

©2001

Douglas Neil Kutach

ALL RIGHTS RESERVED

ENTROPY AND COUNTERFACTUAL ASYMMETRY

by

DOUGLAS NEIL KUTACH

A Dissertation submitted to the
Graduate School—New Brunswick
Rutgers, The State University of New Jersey
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
Graduate Program in Philosophy
written under the direction of
Professor Tim Maudlin
and approved by

New Brunswick, New Jersey

January 2001

ABSTRACT OF THE DISSERTATION

Entropy and Counterfactual Asymmetry

By DOUGLAS NEIL KUTACH

Dissertation Director

Professor Tim Maudlin

I assess the thesis that counterfactual asymmetries are explained by an asymmetry of the global entropy at the temporal boundaries of the universe by developing a new semantic element for counterfactuals called objective assertibility and a method of evaluating counterfactuals that constrains consideration to possibilities where the early universe has low entropy. The resulting theory vindicates the common practice of holding the past mostly fixed under counterfactual supposition while at the same time allowing the counterfactual's antecedent to obtain by a natural physical development. While excellent for evaluating a wide variety of ordinary counterfactuals, it does not fully succeed as an explanation of counterfactual asymmetry.

CHAPTER II

COUNTERFACTUAL CONDITIONALS AND THE DIRECTION OF TIME

The following discussion is an exploration of how physics should inform theorizing about counterfactuals. Specifically, it addresses the form a theory of counterfactuals should take if we take seriously the conjecture that the various temporal asymmetries evident in many forms are fundamentally due to a simple physical asymmetry, the entropy asymmetry. My approach is to try to take on the one hand the universal increase of entropy and on the other the apparent openness of the future and fixedness of the past, and create a theory of counterfactuals that connects the two in a way that explains many of the recognized temporal asymmetries of our world. What will be gained by such a theory is a better understanding of the connections among the many 'arrows of time.'

Study of the direction of time in philosophy of physics has different aims. One is to determine whether, as far as physics is concerned, time is directed. In other words, are the laws of physics such that they require a temporal structure that essentially distinguishes between past and future? A second is to explain how certain physical facts fit in a unified explanation of various temporal asymmetries. A number of *prima facie* distinct asymmetries have been distinguished and are very well known. Among them are the asymmetries of entropy, radiation, cosmology, biology, knowledge, and free will. The asymmetry of entropy is the fact that entropy of physical systems decreases towards the past but not towards the future. The asymmetry of radiation is the fact that we have lots of objects in the world that have radiation traveling away from them spherically in the forward direction of time, but few if any objects that have radiation converging on them in spherical wave fronts. The cosmological asymmetry is that the universe in the past was very small, but in the future is much larger; simply, that the universe is expanding. The other asymmetries are a bit more nebulous. The biological asymmetry is actually a reference to a number of asymmetric processes like respiration, digestion, and parturition, all of which are asymmetric complex processes with asymmetric parts. The asymmetries of knowledge and free will are closely related to these biological asymmetries, and one can plausibly claim they are special cases. The knowledge asymmetry is a discrepancy between the kind of knowledge we can have of the past, e.g., by way of evidence or records of past events, and the kind of knowledge we can have of the future. The free will asymmetry is the apparent fact that the future is at least somewhat under volitional control. That is, we can influence the future, but not the past.

The asymmetries at the end of the list have traditionally been of great concern to philosophers as they are relevant to some central concepts philosophers have traditionally examined. As such, it would be a significant gain if we could understand the connection between these asymmetries and physical asymmetries like entropy. In this dissertation, I attempt to achieve part of this gain by constructing an overarching

framework in which the asymmetries of knowledge, free will, and causation are explained by the physical asymmetries. This framework is one in which counterfactual conditionals play a central role. The higher level asymmetries are explained by a counterfactual asymmetry that in turn is explained by lower level asymmetries.

The evidence used as a touchstone for theorizing about counterfactuals must be carefully selected. Because the aim is to illuminate a particular concept and its role in the overall explanation of asymmetries, the concept that needs to be considered is one that is implicit both in the context of fundamental physics and also in some significant subset of daily usage. If one wants a theory of the natural language conditional for its own sake, then one should investigate thoroughly the conditions of its usage in natural language, but such an endeavor can only by fortune be of great relevance to the philosophy of science. What is of interest is not the counterfactual conditional *per se*, but the conditional as it embodies certain scientific commitments. We should rule out of consideration, if at all possible, any concept of the counterfactual that does not fit with other most central scientific commitments. To be explicit, we think that if there are fundamental laws of nature, then these laws tell us what happens not only in our world but in other worlds where matter is placed differently. If a counterfactual is of the form “If *A* had obtained then *C* would have obtained,” and the fundamental laws have it that whenever *A* obtains, *C* obtains, that counterfactual must be true. The commitment to a physics-friendly concept of counterfactuals is not to be taken as a primitive dogma, as it may turn out upon investigation that no such concept is feasible, in which case this principle should be abandoned, but before investigation it is reasonable to maintain such a constraint because it is more central to our understanding than other viable alternatives like judgments of similarity among possible worlds. We should even be willing to set aside seemingly well-grounded principles of regularity that are a part of our folk understanding of nature, for example that scrambled eggs don’t become unscrambled, if the fundamental laws make it the case that had I scrambled a egg in a certain way, it would have unscrambled itself. One of the main reasons for trying to discover fundamental laws is that they can inform our understanding of the non-fundamental regularities. Upon learning that the fundamental laws allow the possibility (and in some circumstances the certainty) that an egg unscramble itself, we shouldn’t immediately dismiss the fundamental laws, but seek a different way of understanding our previous principles. Fortunately, there is a reasonable conception of counterfactuals that is friendly both to the fundamental physics and to more familiar contexts as well, as I hope to be able to demonstrate.

At the same time, our more limited concept of counterfactual must also make contact with objects not of fundamental physics. Otherwise, the purported role of the counterfactual in the explanation of temporal asymmetries would never be filled. There exists a large body of scholarship connecting counterfactuals to knowledge, causation and free will. For instance, many attempts have been made to analyze these concepts in terms of counterfactuals in one way or another. These attempts usually involve setting out necessary and sufficient conditions, at least one of which is expressed as a counterfactual conditional. A great deal of grief has usually come to such analyses from the notorious vagueness of counterfactuals, so that the conditions expressed as counterfactuals cannot be taken at face value but need to be understood within some specific context or with some specific standard of evaluation. One fairly well known

example of this is David Lewis' (1973b) counterfactual analysis of causation. To avoid generating the unseemly result that (what we accurately call) the effect often causes (what we accurately call) the cause, one needs to adopt a standard for evaluating counterfactuals that rules out this so-called problem of effects. In doing so, some intuitively correct counterfactual conditionals are ruled out: we must say that had the stone not broken the window, it would not have been because the stone wasn't thrown at the window. The stone, surely, would have been thrown at the window; it's just that some miraculous violation of nature's laws would have prevented the stone's seemingly predictable shattering of glass. If you accept this rendering of the counterfactual conditional, it must be on theoretical grounds, for instance to avoid the problem of effects, not on the basis of pre-theoretical intuitions about which counterfactuals are true.

Central to this connection is the vagueness inherent in counterfactual expressions referring to the objects and processes of everyday life. It will be convenient to start by considering counterfactuals with very ordinary antecedents and consequents, of the kind where they refer to physical happenings. These mundane counterfactuals must be among the propositions whose semantic values any good theory of counterfactuals must get right, any theory, that is, that purports to give semantics for the counterfactual conditional implicit in the normal use of counterfactual concepts. They are the conditionals that most clearly express the non-metaphorical use of counterfactuals, and the meaning of counterfactuals that we are after is the meaning that is implicated by standard usage. First, I will review the commonly accepted logic of counterfactuals and second, I will address a potential counterexample to the similarity analysis of counterfactuals.

The Logic of Counterfactuals

It might seem that the best thing to do before engaging in the debate over the logic and semantics of counterfactuals is to be clear about what counterfactuals are. Unfortunately, there exists widespread disagreement over which conditionals properly count as counterfactual conditionals. There is a core idea that seems to be well accepted. Sentences like

If the apple I ate had been green, there would be no more apples around,

are counterfactual conditionals, whereas sentences like

If the apple I ate was green, there are no more apples around,

are, in contrast, called indicative conditionals, or just indicatives. How much a sentence has to be like this one and in what respects is still open to question. For one thing, the name 'counterfactual' conditional is short for 'counter-to-fact' conditional, which indicates that the conditional in question postulates some non-actual state of affairs, i.e., has a false antecedent. One might be tempted into thinking that that means counterfactuals by definition imply the falsity of their antecedents. But arguments have been offered (Goodman, 1965; Ayers, 1965; Adams, 1976) that this would not be a good idea because it would separate out conditionals that share many other features similar to the example above. Borrowing an example (Edgington, 1995), "If you had dropped it, it would have broken." "You're right—I did drop it, and it broke, but I did such a marvelous repair job, you could never tell." Perhaps, it is better to say then, that the

conditionals we are interested in are those that are expressed in the subjunctive mood, regardless of the falsity of the antecedent. This is also problematic. For one thing, indirect speech uses the subjunctive mood. “He said we would meet inside if it rained,” is not an example of the kind of conditional we are interested in. When the sentence is cast in the future tense, the subjunctive and indicative seem to match in meaning: “If it were to rain tomorrow, we would meet inside,” versus “If it rains tomorrow, we will meet inside.” It’s unclear that we want to count the subjunctive as a counterfactual if it has the same truth conditions as the indicative conditional in this case. Ernest Adams (1975) also objects to the common “If I were you, ...” construction being counted as counterfactual on the grounds that it is subjunctive more by its just being an idiom than by it being a truly a counterfactual, although this argument is not very strong.

The consensus among some philosophers (Chisholm, 1946; Lewis, 1973a; Adams, 1975) is not to get bogged down in mire of terminological bickering, but to admit that neither term is adequate, and just to adopt the convention of calling the conditionals we are interested in ‘counterfactuals.’ The term can then be read as if accompanied by a knowing wink. For the sake of tradition, I adopt this convention as well, unless the subtle differences between the indicative conditional and counterfactual need to be carefully distinguished.

There are two broad theoretical approaches to counterfactual conditionals. The older approach, associated with Nelson Goodman because of his pioneering work, is the metalinguistic approach. With the metalinguistic approach, counterfactuals are understood to be a kind of elliptical expression of entailment. That is, conditionals of the form ‘If *A* were the case, then *C* would be the case,’ are taken to be true iff *A* together with some tacit premises entails *C*. Usually these tacit premises include all the laws of nature, plus some additional premises that express the background assumptions under which *A* is presumed to obtain. The difficult part of the program is to say which assumptions should be included among the set of legitimate tacit premises and which shouldn’t. For illustration, assume that Joe did not strike the match and consider

If Joe had struck the match, it would have lit.

Should I assume that the match would have been dry (as was the case), leading me to think that it would have lit? Or, should I assume that the match would not have been smoking (as was the case), leading me to think that it would not have lit? On the one hand, not allowing enough facts as assumptions leads some counterfactuals to come out false where they should come out true because the needed facts that entail the consequent are missing. On the other hand, allowing too many assumptions can lead to inconsistency in the set of assumptions because the totality of facts is inconsistent with the antecedent for any counterfactual that is truly counter-to-fact.

We often have a good intuitive grasp of which assumptions to hold and which to abandon when we are considering counterfactual situations, but to achieve an analysis of counterfactuals without appealing to non-actual situations, one must avoid referring to counterfactuals in a circular way. The game, for Goodman, was to discover a set of principles that will determine the set of assumptions, and hence the truth values of counterfactuals, without using counterfactuals. Goodman’s original project foundered on this so-called cotenability problem, but others have followed in the metalinguistic tradition with similar, albeit less ambitious, aims. Igal Kwart (1975), for one, rightly thinks the cotenability problem is just the problem of determining which counterfactuals

are true, and that there is no reason to think that we should be able to deduce the truth values of counterfactuals from occurrent facts alone. His own ambitious attempt (1975, 1980, 1986, 1991, 1992, 1994a) to treat counterfactuals in terms of probability relations among events, though, does not amount to a successful analysis of counterfactuals either.

The other approach to counterfactuals is based on modeling the conditional in accord with a possible world semantics. A full treatment of the conditional along these lines was first attempted by Robert Stalnaker (1968), and a more general theory that includes Stalnaker's as a special case was developed by David Lewis (1973a). Under Lewis's theory, counterfactuals are treated as a binary connective that is a variably strict conditional. The counterfactual $A \Box \rightarrow C$ is intended to be translated in English as something like, 'Had A been the case, then C would have been the case.' Other translations are possible and sometimes even necessary to account for the appropriate tense, *de dicto/de re* reference problems, etc. The truth conditions for $A \Box \rightarrow C$ are as follows:

If A is impossible, then $A \Box \rightarrow C$ is trivially true in world w .

Otherwise, $A \Box \rightarrow C$ is true in w iff there exists some A & C world with no A & $\sim C$ worlds as similar or more similar to w than it.

