’, *Research on Language and Computation*, 2 (2004), 309–29.

Walters, Lee, ‘Morgenbesser's Coin and Counterfactuals with True Components’, *Proceedings of the Aristotelian Society*, 99 (2009), 365–379.

Walton, Kendall, *Mimesis as Make–Believe. On the Foundations of the Representational Arts* (Cambridge/Mass.: Harvard University Press, 1990).

Wasserman, Ryan, ‘The future similarity objection revisited’, *Synthese*, 150 (2006), 57–67.

Weatherson, Brian, ‘What Good Are Counterexamples’, *Philosophical Studies*, 115 (2003), 1–31.

Williams, J. Robert G., ‘Chances, Counterfactuals, and Similarity’, *Philosophy and Phenomenological Research*, 78 (2008), 385–420.

——‘Defending Conditional Excluded Middle’, *Noûs*, 44 (2004), 650–668.

Williamson, Timothy, *The Philosophy of Philosophy* (Oxford: Blackwell, 2007).

——‘Replies to Ichikawa, Martin and Weinberg’, *Philosophical Studies*, 145 (2009), 465–476.

Won, Chinook, ‘Morgenbesser's Coin, Counterfactuals, and Causal vs. Probabilistic Independence’, *Erkenntnis*, 71 (2009), 345–354.

Wright, Crispin, ‘Keeping Track of Nozick’, *Analysis*, 43 (1983), 134–140.

——‘Comment on Lowe’, *Analysis*, 44 (1984), 183–185.

Yablo, Stephen, ‘Is Conceivability a Guide to Possibility?’, *Philosophy and Phenomenological Research*, 53 (1993), 1–42.

Yalcin, Seth, ‘Epistemic Modals’, *Mind*, 116 (2007), 983–1026.