The root intuition behind the similarity account is to read 'If A were the case, then C would be the case,' as saying 'In all the A -worlds most similar to reality, C holds. The additional complexity in the more formal definition is due to Lewis phrasing the truth condition so that there can be an infinite procession of closer and closer A -worlds. Stalnaker disagrees with this fine point, defending the position that there is always a closest world, a claim embodied in an axiom of counterfactual logic, the Stalnaker assumption.* In general, such disagreement is possible amid broad consensus that the similarity account of counterfactuals is the best account of counterfactuals because there is some flexibility in the axioms one adopts. There is no need at this point to wade through the many combinations of axioms that one could adopt, except to note that a fairly simple combination of axioms is able to generate a logic that matches with our reasoning practices involving counterfactuals over a wide range of contexts. Indeed, no uncontroversial counterexamples to the core logic have been discovered.

Two attitudes towards the similarity accounts are worth distinguishing here. One might understand the similarity relation to be a relation imposed on the worlds by the various axioms of counterfactual logic. For any way of interpreting a set of claims where the axioms of counterfactual logic hold as necessary truths, there exists at least one similarity relation fulfilling the truth conditions of all counterfactuals. One might also go further to claim that this relation matches some interesting concept of similarity that we already possess. In other words, one might claim that the Stalnaker-Lewis logic gives an analysis of counterfactual statements in terms of a similarity relation. Evaluating worlds in terms of similarity in this way promises to avoid Goodman's puzzle about what facts to hold fixed in counterfactual contexts: we hold fixed everything except the minimum changes necessary to accommodate A . Overall similarity between worlds is the standard by which we can judge what changes are minimal. Nothing in the Stalnaker-Lewis

* The Stalnaker Assumption is a special case of the Limit Assumption, which merely claims that there is some set of closest A -worlds.

semantics decides which attitude is appropriate, and Lewis himself vacillates on which one should be adopted. This issue will be taken up in depth in chapter III, but in order to motivate the conclusion there, a potential counterexample to the semantics must be introduced.

The Junky Freezer

There is an abandoned house I pretend to know of that has a junky freezer in the kitchen. The freezer maintains its temperature well enough; its problem is that occasionally, due to its desperate state of disrepair, shudders. Because the ice bin inside is never emptied, every time the freezer shudders, the top-most ice cube falls out of the bin and onto the warm kitchen floor. I went by the house yesterday and observed the inactivity in the kitchen through a window. The freezer didn't shudder yesterday, but I am confident that had an ice cube fallen out of the freezer, it would have melted. We're in the middle of a heat wave, or at least that's what people are calling it.

On the accepted logic, whether the ice cube would have melted is a fact which depends on what happens in the closest possible worlds where the ice cube fell out of the freezer. By my naïve assessment of the counterfactual, if the ice cube had fallen out of the freezer, it would have been because a few seconds earlier, the sporadic malfunctioning freezer would have shuddered, nudging the cube out. Then, the usual forces of nature would slowly infuse the cube with heat, leaving only a puddle behind. If the possible worlds semantics is right, all the closest possible worlds have the cube thus melting.

The problem is, it's possible for the cube not to melt. That is, there are some ways the various atoms in the cube could bounce into the atoms in the kitchen floor and kitchen air such that the cube would maintain its sub-zero temperature or even get colder. These would be situations where by chance many fast atoms in the air heading towards the cube are deflected by other atoms in the air and where by chance many of the slowest atoms in the air reach the cube and are struck by some of the fastest atoms from the cube. These slow atoms are knocked back away from the cube with even greater speed than before. If the preponderance of imbalance between slow and fast molecules hitting the cube is sufficiently great, the air and floor will not transmit heat to the cube, and the cube will not melt. The probability is very low, as one can intuitively judge from all the coincidental collisions that are needed to keep hot atoms away from the cube, but it is a possibility nonetheless.

If even one of these possibilities is among the closest possible worlds, it will turn out false that the ice cube would have melted. What's more, there appears to be good reason to judge these possibilities to be among the closest. The hypothetical ice cube as it is falling out of the freezer is much like any other cube, its macroscopic features are identical to other cubes that will soon melt. Its differences from the other possible cubes that will melt is only a matter of a single atom here or there. These bizarre possibilities, possible worlds that thwart what would otherwise be a simple story about the truth of counterfactuals, do not constitute a counterexample to the possible worlds semantics so much as they indicate that the counterfactuals must be interpreted carefully.

To salvage the common-sense judgment that the ice cube would have melted there are several attitudes one might take:

1. Bizarre worlds are not among the most similar. They are worlds in which strange things happen. They have ice cubes that do not melt, odors that don't diffuse, salts that don't dissolve as they should, etc. If the similarity relation that counterfactuals obey is anything like an intuitive assessment of similarity, these worlds aren't close.
2. You cannot determine outright on the basis of superficial appearance what is most similar to what. Similarity may depend on a great many things, including fine microscopic detail. For all we know, it may be that the junky freezer is just one such case. Once we take into account the fine-grained differences between possible worlds, it is very likely (subjectively) that the ice cube melting worlds will not be among the closest. Thus, it is very likely (subjectively) that the counterfactual is true.
3. Fair enough. The counterfactual is strictly speaking untrue. Nevertheless, it is true that had the ice cube fallen out of the freezer, it almost certainly would have melted. Asserting this new counterfactual is asserting almost the same thing because we are rightly justified in ignoring such small probabilities when we use counterfactuals in our reasoning, so it's easy to see why we might say the untrue counterfactual when we really mean the true one.

All three responses have some plausibility and demand detailed exploration. In chapter III, I will take up all three responses, showing that the first doesn't work, that the second and third do work in junky-freezer cases, but only if we understand the semantics of counterfactuals in a particular way. We must take the similarity relation as a logical relation governing counterfactuals, without any distinct prior content that informs the evaluation of counterfactuals.

If we cannot analyze counterfactuals in terms of an antecedent notion of similarity, we must seek some other way to explain enough of the core intuitions about certain counterfactual judgments so as to permit a good explanation of the 'higher level' asymmetries. To do so, I suggest taking a physical asymmetry which we have overwhelming reason to accept, and then constructing a measure of the acceptability of a counterfactual that has this asymmetry 'built in.' By constraining a theory of counterfactuals on the one side by the physical asymmetry and on the other by its needed role in 'higher level' concepts, an acceptable form of the theory can be discovered.

In chapter IV, I consider David Albert's (2000) explanation of the asymmetries of intervention and knowledge based on the principle that the early universe had low entropy. With some modification, this skeleton of an explanation can be used as a structure for a theory of counterfactuals to flesh out. The rough idea is one of the lessons we learn from the history of the foundations of statistical mechanics is that we must assume that the early universe had very little entropy. We must assume it in order to make statistical mechanics consistent with the enormous number of fundamental beliefs we have about ourselves and about history. If we take this assumption to serve as a constraint on the possible worlds we consider when evaluating counterfactuals, an enormous reduction of the possible things which could have occurred in the past is achieved. With this reduction, hypothetical changes to the present are free to have enormous changes in the future, but constrained to have few changes in the past.

The way this difference is spelled out is in terms of statistical mechanical probabilities. Given the current state of the world, described macroscopically, there is some probability that a future event f , will happen and another probability that a past event p has happened. When considering how both past and future depend on the present, we consider the present state with hypothetical modifications and then conditionalize the statistical mechanical probability measure on the existence of a low entropy past. There are a lot of contexts in which after doing this conditionalization, the probability of p will remain constant and the probability of f won't. But there aren't nearly as many contexts in which the probability of f remains constant while p 's probability changes. This difference between the past and future is one which can be translated into a difference in the semantics of counterfactuals.

By measuring counterfactuals by the objective probability that the consequent obtains, given the antecedent, instead of by their truth conditions given in terms of similarity among possible worlds, the probabilistic asymmetry can be assimilated into a counterfactual asymmetry. In chapters V and VI, I describe this process in detail, but for a preview consider our counterfactual

If the ice cube had fallen out, it would have melted.

It is measured by considering all the worlds that are macroscopically just like the present world at the time when the hypothetical ice falling would have taken place except that instead of the cube being on the top of the pile in the freezer, it's falling out. Then, we consider a certain objective probability measure over the nomologically possible worlds consistent with this modified macrostate, conditionalize on the low entropy past, and determine how likely it is that the ice cube would have melted. This probability is the measure of the counterfactual and corresponds with the degree to which we should agree with the counterfactual. This procedure generalizes to all counterfactuals that are precise enough for such a probability to be determined, which comprise a large class of the counterfactuals possessing the temporal asymmetry under study.

As stated so far, though, this theory cannot be right. There are no worlds that differ only in that the ice cube is falling instead of being perched inside the freezer. For the cube to be falling, the air needs to be displaced, the freezer door needs to be open, there needs to be a shudder of the freezer, presumably with its acoustic accompaniment, etc. What is really needed here is a clarification of what is tacit in assuming that the ice cube is falling out. In part, it is determined by the context of assertion and the meaning of what is asserted, but it also is determined in part by the world. Even if the speaker knew nothing of the decorative magnet on the freezer door, part of what is tacit must be that the magnet moved as it invariably does with the motion of the freezer door. Sorting out these details is not mere nit-picking about the vagueness that governs nearly all counterfactual discourse because it is not clear, given what has been claimed so far, that there are any worlds that can fit a reasonable interpretation of the antecedent and still explain the temporal asymmetry. What's more, as will be seen in chapter VI, there are conflicting intuitions as to how the vagueness can be eliminated and no clear remediation.

Despite the difficulties posed by the vagueness of counterfactuals, the probabilistic theory is able to outperform competitors in the explanation of higher-level asymmetries. In chapter VI, the Entropy Theory of counterfactuals is elaborated and two examples of success are noted: for many ordinary circumstances, it is the case that if the

present were different in some specific way, then the distant past history would have happened essentially the way it actually happened. Yet, for many cases, the near past would have been different in just the way it needs to be different for the present to turn out different from the way it actually is. The theory vindicates the common intuition that present differences arise from past differences in quite ordinary ways. In the end, the theory is flawed in such a way that it cannot adequately explain the asymmetry of influence. Nevertheless, it is an acceptable theory for the evaluation of counterfactuals in ordinary contexts.

Looking at counterfactuals in terms of probabilities is reminiscent of a certain program in the history of theorizing about counterfactuals, the program of analyzing counterfactual conditionals about the past as a kind of indicative conditional. Brian Skyrms (1994) and Ernest Adams (1975, 1976) (see also Ayers, 1965) independently proposed to understand counterfactuals as a kind of “epistemic past tense.” (Adams, 1975, p. 103) Their ideas have been elaborated more recently (Ellis, 1984; Dudman, 1983, 1984, 1988, 1989). Their thinking was that the probability someone assigns to the counterfactual

If that bird were a canary then it would be yellow,
is equal to the probability that might have been attached to

If that bird is a canary then it will be yellow,

on a (possibly hypothetical) prior occasion. Suppose that prior to finding out the color of the bird, a person assigns the indicative conditional above a high probability. Then, if that person finds out that the bird is blue or is not a canary, he or she will presumably abandon affirming the indicative conditional, but continue to agree with the counterfactual conditional. All the evidence that supported assertion of the indicative continues to support the counterfactual, and the blueness of the canary in no way reduces the credibility of the counterfactual. It would be good if this project could be made to work, as it would lend theoretical unity to our understanding of conditionals and would explain a diverse set of observable logical phenomena. In the end though, Adams finds fault with the ‘epistemic past tense’ account of counterfactuals, and does not put forth any specific weaker theory to unify the conditionals.

There nonetheless remains room for unifying the counterfactual and indicative, and at the end of chapter V, I take up the project of showing how my account of counterfactuals explains some of the strong connections between the two kinds of conditionals. If we measure counterfactuals by the probability of *C* given *A*, in the manner prescribed earlier in chapter V, in many situations the indicative and subjunctive conditionals come to have the same probabilities. For example, when one’s epistemic probabilities are maximally informed, that is, when one’s assessment of the likelihood of a certain proposition would not change even if one learned the entire history of the world up to now, then the probability of the counterfactual on my account matches the epistemic probability. Other similar connections can be traced so as to provide good evidence for some kind of theoretical unity. In no case, though, is the indicative conditional reduced to the counterfactual or vice versa. Neither are the two reduced to some more fundamental logical structure. The unity is at best partial. One might think of the unity roughly as a matching of the assertibility conditions while maintaining dissimilarity in the form of their truth conditions.

CHAPTER III

COUNTEREXAMPLES TO ANALYSES OF THE SIMILARITY RELATION

Currently, the most widely received way of understanding counterfactuals is through their truth conditions given in terms of a similarity relation among possible worlds. The reasons in favor of similarity accounts are strong. Stalnaker and Lewis have defended logic systems that justify numerous reasonable inference forms and have no uncontroversial counterexamples. And if the axioms of counterfactual logic are those that Lewis or Stalnaker presents, the semantics for these systems are such that some relation of comparative similarity must exist among possible worlds. (Loewer, 1979)

Lewis (1979) has gone further in presenting an analysis of the similarity relation in terms of the Humean facts of our world. The project of describing more of the similarity relation than is necessary for it to play its role in counterfactual logic promises to inform us not only of the logic of counterfactuals but also of the truth conditions of counterfactuals in terms of facts about our world. That allows us the possibility of accurately judging the truth of many counterfactuals based on evidence of the kind we are usually able to gather. Due to vagueness, the relation can only be fixed to some degree, but it is reasonable to expect that a good analysis should match informed opinion in contexts where the acceptable counterfactuals are fairly clear. One broad context particularly appropriate for such a test is one where the antecedent and consequent are about physical happenings. The reason why we have a good antecedent reckoning of what counterfactuals are acceptable in these contexts is that laws of nature apparently govern the relationships between physical happenings, and we already have a good understanding of these laws—at least in approximation for the medium-sized objects of everyday life.

It will turn out that while our judgments of counterfactuals are governed by a logic that implicitly assumes a relation of comparative similarity, the similarity relation we use *cannot* be explained either as an intuitive judgment of similarity or as a composition of intuitive judgments of similarity. Rather, a different set of principles justifies our assessments of counterfactuals, including principles about chances and probabilities. When we examine how the similarity relation must be, in light of these principles, we are not finding the implicit notion of similarity underlying our judgments of counterfactuals because there is no univocal concept at work, but rather for every counterfactual there exist conventional constraints fixing to some degree or other the set of implicit similarity relations.

Lewis on the Similarity Relation

As described in chapter I, the semantics of counterfactuals given by Stalnaker and Lewis involve a similarity relation among possible worlds. To repeat, the

counterfactual $A \square \rightarrow C$ is non-trivially true in world w iff there is some A & C world with no A & $\sim C$ worlds more similar to w than it. Several restrictions on the similarity relation hold in order to guarantee that the theorems of counterfactual logic are tautologies, but in order to do more than settle some questions of counterfactual logic, more needs to be said about the similarity relation. We would need to know how to judge similarity in specific contexts, and Lewis offers a story about how to measure similarity.

Originally, Lewis was thought by some to have believed that the similarity relation implicit in counterfactual truth reasoning is the same as our offhand judgments about similarity. Yet, the analysis of the similarity relation *cannot* use our pre-theoretical assessment of the overall similarity among worlds. A number of critics (Bennett, 1974; Fine, 1975; Bowie, 1979; Jackson, 1977; Slote, 1978) have presented convincing objections that on an intuitive assessment of overall similarity, claims that we want to come out true, fail to do so on Lewis's account. One among them, Kit Fine imagines a world, w_0 , much like ours where Nixon is poised to press a button reliably linked to holocaust-causing nuclear missiles. In w_0 , Nixon decides at some tense moment not to press the button and as a consequence no holocaust occurs. Given the reliability of the electronics and launching system, the inhabitants of w_0 may rightly claim that if Nixon had pressed the button, a holocaust would have occurred. The future similarity objection to Lewis amounts to the claim that some world where Nixon presses the button and a miracle occurs to prevent a holocaust is closer to w_0 than any world where the button is pressed and a holocaust ensues. If these objectors are right then it turns out that if Nixon had pressed the button, there would have been no holocaust.

Lewis (1979) addresses the objection by rejecting the particular similarity relation that generates Fine's counterexample:

The presence or absence of a nuclear holocaust surely does contribute with overwhelming weight to some prominent similarity relations. (For instance, to one that governs the explicit judgment of similarity in the consequent of "If Nixon had pressed the button, the world would be very different.") But the relation that governs the counterfactual may not be one of these. It may nevertheless be a relation of overall similarity—not because it is likely to guide our explicit judgments of similarity, but rather because it is a resultant, under some system of weights or priorities, of a multitude of relations of similarity in particular respects. (p. 43)

The game, as Lewis sees it, is to take the knowledge we have about how counterfactuals work in specific contexts and see whether there is *any* set of relations that are intuitively respects of similarity that can be ranked or combination with relative weights to justify intuitively reasonable counterfactuals and explain other facts like the temporal asymmetry often seen in ordinary counterfactuals. We don't take some prior concept of similarity and accept or reject the analysis of counterfactuals on the basis of this notion of comparative similarity. Successfully completing this task will leave us with a similarity relation that, in combination with the truth conditions for counterfactuals, will justify the higher-level asymmetries like causation. (This isn't to say that for Lewis the resultant explanation of these asymmetries is *fundamentally* due to the similarity relation. Lewis offers an argument that the asymmetry of counterfactuals is a result of a contingent asymmetry in our world, the overdetermination of the past by the future. I will consider this argument in chapter IV.)

The interesting project for Lewis's analysis of counterfactuals is to characterize the similarity relation that governs judgment of counterfactuals in an informative (read: non-question-begging) way. The overall similarity relation Lewis proposes is defined by a ranked list of priorities:

1. It is of the first importance to avoid big, widespread, diverse violations of law.
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 small, localized, simple violations of law.
4. It is of little or no importance to secure approximate similarity of particular fact, even in matters that concern us greatly.

With this similarity relation, the world w_1 where Nixon presses the button and the holocaust occurs is closer to w_0 than other candidates because it contains only one small miracle to bring about the pressing and has poor approximate match of fact afterwards. There are several potential counter-example worlds compatible with Fine's description of the holocaust. One such world, w_3 , has a small miracle to bring about the pressing and has some other miracle to prevent the holocaust, by an interruption in the electric circuit. The second miracle only generates approximate match after the pressing because traces of the pressing still exist in the light images, fingerprints, memories, heat increase in the button, etc. Due to the additional miracle, w_3 counts as less similar. World w_4 has the same small miracle before the pressing but has another miracle to perform a perfect cover-up. This buys a very large time-span of perfect match but due to the widely spreading traces of the pressing, the perfect match comes at the cost of a large miracle. Hence, w_4 is less similar as well.

Under indeterminism, additional counterexample worlds must be handled. For Lewis's similarity relation to hold up, he needs for the closest worlds to be ones where Nixon presses the button and a holocaust occurs. In the best case, one such world, w_5 , diverges from the actual through some lawful indeterministic process bringing about the button pressing and then develops along a likely evolutionary path towards a holocaust. World w_6 evolves like w_5 but then has a localized divergent evolution disrupting the electric signal, generating an approximate match of physical fact to the actual world at no cost of miracles. Given the priority rankings, w_5 cannot count as closer to w_0 than w_6 . The only way w_5 can count as equally close is if approximate match counts for nothing at all.

Another world, w_7 , proves difficult as well: the evolution of w_7 matches w_5 's except shortly after the button pressing, a very unlikely evolution occurs so as to make w_7 match the actual world from that time on. In Lewis's terminology, w_7 contains a quasi-miracle. A quasi-miracle is a lawful indeterministic evolution that is distinguished by its bringing about physical states that would normally be produced through quite different mechanisms. A quasi-miracle looks like nature conspired to cover up the usual traces that events leave in the future. In w_7 , the quasi-miracle eliminates all the light images, fingerprints, memories, heat increase of the button, etc. so that perfect match with w_0 is established. Lewis suggests modifying the priority ranking so that quasi-miracles count heavily against similarity, presumably equal or nearly equal in weight to the large miracles usually required to create perfect match.

Lewis defends the introduction of quasi-miracles into the priority list from an appearance of inconsistency. With the modified priorities, we can definitely say that if

Nixon had pressed the button, there would not have been a convergence quasi-miracle. Yet, up until the supposed time, there is some positive chance in w_5 of such a quasi-miracle. The apparent conflict between the claim that the quasi-miracle would have had a chance and the claim that it definitely would not have occurred is abated because Lewis does not rule out the *chance* of a quasi-miracle, only the fulfillment of that chance. The weighting system is such that if Nixon had pressed the button, the positive chance of convergence in the closest worlds would exist unfulfilled.

The recourse to introducing quasi-miracles to the priority ranking has met with some disfavor because what events count as quasi-miracles seems to depend on what is surprising or conspiratorial to us human beings, whereas the truth conditions for counterfactuals should depend on non-anthropological features of the world. So far though, no one has provided a decisive argument against the introduction of quasi-miracles. What is fair to say, is that quasi-miracles add more complexity to the account in a way that detracts strength. Briefly, we will find that still further adjustments to the weighting system are necessary, further weakening the analysis. At a certain point, the additional refinements become mere gerrymandering, fitting the similarity metric to be whatever it needs to be to agree with intuitive judgments in clear cases. Because the notoriously shifting contexts implicit in the assessment of vague counterfactuals keeps the clear cases to a minimum, there is no shortage of opportunity for crooked boundaries.

Counterexamples from Statistical Mechanics

In chapter I, the example of the junky freezer was introduced as a potentially troubling counterexample to the standard way of thinking about counterfactuals. Remember that the facts of the case are this: there is a warm kitchen with a freezer that habitually, in a way that appears random, on occasion, due to some small localized electrical fault, shudders in a way that invariably knocks the door open briefly allowing the top-most ice cube to fall out of the ice box towards the empty kitchen floor. One day, there was no shuddering of the freezer and no cube fell out. We should all agree with the following:

If the ice cube had fallen out, it would have melted.

If the standard way of thinking about counterfactuals is right, then this sentence had better come out true, or at least (in some sense to be explored later) probably true. During the discussion of this example, I claimed that given what we know about natural mechanisms, there is a possibility that the ice cube would not melt. (This possibility is not a feature of my under-description of the example. That is, I am not referring among other things to the possibility that someone could have entered the kitchen and place the cube back in the freezer.)

Statistical mechanics has been able to account for this and many other aspects of thermodynamic behavior as well as other physical processes by understanding the behavior of gross material bodies in terms of their atomic constituents. One aspect that becomes puzzling once one adopts this perspective is the origin and explanation of irreversible physical processes. One way that this puzzle arises is when the microdynamical laws that describe the evolution through time of a system are weakly time-reversal invariant. By weakly time-reversal invariant, I mean that for any continuous set

of physical states $[p(t_1), p(t_2)]$, some continuous set of states $[p^\dagger(t_1), p^\dagger(t_2)]$ is dynamically possible, where $p^\dagger(t)$ is some state macroscopically identical to $p(t)$ for all t in $[t_1, t_2]$. It follows that for any system adhering to weak time-reversal invariance, if there is a physical evolution that involves an entropy increase, there exists a physically possible evolution that involves a net entropy decrease. One can demonstrate weak time-reversal invariance for many of our best working theories of fundamental physics: classical mechanics, classical electromagnetic theory, quantum mechanics, and general relativity. Thus, we have some reason to believe that the operative laws of our world obey weak time-reversal invariance. Because of the observed entropy increase of all the physical systems we typically encounter, we then know that there are some physical states that are macroscopically identical to actual states but that evolve in entropy-decreasing ways.

A good explanation for the observed monotonic increase of entropy was given by Boltzmann: for a reasonable probability measure of microscopic physical states, any set of states defined by macroscopic constraints that are not at equilibrium will have many more microstates evolving towards equilibrium than away from equilibrium. The equilibrium state for an isolated system is where it has maximal entropy. Thus, despite the one to one correspondence between states that evolve towards higher entropy and those that evolve towards lower, the preponderance of off-equilibrium states tend towards higher entropy. The increase of entropy for isolated non-equilibrium states is highly probable but not certain. This explanation of entropy increase means that the second law of thermodynamics, which asserts the universality of entropy increase, should be understood probabilistically.

Not only does entropy increase have possible exceptions, but also other apparently typical phenomena that doesn't *necessarily* fit strictly into the category of entropy-increase. Deflated rubber balloons can spontaneously inflate, lukewarm water freeze, solid metal bars crumble into metallic dust. We may reasonably denote worlds that contain such processes bizarre worlds. Given the wide variety of circumstances involving physical objects subject to such processes, we are rarely if ever free from the positive probability that bizarre evolutions occur. While the ubiquity of these chances for strange behavior usually does not call for attention due to their low probability, for someone interested in the truth-conditions of counterfactuals, these possibilities *are* troublesome. For if just one of the closest worlds is one of these bizarre worlds, many of the counterfactuals we routinely accept will turn out untrue.

The Deterministic Case

In the context of determinism, the bizarre world where the cube grows larger differs from the actual world in that it has perfect match of fact up until the one small miracle that launches the shudder of the freezer, no large miracles, and some degree of approximate match after the miracle. The more mundane worlds where the cube melts differ from actuality with perfect match up until the single small miracle, have no large miracles and have some degree of approximate match afterwards. So far, they are equivalent. Thus, we must look closer at the respects of similarity to find some way to judge the mundane worlds as more similar to actuality than the bizarre worlds.

Perhaps the correct response to the counterexample is not to be worried because the circumstances that permit the troublesome evolutions are so rare as to make it unlikely that the bizarre worlds count among the closest. Even if there is nothing ruling

out the bizarre worlds being among the closest, there is so far no reason to think they are equally close to more mundane worlds, so maybe we can ignore them. We may have good reason to do so if our similarity metric is very fine-grained so that the slightest details in how the miracle occurs makes some difference to the similarity. One special case of this is when there exists a limit world: when there is a closest A & C world unmatched in closeness by an A & $\sim C$ world. Other cases exist even without a limit world. If in the limit as one approaches the actual world, the kinds of A & C worlds one encounters become qualitatively more similar in their historical development, then the bizarre worlds may very well be left out. For us safely to ignore the bizarre worlds, there needs to be some reasonable sense in which the probability of the limit world's being bizarre is low. This kind of probability cannot be some objective chance because it is the probability of some unlawful process taking place, which objective chances do not give. The kind of probability we need is some objective probability measure over possible worlds because we need to make sense of the absolute probability of the limit world (or some set of closest or limiting worlds) being in the class of bizarre worlds. We already have some objective probability measure—the statistical mechanical measure that figures in Boltzmann's explanation of the entropy increase—that gives the appropriate low probability, so it might seem that we can put this probability measure to work. The problem is that the statistical mechanical measure is defined over a class of nomologically possible states constrained by certain macroscopic parameters. Yet, in the deterministic case, we must consider worlds that all contain miracles, and no statistical mechanical measure exists over miraculous worlds. Thus, we have no way of making sense of the probability of the limit world being bizarre.

Another way to count the bizarre worlds as rare may be to say that it is subjectively improbable that the ice cube would have stayed frozen. That is, we strongly believe that ice cubes melt in warm kitchens, and if sophisticated, we strongly, but not with total certainty, believe that a given ice cube in a warm room left undisturbed about which we have no further information will melt. By using this probability in evaluating what would have happened had an ice cube been on the kitchen floor, we will find the bizarre possibility to be overwhelmingly unlikely. Or rather, we *should* find the possibility to be overwhelmingly unlikely. The problem with using subjective probabilities to indicate the truth values of counterfactuals is that they are not normative. What would have happened is not a matter of subjective belief about what the probabilities are, but is a matter of objective fact. We need an *objective* probability whereby the bizarre worlds are rare.

The Indeterministic Case

In the indeterministic case, we have three ways to rule out a bizarre world being among the closest. First, there are material facts about how the physical state of the world evolves. Remember that criteria (4) allows for the possibility that approximate match of fact may count for something. If it does count for something, then perhaps we can rule out the bizarre worlds because they differ from ours in approximate fact. Unfortunately, this strategy has little to commend it because the bizarre worlds are little different from the mundane world with regard to approximating actuality. In the actual world, the ice cube did not fall out of the freezer, so both the worlds where the cube melts and those where it does not melt match the actual world only approximately. As to which approximates our world more, we might note that many bizarre worlds (perhaps

due to the bizarre physical evolution) have greater differences the mundane world—in some worlds, the cube's not melting causes a significant reaction among observers, leading to widespread reports of the strange occurrence. However, there are also worlds that approximate actuality better than the mundane worlds. There are bizarre worlds where the cube falls out of the freezer, but then spontaneously springs back in. These bizarre worlds rule out any hope that criteria (4) will provide the necessary principle.

We may wish to rule out the bizarre worlds just because of their bizarreness. In other worlds, we might be able to use criteria (1) as it was amended to count worlds with quasi-miracles as a respect of great dissimilarity. Remember that quasi-miracles are defined by their having the appearance of a conspiracy of chance outcomes to produce what would normally come about through different means. The worlds where the ice cube stays frozen may count as a world with a quasi-miracle and hence count as very dissimilar from our own world. The problem, though, with counting quasi-miracles as a weighty respect of dissimilarity for ruling out bizarre worlds is that they do not weigh enough. When we were trying to rule out convergence quasi-miracles, there was some reason to count w_7 as less similar to w_0 than w_5 . Whatever the future of w_0 is like, w_7 has one more quasi-miracle than w_0 because it matches w_0 's history perfectly after the reconvergence quasi-miracle, and almost every world like w_5 has no more quasi-miracles than w_0 . The matter is different when measuring bizarre worlds like our ice-not-melting world. Suppose that the actual world has a single quasi-miracle in its future. Because the future of ice-not-melting worlds will not match the actual world, they need not (and most will not) have our world's quasi-miracle. A typical ice-not-melting world will have a single quasi-miracle, just like our world. With respect to the prevalence of quasi-miracles, it is more similar to our world than a world with no quasi-miracles. It remains that one might hold that the general paucity of quasi-miracles in the actual world is still good reason to count any specific quasi-miracle as a mark of dissimilarity, but on what grounds? ** The lack of any apparent counterexamples should lead one to infer that the heavy weight of dissimilarity given to quasi-miracles doesn't allow one to rule out bizarre worlds from being among the closest antecedent-worlds.

The third way to rule out the bizarre worlds is to use facts about how likely the bizarre world is to be among the closest worlds. With determinism, we had no good way to make sense of the probability claim that objectively, the likelihood that the closest world is bizarre is low. If indeterminism is true, the situation is better in that there is some sensible probability if transition probabilities are defined for the required divergent evolution. In such cases, the reasonable probability measure to use is the measure that is established by transition laws defined over all the worlds that share a history up until just before the antecedent. There is a class of theories that are good examples of such varieties of indeterminism, namely quantum mechanics interpreted along the lines of the Ghirardi, Rimini, and Weber's (GRW) interpretation (1986) and similar variants. GRW defines a probability per unit of time of the full universe spontaneously changing into a given state. With this measure, the probability of a given world being bizarre is very small. As noted before, the low probability of the closest world or worlds being bizarre gives good reason to infer that if the counterfactual itself is probably not falsified by a bizarre world. This seems to offer some hope that if the laws are like those postulated by the GRW interpretation of quantum mechanics, we can count the problematic counterfactuals as unlikely to be true.

Unfortunately, this strategy will not succeed. To see how, consider a chancy device that when activated issues exactly one of two possible distinct outcomes, e or f . If the device is not activated, neither e nor f ensues. The circumstances and laws of nature are such that whenever the device is activated, there is a chance, c , of e occurring. (To rule out the relevance of the fourth respect of similarity, let the device's outcome have no additional consequences significant for the similarity relation. If you like, let e and f share exactly the same consequences after their occurrence.) In addition, assume that the microscopic details of how the device is activated have no bearing on the outcome; the outcome is based solely on a lawful chance process.

Now consider the conditional, "If the device had been activated, e would have occurred." If one admits no more than two truth values, either one can say that this conditional is determinately false, or one can say that its truth value is in some way underdetermined. Evaluating the expression as false is certainly due to holding that the similarity relation counts all the chance outcomes as equally similar. For if the chance c were anything less than certainty, the conditional would not be *determinately* false. There might be some subtle feature of the relevant under-described device-activated worlds that would have f occurring in the nearest worlds, and if so, we will be unable to rule out the bizarre worlds on the basis of their having low positive probability.

If one claims that the counterfactual is indeterminate, that is because one thinks that there are additional features that the similarity relation picks out, features that don't appear in Lewis's priority ranking of respects of similarity. What could these respects be like? The reason for thinking that the counterfactual is indeterminate is presumably because one wants to account for the fact that we think it is likely to degree c , that e would occur. That is, we want it to be the case that there is a probability c that the similarity metric combined with the antecedent implies that e occurs. No matter how we want to interpret this probability, objectively or subjectively, there are no plausible pre-theoretical respects of similarity that could generate such a result, regardless of priority ranking or weight.

The implausibility of using a prior notion of similarity alone to select out one determinate result from among many possible chance outcomes can be expressed more generally. Pick two non-actual worlds that differ in similarity to actuality. Take A to be the proposition true in just the world that is more dissimilar and let B be the proposition true in the more similar world. According to the account, $(A \vee B) \Box \rightarrow B$ is true because the B -world is closer. But compare this to the way the counterfactual is used. I give you two distinct full descriptions of the way the world could have been, A and B , and the information that A is far more likely to have been true, given some facts about chances. By taking into account the chances, you have good reason to think that $(A \vee B) \Box \rightarrow A$ is more likely to be true.

Bennett on the Similarity Relation

Jonathan Bennett (1984) has also argued that Lewis's analysis of the similarity metric fails although he presents different arguments. For example, Bennett agrees with John Pollock's argument (Nute, 1980) that attacks the need to have perfect match as long as possible. A patron at a club checks his coat in at the beginning of the evening and picks up his coat at midnight to leave. A steady stream of potential coat thieves passed out of the club all night long. Yet, if his coat had been stolen, it would have been stolen

just before midnight because this would have preserved the greatest region of perfect match.

In response, Bennett offers a different account of the similarity relation. He argues that we should take as a starting point a relation of *T*-closeness. *T*-closeness is a relation between worlds that holds in virtue of their properties over some time period or moment *T*. Bennett leaves open what the *T*-closeness relation is, but does claim that *T*-closeness is somehow related to *T*-similarity, which is the degree to which the various states of affairs at *T* at different worlds match. He cryptically adds that he is assuming that if *T*-closeness “doesn’t consist in *T*-similarity, then it at least implies it.” (p. 73) I think what we may take this to mean is that we should take some relation that counts formally as a similarity relation as primitive and call it *T*-closeness, and if all goes well, a ranking of worlds by *T*-closeness will generate a ranking of worlds that matches our intuitive assessment of overall similarity, whatever that may be.

Bennett’s theory is that $A \Box \rightarrow C$ is true iff *C* is equivalent to some conjunction (*R* & *S*) such that *R* is true at all the *T*-closest causally possible *A*-worlds, and *S* is true at some of those worlds and also at the actual world. (p. 76) For simple cases like the ones we want to consider, this reduces to $A \Box \rightarrow C$ being true iff all the *T*-closest *A*-worlds are *C*-worlds. By causally possible, Bennett has in mind at least the following: that the only *A*-worlds that should be considered are ones where all the laws of the actual world hold.

Bennett sees his theory as having several clear benefits over Lewis’s. For one, no appeal to miracles is made. It is odd on the face of it, that miracles have any role to play in the evaluation of ordinary counterfactuals as Lewis thinks they do. This oddness becomes downright counterintuitive when one considers the issue explicitly:

If Dukakis had won in 1988, some violation in the laws of physics would have had to have occurred.

Of course, this counterfactual is true only if determinism is true. Moreover, if determinism is true, the only alternative that would have Dukakis winning would be if the historical facts (arbitrarily long ago) were different. Since a choice must be made between preserving history and violating the laws of nature, it is not so surprising that miracles are allowed among the closest worlds. Lewis (1979) argues that we should opt for miracles because on the alternative, we cannot be certain that small differences in nearby times can be kept small in the past. It is possible (and indeed very likely, given some arguments to be made in chapter IV) that if we consider only worlds with laws like ours we will be considering worlds that have very different past histories. Thus, any counter-intuitiveness in the preceding example is equaled by the likely truth of some counterfactuals like the following:

If Dukakis had won in 1988, all the books about ancient Babylon would be wrong.

Regardless, Bennett sees some gain from the simplicity of not needing the various priorities that rank the worlds in terms of the size of miracles they possess.

Bennett also thinks that his theory is superior because it is able to provide a better explanation of backwards-looking counterfactuals, counterfactuals where the consequent is about a time before the antecedent. Although this issue will be addressed in great detail in chapter VI, part of Bennett’s argument is worth exploring here because it’s

possible that his theory can be parlayed into an argument that Lewis's account of backward-looking counterfactuals does not come out right, even though Bennett himself thinks that Lewis is safe here. Switching to one of Bennett's examples,

If Stevenson had been president in February 1953, then at his death the obituaries would have spoken of him as Eisenhower's liberal successor.

Bennett thinks this counterfactual is clearly false because if Stevenson were president in 1953, it would have been because he won the election in 1952 over Eisenhower. Bennett provides it as a counterexample to Frank Jackson (1977) who thinks that we should consider worlds that are similar to the actual world up until the time pertaining to the antecedent and then afterwards by the laws of nature. In this case, that would involve Eisenhower being elected, inaugurated, and then deposed or killed off somehow. Then, something odd with American politics would have occurred to allow Stevenson to be president instead of the vice-president taking over.

Lewis thinks that the transition between perfect match of fact and the antecedent's happening should take place as late as possible to achieve as much match as possible, but not so late as to have a great discontinuity. For instance, things will not instantly teleport themselves around to be as the antecedent says they are, but in some unspecified way make a smoother transition. Applying Lewis's characterization to the Stevenson example is not easy. On the one hand, the right thing to say seems to be that a small miracle would have occurred before the election, perhaps a brain miracle in Stevenson that caused a more persuasive speech, perhaps a subtle blunder by Eisenhower's staff member, a leaked indiscretion. The consequences of this small difference would have cascaded into Stevenson's election and the counterfactual would therefore be false. On the other hand, this way of resolving what happened does not lend itself to application in explaining causation, which is a main reason for Lewis providing a priority ranking. For causation, under Lewis's counterfactual account, to be asymmetric, it had better not be that thunder causes lightning. To avoid such cases of potential backwards causation, Lewis argues that on the standard resolution, it always turns out false that had the effect not happened, the cause would not have happened. This means that at least in the conditionals relevant to causation, we must use a resolution that does not trace back present states lawfully very far into the past. If the thunder had not occurred, the lightning still would have struck, but a miracle then would have occurred, preempting its noise. There is thus a discrepancy between the cases where Lewis wants to assume the miracle happened a while back so as to produce a plausible evolution to the events posited by the antecedent, and the cases where he wants to assume the miracle happened very recently to forestall backwards causation.

For all these problems of Lewis's theory, Bennett's partial theory is in no better condition. For one thing, there are simple counterexamples to his truth condition. Take a low probability event that actually occurred, say the roulette wheel ending up on 00. It turns out true on Bennett's theory, that had the wheel been spun with a different speed, the ball would have ended up on 00. For another, no prior notion of similarity plays any role in the theory. Which Stevenson-as-president-worlds are T-closest to actuality? For Bennett's sake, not the ones where last month's newspapers have stories about Eisenhower being inaugurated, but if widespread changes in the detailed physical makeup of the world do not count against similarity, what does? We are not given any reason to think that some notion can explain which worlds are T-closest.

Conclusion

It has turned out upon examination, analysis of the similarity metric—even when restricted just to physical contexts—is beset with insuperable difficulties. The principal attempts to provide such an analysis can be shown to be faulty. Specifically, David Lewis’s formidable (1979) analysis doesn’t adequately address cases where the operative laws involve chancy processes, including those processes whose chances are given by statistical mechanical probabilities, which comprise a large number of cases. The failure of Lewis’s particular account of the similarity relation is not due to its contradicting some intuitive assessment of certain counterfactuals but rather as an inability to pronounce any definitive judgment as to the truth values of counterfactuals in such situations. The particular similarity relation Lewis selects is simply insufficient to align the probabilities of the truth conditions of counterfactuals with the chances. Jonathan Bennett’s attempt to modify the analysis of the similarity relation fares no better as far as these cases are concerned. This failure over such a significant set of cases provides reason to believe that the similarity relation is a semantic primitive for the Lewis-Stalnaker counterfactual logic, and is not suitable for an analysis of the counterfactual conditional. This has the consequence of vitiating any explanation of the counterfactual asymmetry in terms of some feature of an independent similarity relation.

At stake is the source of the explanation implicit when an asymmetry of counterfactuals is used to explain a higher level asymmetry. The right conclusion to draw is that the similarity relation does govern the logic of counterfactuals, but it is a primitive relation, providing no basis for the analysis of the meaning of the counterfactual conditional. Rather, when people evaluate a counterfactual, or propose one, they tacitly assume some set of possible worlds where the antecedent is true, not necessarily worlds that count as the most similar. Certain consequences then follow from the choice of possible worlds. For instance, if one considers only worlds where a sample of gold is placed in a very hot environment and where the usual generalizations are operative, it follows that the gold melts in all those worlds. Evaluating the counterfactual amounts to evaluating the consequent, restricting one’s consideration to such worlds. If the consequent obtains in all the worlds, we say the counterfactual is true. If it obtains in none, we call it false. In the general case, more complicated rules apply. How worlds rank in terms of similarity is in part determined by the various logical implications of the axioms of counterfactual logic together and in part by certain tacit background assumptions.

Lewis’s Explanation of the Counterfactual Asymmetry

If the previous arguments about the similarity are correct, they should undermine one’s confidence in Lewis’s explanation of the counterfactual asymmetry. The following is an explanation of where the explanation goes wrong.

Lewis defines a determinant of the fact F as a minimal set of conditions jointly sufficient for F , given the laws of nature. All the facts in deterministic worlds, by definition, have at least one determinant at any time, namely the entire physical state at that time. It is possible that a fact have more than one distinct determinant at a given time. If so, the fact is *overdetermined*. If a given fact precedes more than one of its determinants, we say it is overdetermined by future facts, whereas if multiple determinants precede the fact in question, it is overdetermined by the past. Lewis thinks

that worlds like ours are properly characterized as having a predominance of events that are overdetermined by the future but not the past. One purportedly typical example of an overdetermining future fact is a part of an outgoing ripple from a stone's having been cast in a pond. Considering a time a few seconds after splashdown, there are numerous macroscopically distinct regions of the pond's surface that contain disturbances in the form of arced wave fronts receding from where the stone entered the pond. Lewis claims that the facts in each of these regions independently determine that the stone fell into the pond where it did. If one examines the portions of the world a few seconds before the stone entered, very few distinct regions can be found bearing facts that determine the stone's impact. Perhaps there are a few in the brain of the thrower, but there probably aren't any outside the brain. This asymmetry appears to be fairly general since in most real examples, there is some presence of radiation, whether it is sound, light, or something else.

The supposed overdetermination asymmetry is translated into a counterfactual asymmetry by way of the priority ranking of various respects of similarity (c.f., p. 15). For simplicity, consider a counterfactual of the form, "If c had happened, then ...," where c is some event. Because perfect match counts the most towards similarity, the closest worlds that at some time are different from actuality are going to be worlds that either match the actual world in the past up until some miracle just before the time when c happens, or in the future back until just after c happens. In either case, there will be a gap of time between the miracle and c 's occurrence in order to avoid large discontinuous or abrupt changes in the world. Because of this time gap, there is a difference between the kind of miracle that is required to bring about c . After c 's occurrence, there are a lot of determinants of c , and because each determinant's connection to the determined fact is nomological, it takes either a lot of little miracles or one large miracle to break that connection. Before c 's occurrence, there are few determinants, and hence only a small miracle or two is needed to bring about c . Because worlds with only a small miracle are closer than those with a large miracle, the worlds where the miracle precedes c are closer than those where the miracle follows c . Thus, the closest worlds are those that perfectly match the past of our world up until just before c , and then diverge according to the laws of nature.

Previously, several arguments were advanced to demonstrate the inadequacies of Lewis's priority ranking of respects of similarity. While those arguments do undermine the above attempt to explain the counterfactual, there is something right in Lewis's argument. To see this, a bit of clarification is in order. First, we should not understand Lewis's characterization of a determinant as being nomological determination because if we do there is no hope of getting any overdetermination asymmetry. While the complete physical state of the world at a time does nomologically determine facts both past and future, there is no plausible candidate for a physical law whereby spatially distinct parts of a physical state independently determine an event's existence. The bit of wave does not determine by itself that a stone splashed into the pond. For that, one would need the bit of wave together with a lot of detail about the surrounding environment. Classically, this would involve at least the entire physical state together with some additional restrictions to ensure determinism. Relativistically, this would require specifying the physics on some sufficiently large space-like hyper-surface. For the pond example where the post-splash time interval is roughly on the order of a second, this surface would

involve more than the physical state of the entire Earth. For almost any ordinary situation, the laws of nature by themselves plus localized bits of fact determine neither previous nor future facts.

To give Lewis's explanation a fighting chance at survival, one needs to understand determination as a weaker connection than nomological determination. The best reconstruction of what the explanation needs is a notion of determination that is something like a strong evidential connection. The piece of the pond ripple determines the existence of some pond disturbance at a particular location in the sense that the existence of a piece of ripple when conjoined with a set of claims about typical regularities (e.g. that water disturbances almost always dissipate energy over time, that heat almost always flows from hot to cold, etc.) implies the high likelihood that there was a pond disturbance. The degree of specificity with which this determination occurs depends on the determinant itself as well as the richness of the set of background conditions one conjoins. So a piece of pond ripple is a determinant perhaps of some pond disturbance but not a determinant of a stone falling in the pond.

Part of the appeal of Lewis's story comes from the fact that more often than not, the antecedents people typically hypothesize involve events that have an asymmetry of such determinants. Often the antecedent in itself is essentially time-asymmetric. Such antecedents typically concern events involving some radiation of sound or light which plausibly count as determinants of those events. Thus, merely the fact that one quantifies only over worlds where the antecedent is true, often insures that any post-antecedent convergence miracle must be larger than a pre-antecedent divergence miracle: throwing a stone into a pond typically has implications about a smaller volume at the beginning of the throw, namely that of the thrower, than at the end of the throw, which includes the stone in the pond, its ripples, as well as all the fleeing light images of the throw. For this reason, it is not terribly surprising that the priority ranking for worlds, by placing great importance on the size of miracles, is able to reflect this asymmetry. The victory is empty, however, because the asymmetry in these cases does not necessarily arise from some notable feature of the world, but is a feature of the kinds of situations that humans distinguish for practical reasons.

By setting aside counterfactuals that possess asymmetry in virtue of their asymmetric antecedents, one can see that Lewis's proposed explanation of the counterfactual asymmetry breaks down. That is, in the cases where extraneous sources of asymmetry are eliminated, the priority ranking of respects of similarity fails to deliver the required asymmetry. This is most perspicuously the case when the antecedent involves a highly localized event. Take an incandescent light bulb that was on continuously from time t_0 to t_1 to t_2 . Under plausible circumstances that we can assume, the following should come out true:

Had the light bulb burned out at t_1 , the room would be dark at t_2 .

For the light bulb to burn out, all that is needed is a small miracle somewhere in the thin fiber that constitutes the lighting element. This miracle would be equally large no matter whether it occurred before or after t_1 . What's more, if we believe in the current evidence for how the universe has evolved, including the big bang, it is reasonable to assume that the universe will exist longer than its current age of 10 billion years or so. That means that if perfect match counts for a lot, the closest worlds are worlds that match reality in their future history, i.e. in a history where the light goes out at t_1 are worlds where the

light was out at t_0 but not at t_2 . It is true that these worlds are radically unlike ours in that they possess radiation that looks like evidence for a light being on, even when it is not. Because these worlds perfectly match our world in future fact, people in these worlds who were in the room from t_0 to t_2 will think that the light has always been on. Video cameras recording the happenings will show evidence only of the light being on, etc. Although these worlds count as dissimilar from reality on some intuitive judgment of similarity, remember that we are not judging by some pre-theoretical notion, but by the priority ranking that Lewis advanced.

The problem is more general, though, occurring in cases even when the antecedent is about a spatially larger region. In Lewis's previous account of causation, in order to solve the problem of effects he needed to resolve the vagueness inherent in when the divergence miracle occurs by making the miracle happen after the cause but only a very short time before the effect would have occurred. This allows him to claim that if the effect hadn't happened, the cause still would have happened, but miraculously would have failed to produce its expected effect. This resolution avoids the potentially untoward consequence of Lewis's theory that causation is symmetric, that the effect causes the cause. Yet, to the extent that the miracle needs to happen a very short time before the effect would have occurred, the size of the miracle must grow to the size of the affected region.

If the truck were now on the other side of the street, ...,

when evaluated in a context that makes Lewis's story about causation come out right, requires a miracle that covers both sides of the street so that the truck can be in a continuous manner shifted over in a very short time. In such a context, there is no reason to think that a post-antecedent miracle could not just as well shift the truck's position so that perfect match exists post-antecedently. That is, there is no reason to think that later miracles need be any larger than earlier miracles and hence no reason to think that the problem of effects has been overcome.

Because these counterexamples demonstrate the inadequacy of Lewis's explanation of the counterfactual asymmetry, it is mysterious perhaps why the account enjoys as much success as it does. Why is it that so often for ordinary counterfactuals, the post-antecedent state takes up a larger spatial volume than the pre-antecedent state? The answer is that first, for obvious practical reasons, we humans tend to choose our antecedents such that they are spatially localized, and second, that our environment is full of radiative processes. Thus, any antecedent that involves ordinary processes will implicitly involve the accompanying radiative processes.

This fact often stands in the way of straightforward counterexamples to Lewis. Adam Elga has proposed a counterexample to Lewis's priority ranking of respects of similarity where there is an ordinary asymmetric process, Greta frying an egg that she cracked at $t = 0$. The counterfactual to be evaluated is

If Greta had not cracked the egg, then at $t = 5$ minutes later, there would not be fried egg on the pan.

The closest world where Greta doesn't fry an egg, Elga argues, is a world where there is fried egg on the pan, and perfect match of fact throughout later history, but also a small miracle somewhere in the 5 minute interval. The history preceding the miracle, just follows from the deterministic laws and the state of the world immediately preceding the

miracle. The miracle can be rigged easily so that it is false that Greta cracks the egg. Thus, the counterfactual comes out false on Lewis's account, when it quite plainly is true.

The important fact to note about this counterfactual, is that it is crucial that the antecedent is expressed as a negation, because it allows all sorts of possible worlds to count as antecedent-fulfilling, including ones where Greta does not exist. If the antecedent had been expressed as, "If Greta had done something other than crack the egg," a translation at which no one in any ordinary circumstance would balk, the counterexample world cannot be constructed. This is because Greta's doing something else, under any reasonably assumed context, involves her moving about in some normal way producing all the radiation of heat, light and sound that a person normally emits, the totality of which cannot be erased with a small miracle. Elga's counterexample is effective only for those who accept a simplistic understanding of the antecedent-fulfilling worlds as those where the antecedent, divorced from the context, is strictly true. Thus, his counterexample bolsters the argument on page 20 that demonstrated the silliness of applying the standard resolution of vagueness without regard to context.

The problem with Lewis's explanation of the counterfactual asymmetry are twofold. First, the explanation relies on the priority ranking of respects of similarity, which have been shown to have many problems, and second, it succeeds only where the antecedent or context already implicitly involve asymmetric physics. This leads one to reasonably conclude that the source of the counterfactual asymmetry isn't tied up in an asymmetry of *counterfactuals*, but is rather a counterfactual concerning asymmetric antecedents. Lewis is on to something with his thesis that the asymmetry is a result of a general overdetermination of the past by the future. The sense in which this is correct is that the general presence of radiation in the localized processes that humans care about, generates a kind of overdetermination, so that most typical antecedents will be temporally asymmetric, and lead in Lewis's account to a counterfactual asymmetry. While the accumulated arguments against Lewis's priority ranking are sufficient to demonstrate the need for some other account of the counterfactual asymmetry, it will be difficult to explain in simple physical terms the localized overdetermination that Lewis points out.

CHAPTER IV

DAVID ALBERT'S EXPLANATION OF THREE ASYMMETRIES VIA STATISTICAL MECHANICS

The central lesson of the history of attempts to make the apparatus of statistical mechanics consistent with our common sense beliefs about history is that we need to accept that the universe a long time ago had a very low entropy. This principle is virtually forced upon us if we want to take our memories, writings, and evidence for our scientific theories like statistical mechanics seriously. The gain from accepting this low entropy hypothesis is mainly that it rescues statistical mechanics from a head-on collision with accepted historical fact. That is, accepting the low entropy hypothesis is what we need to make statistical mechanics compatible with our common sense beliefs about the past. Yet, it does not offer much, if anything, in the way of predictions beyond what we already take, on other grounds, to be true of the world. What would be nice is if, once having been made to accept the low entropy hypothesis, we could put it to some productive use. One potential application is in explaining other asymmetries such as the asymmetries of influence, knowledge, and causation. Although there is a close connection between the fact that the universe long ago had low entropy and our seeming freedom to alter the future, in the end, I believe that the explanation for these asymmetries in terms of the low entropy hypothesis is at best partial.

The low entropy hypothesis is often presented as a claim about facts about the early universe. Further, this outlook often assumes that our universe is approximated by one of the various classical relativistic big-bang models. Assuming the existence of the big bang, we can pose the low entropy hypothesis as a claim about the early universe, presumably sometime at or before the change from a radiation-dominant phase to a matter-dominant phase. The claim is simply that at the big bang, the universe was in some low entropy macrostate. The hypothesis does not require the existence of the big bang or even the truth of general relativity. For instance, in a classical eternal cosmology, it can be expressed merely as a constraint on the physical state of the universe at some time, t , long ago, perhaps 10 or 20 billion years. In such a case we may often safely ignore what happened previous to t , as facts about times before t are likely to be empirically accessible to us only through the difference they make on the physical state at t . As such, I will adopt the convention of speaking of the past low entropy state as a fact about the early universe without assuming that the history of the universe is finite.

Statistical Mechanics and the Direction of Time

For the sake of later discussion, it is important to know why we have been put in the position where we must accept the low entropy hypothesis. The short story is as follows. Although we in the 20th century have not struck upon a true fundamental theory

of everything physical, we have found false theories that generate approximately the correct experimental results in some restricted domains and have the form of a fundamental theory. Some examples include classical mechanics, classical electromagnetism, non-relativistic quantum mechanics, and the general theory of relativity. These theories are our best proxies for a true theory of everything and are worth taking seriously as true theories for worlds that are in important ways like our own. One feature of these theories is that when properly understood as fundamental theories, they give dynamics for the microscopic constituents of the world, and in virtue of giving the dynamics at the microscopic level, they give dynamics for the macroscopic level. Yet the complexity of the dynamical equations is such that only in special idealized situations (or approximations to these situations) can solutions for macroscopic evolution be given. One of the great successes of science has been the application of statistical mechanics to bridge the gap between the micro and macro worlds in order to allow for explanations of macroscopic behavior in terms of the microscopic constituents of matter without knowledge of the precise details of how that microscopic matter is configured.

We can apply statistical mechanics to the special case of evaluating the evolution of macroscopic states under the fundamental dynamics that govern microscopic states. Consider the possible evolution of a system from macroscopic state A to macroscopic state B . To determine how likely A is to go to B , we consider all the microstates consistent with whatever macroscopic constraints exist on A ; that is, we consider a phase space with the set of states that count as A states. Because the microscopic dynamics defines an evolution on this space, we thereby have defined for any time t , the class of states that A evolves into after time t . If the phase space admits of an appropriate probability measure, then we can sensibly answer how likely it is that A will go to B after time t . It is just the relative proportion of A -after- t states that are B states.

One of the important consequences of this procedure of inference is that it turns out that, if A is not near the maximum entropy possible given the macroscopic constraints, the states A evolves into after a time Δt will almost all be of higher entropy than A . On the one hand, this is a good result, because it explains why entropy for isolated systems will almost invariably increase. That is, it explains why the second law of thermodynamics is such a successful approximation to the truth. On the other hand, this is a bad result because the increase holds whether or not Δt is negative or positive. The inference procedure predicts that if a system is in a non-maximal entropy state (or equivalently, not at equilibrium), then it almost certainly evolved from a state of higher entropy. The consequences of this result are enormous. For one thing, it shows that the second law of thermodynamics is false because it is highly likely, given the way the world is now, that the world was higher entropy in the past and thus that entropy has been decreasing to get to its current low state. More important, it implies that almost everything we take ourselves to know about the past is false. To see this, just note that the state of the world and of your local surroundings 5 minutes ago had lower entropy than the world and surroundings have right now. Applying the inference procedure to the world's current macrostate leads to a probability distribution that has a miniscule probability that the world even approximately matches the way you remember it being. Moreover, this procedure is self-undermining, in that it grants very low probability to the

existence of all the various experiments that gave evidence of the adequacy of statistical mechanical predictions.

One way to remedy this is to adopt as an additional principle to our inference procedure that the early universe was in a low entropy macrostate, i.e. the low entropy hypothesis. This immediately solves all the problems above. It is no longer the case that most likely the macrostate of the world 5 minutes ago was of higher entropy than it is now. For, given the initial macrostate, it is far more likely that the universe as a whole and its isolated parts in particular increased monotonically in entropy than that it increased to an amount significantly greater than the current entropy and then made an extremely improbable deviation towards lower entropy.* By making statistical mechanics compatible with our accepted beliefs about history, the low entropy hypothesis also removes the likelihood that statistical mechanics undermines its own credibility. Furthermore, the addition of the low entropy hypothesis to the procedures of inference do not in any way ruin the good result that it is highly likely that the evolution of isolated macroscopic systems towards the future will almost invariably increase in entropy.

To see that we *need* to adopt the low entropy hypothesis, several alternative strategies need to be considered. First, it is highly implausible that a mere restriction on the fundamental dynamics will save statistical mechanics. This would require a dynamics that essentially raises entropy in the forward time sense but decreases it in the backward sense. Certainly none of the theories that approach being fundamental theories have this feature. Even indeterministic theories that typically make very different predictions about the forward and backward evolutions of a system, do not have any mechanism to ensure the decrease of entropy towards the past. Often, they offer no complete theory about backward evolutions. In the GRW interpretation of quantum mechanics (c.f., p.19), for example, almost any past state is nomologically compatible with the current physical state. Most past states have a very small chance of giving rise to the current state, admittedly, but there is no way to calculate how likely each past state was, given the current state. The implausibility of accounting for the past low entropy in dynamical terms is clearest when one considers the kinds of theories dynamical theories are. Fundamental dynamical theories, if we are to be able to recognize their truth, or at least their accuracy, must hold true over a range of various physical microscopic configurations. But for the dynamics to make a system evolve towards lower entropy, it needs to coordinate large numbers of different particles in a way that is highly sensitive to the particles' dynamic variables. So a dynamics that could force evolution in the past towards low entropy would be of the kind that would be very difficult for us to recognize as a dynamics. (One could contrive some dynamics that can accomplish this goal, like a system where every particle's speed is an increasing function of the age of the universe. As one traces the motion of particles back in time under the influence of this dynamics, one gets ever slower, ever less energetic, particle collections with ever lower entropy. We can dismiss these dynamics if we assume certain plausible constraints such as energy conservation.)

* This way of putting it is an overstatement. There is certainly no proof for realistic systems that the probabilities *have* to work out that way. It is more accurately, a very plausible conjecture that evolutions with large-scale entropy decreases are less likely than the actual evolution.

Another possibility is that some constraint on physical states exists other than the constraint posed by the low entropy hypothesis. If this constraint were to entail the early low entropy, then then we could understand this other constraint as providing a deeper understanding of various temporal symmetries, so it could be seen as a refinement of the low entropy hypothesis rather than a hostile competitor. Alternatively, one might think that there could be some weaker constraint than the claim that the early universe had low entropy. This possibility is not realistic though, for the mundane reason that the low entropy hypothesis is so vague that there is no room left for a distinctly different hypothesis that can play the needed role in explaining asymmetries. If we are to take our observations of far away galaxies as credible evidence of the nature of the early universe, say at $t \sim 100$ million years after the big bang, then we must suppose that entropy was low at least that far back in time. This constraint is certainly weaker than the constraint that the universe be in a still lower entropy state at $t \sim 1$ year, but it is no different *in kind* from the stronger hypothesis so we may as well take the low entropy hypothesis to be the weaker if circumstances require.

The Knowledge Asymmetry

The knowledge asymmetry is evident in the remarkable difference between our epistemological access to past facts compared to our access to future facts. While it is unclear exactly which epistemological differences are in need of explanation, one reasonable conjecture worth examination is that the difference is constituted by a difference in the kinds of resources we can use to make successful inferences. David Albert claims there is a limit on what we can infer about future facts that does not exist for past facts. Almost everything we can know right now about the future, he argues, can be inferred in principle from what we know about the present macroscopic state of the world in combination with the existing dynamical laws and some general statistical features about the microscopic distribution of particles. We know far more about the past, though, than what can be inferred from this same information.

The additional inferential resources are in the form of records. Records allow us to learn details about the past way beyond what the dynamics allows us to infer. Albert recognizes under the name of a ‘record’ anything whose (not necessarily complete) state at two different times allows one to infer something reliably about the intervening time interval. R is a record of some state $f(t)$ at some time t in the interval between t_1 and t_2 iff R 's states at t_1 and t_2 are jointly reliably correlated with $f(t)$. A record is a relation between the state of a measured system and two states of the measuring device, a ready-state and an indication-state. Albert finds it important to distinguish this notion of record from some notion wherein the recorded state at one time is reliably correlated with the state of the measured system at another time. A fossilized dinosaur bone is a record of a dinosaur having existed sometime long ago. One might try to understand this record as simply a relation between the fossil and the dinosaur. This is a mistake Albert thinks, because the only way the fossil counts as reliable indicator of the dinosaur is that it is one temporal end of an historical process whose other temporal end is a muddy patch of earth ready to envelop whatever bones may fall in it. Without the rock surrounding the fossil having been soft enough long ago, the presence of the fossil does not correlate reliably with the previous existence of a dinosaur.

The importance of this distinction becomes apparent when one examines the inferential resources we possess. Our knowledge of the current state of the world can be understood as the knowledge that certain macroscopic facts presently obtain. (Anything we know about the microscopic nature of the present is only in virtue of our microscopes and Geiger counters, etc. reliably correlating microscopic facts with macroscopically discernable states.) Our knowledge about non-current states is accessible to us through the fundamental dynamical laws that tell us how microscopic states evolve into the future, and through certain rules of statistical mechanics that tell us the appropriate probability distribution for the microscopic states, given the macroscopic constraints. With this information, one has access in principle to the probability of any supposed fact about the future.

However, if we use this information to make inferences about the past, we must be very careful. If we are naïve enough to take the same kind of rules for determining the statistical distribution of microstates that we use for prediction, and use them for retrodiction, we will make widespread, demonstrably erroneous inferences. If the dynamical laws are the kinds of constraints that exist in deterministic fundamental theories like classical mechanics and general relativity and in Bohm's interpretation of quantum mechanics, then by reversibility arguments, one can show that among the inferences one would draw, there are claims to the effect that the past was of significantly higher entropy than it is now. In conjunction with what we know to be true about the current state, this implies that entropy decreased significantly, contrary to the second law of thermodynamics.

The way to be less naïve is to consider only probability distributions that arise from a low entropy past. To eliminate any conflict between what we obviously know to be true about the past and what this procedure tells us is likely to be true about the past, we may accept the low entropy hypothesis, the claim that one temporal end of the world's history (the one we call 'the past') is a very low entropy macroscopic state. When we take our original probability distribution and conditionalize on this low entropy hypothesis, we are able to avoid the conflict between our large base of beliefs about the past and what statistical mechanics tells us about the world.

Albert's interesting, substantive claim is that the low entropy hypothesis is the *only* additional claim we need to use in making inferences in order to explain the knowledge asymmetry. The explanation is that the low entropy constraint allows for the possibility of records by severely restricting the space of possible past histories consistent with the current facts. Because (so far as we know) only one temporal end of the universe is strongly restricted, there is an asymmetry in the kind of inferential power we have, an asymmetry of records. These records give us substantially more information than mere prediction or retrodiction via laws, so we can learn much more about the past and know past facts in a different way.

Albert does not detail how the early low entropy in our world is the source of the kinds of things we take to be records, but provides an idealized physics example of billiard balls on a table: 15 billiard balls are located on a normal billiard table. At time $t = 10$ seconds, the balls have various scattered positions and velocities which are known to us. Among these facts is the fact that the orange ball is now moving. If we want to determine whether the orange ball will avoid a collision over the next 10 seconds, we need to use all of our knowledge of the current positions and velocities of the balls in a

calculation of the future trajectory of the orange ball. The same goes if we want to know whether in the past 10 seconds, the orange ball avoided a collision.

However, our epistemological situation changes if we suppose that we know something else, that at time $t = 0$, a time we should think of as analogous to the big bang, the orange ball is at rest. If we know that the orange ball was at rest at $t = 0$ and that it is moving at $t = 10$, we can infer immediately that it underwent a collision without needing to use the information available about the other balls in a calculation. The additional fact we know about $t = 0$ constrains the possible trajectories of the orange ball to just those which involve collisions between $t = 0$ and $t = 10$, and knowing this fact allows us knowledge which would otherwise need to be derived from knowledge of the entire collection. With knowledge of the orange ball's motion at $t = 0$, we can infer that it had a collision in the first 10 seconds from only a small amount of data from the $t = 10$ physical state, i.e., that the orange ball is moving at $t = 10$. With this kind of knowledge of the $t = 0$ state, we can know facts about the past in a way that requires little knowledge of the present ($t = 10$) macrostate. Specifically, we can know these facts without knowing enough about the present macrostate to retrodict these facts, or even to attach any probability via some predictive/retrodictive procedure to the obtaining of these facts. So we can know things via small portions of the macrostate, i.e. records.

The billiard ball example demonstrates in a very simple way how an asymmetric restriction on a set of data can lead to some asymmetry of knowledge. While the example works well for this idealized case, its value is only as good as its applicability to realistic cases of knowledge. Judging its quality, it is important to keep the goal in mind. We would like to understand what physical facts we can point to as the source of the knowledge asymmetry. At a minimum, we want some story about how the low entropy past makes a knowledge asymmetry possible. To see that the explanation generated by the low entropy hypothesis reaches this minimum goal, it is enough to recognize that the low entropy condition, by placing very strong constraints on the kinds of worlds that can legitimately give rise to the known current facts.

But beyond making plausible the possible existence of *some* knowledge asymmetry, we also want to know how certain key physical facts account for the particular kind of knowledge asymmetry we have. Albert characterizes the asymmetry as follows:

Start with a probability distribution which is uniform—on the standard measure—over the world's present macrocondition. Conditionalize that distribution on all we take ourselves to know of the world's entire macroscopic history (and this will amount to precisely the same thing—if you think it over—as conditionalizing it on the past-hypothesis). Then evolve this conditionalized present-distribution, by means of the equations of motion, into the future.

THIS WILL YIELD (AMONG OTHER INFORMATION) EVERYTHING WE TAKE OURSELVES TO KNOW OF THE FUTURE.

Conversely:

Start with the same uniform probability distribution over the present macrocondition. Conditionalize this distribution on everything we take ourselves to know of the world's entire macroscopic future history.... Then evolve this conditionalized present-distribution, by means of the equations of motion, into the past.

THIS WILL YIELD IMMENSELY LESS THAN WE TAKE OURSELVES TO
KNOW OF THE PAST.

This is a fair characterization of the respect in which our knowledge of the past is superior to our knowledge of the future: our knowledge of the future is limited to what we can infer by evolving the probability distribution conditionalized on knowledge of the past, whereas the past is not so limited by knowledge of the future.

On closer examination, this characterization does not tell us much. It says essentially that there is knowledge above and beyond what we could get from predicting and retrodicting from the current macrostate. This knowledge is effectively knowledge of the past in the sense that any thing we know about the future above and beyond what the present macrostate and laws tells us is due to the fact that we know something additional about the past. All this distinction tells us is that we have a kind of knowledge different from the prediction/retrodiction type, namely knowledge via records, and we have this knowledge in a direct way only for past facts. The importance of this characterization is equal to the amount it tells us beyond what was already obvious, that we know a lot more about the past than the future and in a more detailed way. Without more information about the source and character of this additional knowledge, the value of the characterization is that by framing the difference in terms of prediction/retrodiction vs. knowledge by records, it affirms that the important difference between knowledge of the future and knowledge of the past is with respect only to knowledge via records. What it does is clarify the problem to be addressed.

It's reasonably clear, though, that Albert thinks he has done something more than just set up the problem. He thinks he has shown that the low entropy hypothesis (past-hypothesis) is the lone, crucial physical fact responsible for the knowledge asymmetry. (Of course, it's also crucial that at some time in history, some kind(s) of being somehow gain the physical mechanisms necessary for them to know things, but this is beside the point.) The identification of the early low entropy as the source of the knowledge asymmetry, of course, is a much stronger claim and certainly worth a close examination.

Remember that Albert finds it important to recognize that records, or measurements, are essentially triadic: in order to measure some physical system, one's measurement device must be in some ready-state before the measurement and in a measurement-indicating-state afterwards. One needs two states of the measuring device to make inferences about some intervening phenomena. So, a dinosaur-looking fossil is not a record of a dinosaur unless there was the right kind of conditions for the bone to be preserved. While all this is true, it is not a strong indictment of the view that records can be seen as relations of the measuring device and measured system. Of course, the state of some measuring device does not imply by itself or in addition to the dynamical laws of nature, anything about some measured system. As everyone recognizes, the device must interact with the measured system in some *appropriate* way. It is just part of what that appropriate way is, that the measuring device at some time be in a ready state, as the rock long ago was soft and muddy. Other conditions apply as well, e.g., that the Earth not be consumed by a great conflagration. In light of this, we should take Albert's recommendation for how to understand the logic of a record as an incorporation into the notion of a record, one of the conditions of 'appropriateness' of the measuring-device measured-system interaction.

Because there are additional conditions of appropriateness for objects to count as reliable records, merely adding the low entropy hypothesis to the prediction/retrodiction inference procedure is insufficient to allow us to infer all we know about the past. To see this, one needs only to note how in many cases we can have reliable beliefs about the past even when there are no records of the conditions that made those beliefs reliable. Consider Bill, an ordinary person, who was unfortunate enough to be victimized by error-prone bureaucracy and was confined in a deep underground bunker with 99 delusional patients who have consistent but systematically erroneous beliefs about all sorts of historical facts. What is at one time unfortunate can be a blessing later, as Bill found out when they finally escaped to a surface world wiped clean of any macroscopic traces of previous human existence by some appropriately cataclysmic disaster. Bill knows a lot of detailed facts about events before the holocaust, but Albert's inference procedure will not allow inference to these facts primarily because Bill's brain contains the only record of these events. If one takes the current macrostate and conditionalizes on the low entropy hypothesis, and if we are generous towards Albert's account, the bulk of the remaining possibilities will involve human beings somehow coming into existence through some more or less natural, probable evolutionary process. In no case, though, is it likely that the most probable physical evolutions are just those that make true all of the facts Bill knows. Even if the most likely way for Bill to get the beliefs he has is to have the events he knows took place actually happen, which is very dubious, we still have 99 other survivors whose brains can be equally plausible candidates for the true record of the past.

What this shows is that the low entropy hypothesis, while arguably a necessary condition for our knowledge of any particular fact, is not sufficient for it even in conjunction with the entire present macrostate. So the low entropy of the early universe cannot fully account for the knowledge asymmetry in the sense that its truth does not make probable the wide range of detailed knowledge people have.

Albert's reaction to this counterexample is to say that it overestimates the aim of his explanation. Albert thinks that the data to be explained—the set of things people can be said to know—are something like all the claims that a person is in a position to defend, given their current circumstances. So Bill's accurate beliefs do not count as “what he can be said to know” because he is in no better position than the others to defend his beliefs. The worry with this response is that, without some independent grasp on what beliefs count as the defensible ones, the explanation lacks content. It is far from clear that the defensible beliefs can be selected without circular appeal to the laws and principles that underlie the inference procedure that defines what is knowable.

Despite the inability of the low entropy hypothesis to explain the breadth of our knowledge via records, there is something to the low entropy hypothesis. On the one hand, because it severely constrains the physically possible evolutions consistent with our present macrostate, it does make plausible the possibility that one can infer past facts from the present way beyond what retrodiction allows. On the other hand, it doesn't allow all past facts to be inferred, not even the small subset of facts that count as known facts. So it is an open question, how much of a constraint the low entropy provides, and how much it can inform us about the knowledge asymmetry. Unfortunately, for most any realistic case, the best information we could get would be merely speculative conjecture about what evolutions are more likely than others.

The Asymmetry of Influence

Another asymmetry is exhibited in the behavior of decision making creatures: it is commonly accepted in our unreflective moments that humans and other creatures like us have the ability to change the future in a way that bricks cannot, yet we lack any such power with respect to the past. After a long history of philosophical investigation into the nature of human action, we have gained some insight despite the seemingly intractable debate over the nature of freedom. There are all sorts of ways humans influence the world. One way is simply that people figure in the dynamical evolution of the physics of the world in virtue of their having physical bodies. This is obviously not the relevant kind of change because bricks have this ability as well. What is different about humans, as it is commonly expressed, is that humans have the power to do things other than what they actually do—for shorthand, the power to do otherwise. The exact way this power is explicated leads to central philosophical differences over free will.

Incompatibilists, those who hold that necessarily free will does not exist in deterministic worlds, often argue that we should understand the power to do otherwise as implying the possibility of doing otherwise in the very same circumstances, all the way down to the last microscopic detail. After all, bricks have the power to do otherwise in the sense that they can bring about effects other than those they actually bring about if they were to have different circumstances (e.g., different velocities). This is forbidden in deterministic worlds, so it follows that we are necessarily not free if determinism is true. Compatibilists, those who hold that free will can exist in some deterministic worlds, often hold that we should understand the power to do otherwise less restrictively. Being free implies that there exist very similar circumstances, e.g., circumstances exactly the same as in actuality except that I had different desires or motives, I would have done otherwise. Both senses of the power to do otherwise are cogent. The crux of the debate between compatibilists and incompatibilists is over which sense is appropriate for whatever job we need the concept of free will to perform. For example, we seem to need a notion of free will to explain culpability, so the heart of the debate over free will should be focused on what notion of free will can adequately fit in an explanation of culpability.

Fortunately, we can circumvent this fracas because the explanation of the asymmetry of free will is an explanation of the intuition that people commonly have that they are free to alter the future but not the past. Whether this intuition captures a sense of freedom necessary for ethics is irrelevant. All that needs explaining is *some* notion of free will that captures a good part of what we intend by freedom. Even if free will is an illusion, we can still explain the lure of that illusion. If the compatibilist notion of freewill, as described above, can be shown to serve in an adequate explanation of the asymmetry of free will, then the incompatibilist notion will easily fit into the explanation as well, assuming plausibly, that the explanation in question holds in indeterministic cases.

In fact, for an explanation of the free will asymmetry we need only consider certain clear-cut cases of compatibilist freedom. There are normally understood to be several necessary conditions for a person's action to be free. It must be uncoerced, the person must not be brainwashed or influenced by mind-altering drugs, etc. We can ignore all these conditions that deny a person free will and instead focus on one central necessary condition that in circumstances with no mitigating factors will also be

sufficient for free will. This condition for free agency is expressed by the following counterfactual:

If the agent in question had decided to act in a way other than she did in fact decide, she would have acted otherwise.

(The condition is presented in the past tense because it is in the past tense that the counterfactual conditional is most clearly distinguishable from the indicative conditional. It can be phrased in the present and future tense as well so long as one does not confuse the counterfactual conditional with other conditionals that many people think to underlie the usage of indicative conditional, namely the material conditional.) Because this condition is a criterion for an agent's action making a difference or influencing an outcome, the associated asymmetry is known as the influence asymmetry. Thus, the asymmetry of free will can be understood as a special case of the influence asymmetry, which includes events other than decisions that can influence outcomes.

There is room for disagreement with the assertion that the influence asymmetry is constituted by a counterfactual difference. There is a distinction famously made by John Locke between actions that are voluntary and those that are done with liberty. In his example, a sleeping person is carried to a room and locked inside. Upon waking, the person finds someone in the room whom he or she wanted to spend time. The person stays in the room voluntarily because the person's will, acting in accordance with desires, dictates staying. The person does not act with liberty, though, because he or she could not leave if that were desired.

One might wish to associate the notion of influence with whatever issues from the will, instead of pinning influence on counterfactual dependence. This is central to arguments that use the so-called Frankfurt cases to show that the agent is free despite the relevant counterfactual being false. These are situations where some device interrupts the usual counterfactual dependence. In reality, Amy considered the options of vanilla and chocolate, decided on chocolate, and ordered chocolate. Normally, we would think it true that had she wanted vanilla, she would have ordered vanilla. But in a Frankfurt case, if she had decided vanilla, she would not have ordered vanilla because some device or other would have detected the desire for vanilla and would have interfered with Amy's thinking so as to make her order chocolate. The intuition of some is that Amy's action was free even though had she wanted to do otherwise, she would not have acted otherwise. This intuition is nothing more than the association of free choice with what is voluntary in Locke's sense.

While the merits of this association can be debated, in what follows, influence is associated with counterfactual dependence. It will turn out, after some theoretical development that there will be an important sense in which there is backwards counterfactual dependence. That is, there will be events in the past that counterfactually depend on events in the present. If these cases are legitimate, it will then become a live issue again. If the commitment is strong to the connection between influence and counterfactual dependence, these cases will demonstrate surprisingly that we can influence the past. If one is committed even more to the idea that in ordinary situations, we cannot influence the past, then these cases refute the claim that counterfactual dependence implies influence. The goal in what follows is to investigate whether the influence asymmetry, i.e. that we can influence the future but not the past, is due to some kind of temporal asymmetry in the counterfactual.

David Albert's Explanation of the Influence Asymmetry

Returning to the billiard ball example, if we were to modify the present ($t = 10$) macrostate by fiddling with the velocity of the red ball or the position of the green ball, we might very well change whether the orange ball has a collision in the next 10 seconds. Indeed, there are many ways we can modify the ($t = 10$) macrostate, many of which will imply that the orange ball will have a collision in the next 10 seconds, and many of which will imply no collision. However, there are very few modifications of the present state that will imply that the orange ball will avoid a collision in the past 10 seconds. Remember that that collision was guaranteed by the initial ($t = 0$) macrostate that had the orange ball being still and the ($t = 10$) macrostate that has the orange ball moving. Only modifications that put the orange ball at rest (among other restrictions) are consistent with the ball having had no collision. In some intuitive sense, there are many more ways to affect whether the ball will have a collision than whether it did have a collision.

The sense in which there are more ways to affect the future is that there is a larger region of space in which localized modifications to the present state (of the kind we can be said to have control over) can affect the future than the region in which localized modifications can affect the past. Albert's claim is that the influence asymmetry is explained by there being more handles—more localized regions which make some macroscopic difference over which we can be said to have control—on the future than on the past. This in turn is explained by some theory of counterfactual conditionals.

We start by taking for granted that there are some things over which we have control. This is necessary because the debate over explanations of an influence asymmetry only make sense if the existence of influence is taken for granted. Details about the source or ground of this control are left aside. All that one needs to assume is that among the things under our control, the predominant physical objects of our control are localized. For example, we may be said to be in control of portions of our brains, or perhaps significant parts of our bodies like fingers and legs. Then, we note that there are a lot of regions that make a difference in the future but not the past. Assuming that there is no surprising negative correlation between something's ability to make a difference and its falling under our control, the things we are able to control will more likely make a difference in the future, but not in the past. Thus, we will have more handles on the future than the past.

The way counterfactuals enter the story is in spelling out what it means for a region to make a difference. Although Albert does not present any specific account of counterfactuals, he clearly thinks some theory of counterfactuals can do the required job of fitting into this explanation. To evaluate his claim, we need to investigate a more explicit theory of evaluating counterfactuals, one that will not only fit in the idealized billiard ball example, but also for more realistic examples. Otherwise, Albert's explanation may meet the same fate as his explanation of the knowledge asymmetry.

For an introductory stab at a more realistic explanation of the influence asymmetry, consider the counterfactual, "If Amy had wanted vanilla, she would have gotten vanilla." Take the macrostate at the time Amy decided what she wanted (which was chocolate) and modify this state by rearranging a few of the physical variables corresponding to what's going on in her brain. Take this modified state and consider all

the microscopic states consistent with it, and the statistical mechanical probability measure over these states. In other words, take all the possible worlds that have the modified macrostate in question and obey the fundamental laws of physics, and the appropriate probability measure over them. Conditionalize this probability distribution on the low entropy hypothesis, and voila! You get some set of worlds. If most of these worlds have Amy getting vanilla, then the counterfactual is one worth assenting to. If she does not, then the counterfactual is worth denying. The kinds of situations where it is worth accepting, if Albert is correct, should be exactly those where we would intuitively accept the counterfactual, i.e., situations where Amy has the ability to communicate with someone selling the goods, she has the requisite money, the retailer has sufficient stock of vanilla, etc.

Now consider whether Amy, after eating the chocolate, can influence her previous choice. She does this successfully only if the following counterfactual is true for some action X that Amy can perform:

If Amy were to X right now, she would have ordered vanilla a few minutes ago.

Evaluating this class of counterfactuals with the exact same procedure leads to some quite different results. Taking the current macrostate, modifying it so that Amy's action is different and then taking all the possible worlds corresponding to all the possible microstates consistent with that macrostate, and then conditionalizing on the low entropy hypothesis, does not lead to her ordering vanilla in the past. Despite any modifications one wants to make concerning Amy's mental state, there are chocolate stains on her lips, missing chocolate from the confectionery supply, memories of bystanders, and images of her mouthing the word 'chocolate' traveling deep into space all exist in the current state of the world. These facts about the current state, in conjunction with the truth of the low entropy hypothesis, make it the case that she very probably ordered chocolate. So, very probably, the counterfactual is false.

These examples, if realistic should reasonably generalize so that counterfactuals of the kind that are implicit in the exercise of free actions will turn out at least very probably false when the action precedes the decision. The counterfactuals where the action follows the decision will turn out true in many cases: true in the cases paradigmatic of free will, and false in cases where the will is not efficacious. If this works out, it will demonstrate a counterfactual dependence, at least in the cases of human actions, that is asymmetric, and will therefore explain why we can influence the future but not the past.

The explanation of how Albert's general strategy would have to work, implicitly relied on some controversial assumptions that need to be made explicit. For one, the statistical mechanical probability distribution was associated with a degree of assent to a counterfactual, yet the applicability of degrees of assent to counterfactuals is mostly terra incognita. For another, Albert has only given a single idealized example that he hopes can be generalized, and there is no general argument as to why the idealized result should hold in realistic cases. In the following two chapters, these two issues will be discussed.

CHAPTER V

THE PROBABILISTIC THEORY OF COUNTERFACTUALS

The English conditional is commonly classified by philosophers into two kinds, the indicative and the subjunctive. Motivation for this distinction comes partly from the many clear cases where our attitudes between an indicative conditional and its subjunctive counterpart differ greatly, but also from the significant success of quite different theoretical treatments for each kind. The logic of subjunctive conditionals is modeled with a relation of comparative similarity, and the logic of indicative conditionals is modeled with assertibility ratings given by subjunctive conditional probabilities. The division, while commonly accepted, has yet to be explained to satisfaction. Disagreement exists over where the boundary between the two is properly drawn (Bennett, 1988), and even whether there is a boundary at all (Edgington, 1995). The issue is not a merely a taxonomic squabble. A successful solution promises to give a coherent theory of conditionals generally that explains why the subjunctive and indicative agree and disagree in meaning where they do. With this end in mind, I suggest an explanation for the unity of conditionals, to the extent that they are unified, in terms of a certain theory of counterfactual conditionals. Under the theory I present, counterfactual conditionals are evaluated by their objective assertibility. It turns out that in circumstances where one would normally judge a subjunctive and indicative conditional to be synonymous, the assertibility one would assign to the subjunctive equals the assertibility of the indicative, and where one would judge them non-synonymous, the assertibilities come apart. This theory does not pretend to treat counterfactuals generally, but merely to model the evaluation of counterfactuals that involve physical happenings. The discussion will focus on fairly clear counterfactuals about happenings that can be localized in space and time, but the theory applies more broadly.

The Dual Approach

The simplest argument that indicatives and subjunctives must differ can be lifted from Adams (1970) with a harmless modification. For the same antecedent and consequent, we can construct two distinct sentences:

- (1) If Booth didn't kill Lincoln, someone else did.
- (2) If Booth hadn't killed Lincoln, someone else would have.

Historical evidence leads us reasonably to agree with (1) and disagree with (2). Since (1) and (2) have different degrees of plausibility, they must differ in meaning. The meaning of the conditional statement is constituted by the meaning of the conditional connective

and the meanings of the antecedent and consequent. Since (1) and (2) do not differ in the meanings of their antecedents nor their consequents, they must differ in the meaning of their conditional connectives.

Another argument in favor of the dual account is that important advances in the understanding of conditionals have come in two parts. Indicative conditionals have found an extremely successful model in epistemic conditional probabilities that give their assertibility conditions. The assertibility of a conditional, “If A , then C ,” for a person P is the degree to which P is willing to assert this conditional. Two clarifications of the concept of assertibility are helpful. Originally, it was essential that the notion of assertibility involve public linguistic production because it was part of Grice’s (1975) explanation of the paradoxes of the material conditional: We really do believe that if the moon is made of cheese, then Elvis is alive because we believe ‘The moon is made of cheese \supset Elvis is alive’, but we would never *say* that because it would mislead listeners who quite reasonably expect us to avoid so understating our beliefs. These days, because assertibility plays a useful theoretical role outside Grice’s explanation, assertibility is a term for a broad collection of epistemic attitudes including willingness to assent and willingness to believe. Also, the concept of assertibility departs from its non-technical meaning by abstracting away from many context dependent factors that affect one’s willingness to assert. A conditional involving profanity or sensitive secrets is not for that reason any less (or more) assertible.

The great success of Adams (1965, 1966, 1975, 1996) was to show that while the assertibility of most propositions is equal to the degree the speaker thinks the proposition true, the assertibility of a conditional expression is equal to his or her subjective conditional probability of the consequent given the antecedent. His thesis is backed strongly by two kinds of evidence: First, by direct appeal to intuition, one can see in many examples that assertibility values close to one are associated with strong likelihood to assert the conditional, and values close to zero are associated with strong unwillingness to assert. Second, there is an excellent correspondence between the inferences we intuitively judge to be good and the inferences that preserve high assertibility.

The conditional probability account is well established for indicative conditionals but not for counterfactual conditionals. Adams (1975, 1976, 1993) investigated a line of research to explain counterfactual conditionals as a kind of indicative conditional, but found convincing counterexamples to his proposed account. However, subjunctive conditionals have been well modeled (Stalnaker, 1969; Lewis, 1973a) by giving truth conditions in terms of possible worlds. Lewis’s more general theory models them as variably strict modal operators whose semantics is given by a relation of comparative similarity. The Stalnaker-Lewis conditional and Adams conditional are difficult to assimilate in virtue of their significantly different theoretical underpinnings. For example, the assertibility conditions for indicatives are not directly related to their truth conditions if they even have any, and subjunctive conditionals, unlike indicatives, are thought to have assertibilities equal to the probability that they are true.

The Unified Approach

One striking piece of evidence for the unity of conditionals is the apparent synonymy of indicatives and subjunctives that are purely about the future.

(3) If Smith wins the award next year, he will retire,

and

(4) If Smith were to win the award next year, he would retire.

Examples like these are abundant, but not all sentences of those forms are. Edgington (1995) gives an example where we have convincing evidence, based perhaps on a snitch, that one prisoner, either Smith or Jones, will attempt an escape tonight and less convincing evidence, based on character assessment, that it will be Jones. Thinking that Smith is not daring enough, we disagree with

(5) If Jones were not to try to escape tonight, Smith would,

but knowing that one of them will try, we agree with

(6) If Jones doesn't try to escape tonight, Smith will.

Although (5) and (6) are distinct in meaning, they are exceptional cases. The widespread existence of pairs like (3) and (4) points toward the existence of some unified account of conditionals. A full explanation of the unity, of course, will need to explain the relevant difference between these two pairs.

Edgington offers additional circumstantial evidence for the unified theory: We are arguing about whether you will be ill if you eat this apple. You throw away the apple, and we continue to argue about whether you would have been ill if you had eaten the apple. Someone who leaves before seeing you throw it away argues later that if you ate the apple, then you were ill. Throughout, the subject matter appears to remain constant, and the dual approach needs to explain how the same evidence base can justify the same conclusions for conditionals that are evaluated by such different criteria.

To show how conditionals are unified, I will first identify a certain ambiguity in the counterfactual conditional, a weak reading and a strong reading, and then I will develop a semantic element called objective assertibility based on the weak reading. After justifying a theory of objective assertibility on some independent grounds, I will use it to explain the above cases.

The Strong and Weak Counterfactual Readings

The indicative and subjunctive conditionals are each ambiguous. This fact is widely recognized concerning the indicative. "If the barometric pressure drops, it will rain," can be understood as a kind of strict conditional, i.e. as implying some kind of necessary connection between the pressure and rain. On an alternative reading, the sentence expresses a conditional single case predication, made true if the pressure drops and it rains and false if the pressure drops and it doesn't rain. Regardless of whether this second disambiguation is modeled with the truth-functional ' \supset ' operator or with some logically stronger connective, it involves no notion of necessity.

An analogous ambiguity exists for conditional expressions in the subjunctive mood, although it has been unrecognized by philosophers. Specifically, there are two

different disambiguations worthy of examination for subjunctive conditionals like “If the pressure had dropped, it would have rained.” On what we may call the strong reading, this sentence is synonymous with “If the pressure had dropped, it must have rained.” The strong reading has been thoroughly investigated by philosophers. Older treatments (Goodman, 1947) took certain kinds of subjunctive conditionals, the counter-to-fact ones, to be a kind of elliptical argument. Counterfactuals of the form “If A were the case, then C would be the case,” are true on this account iff A together with some background assumptions usually including the laws of nature entail C . The semantics developed by Lewis (1973a) takes the truth of the counterfactual ‘ $A \Box \rightarrow C$ ’ to be given roughly by the truth values of C in all the closest A -worlds. Specifically, $A \Box \rightarrow C$ is true iff there is some $(A \& C)$ -world such that there are no $(A \& \sim C)$ -worlds at least as similar to actuality than it. In both Lewis’s and Goodman’s accounts, a kind of necessity is implied by the conditional: modifying the world to accommodate A necessarily makes C obtain as well.

On the weak reading, the subjunctive conditional is taken to express what *would* be true if A were true, not what would *have* to be true if A were true. The difference reveals itself through several pieces of circumstantial evidence. For one, some people make a special point of introducing necessity into reasoning about subjunctive conditionals in order to clarify the role of modality: One person says, “If she had missed the bus, she would have called.” Another rejects the conditional by saying “Not necessarily.” This denial amounts to a claim that her calling is not entailed by her missing the bus while still leaving open the possibility that she missed the bus and called. That is, it could still be true that she would have called; what is false is that it would *have* to be the case that she called. In addition, some people have pointed out that forward-looking counterfactuals, i.e. counterfactuals where the antecedent concerns a time previous to the time concerned by the consequent, are usually expressed in a simple counterfactual mode: If the cannon had fired, we would have heard it. However, backward-looking counterfactuals are usually expressed using a necessitation modality: If we had heard the cannon, it would have to have been fired.

The best evidence for the weak reading comes from its function in making probabilistic counterfactuals intelligible. We very often use counterfactual conditionals that refer to physical happenings whether explicitly or tacitly, and almost always, these conditionals involve physical probabilities. For example, if we are in an uninhabited warm kitchen with all the typical background conditions that obtain around kitchens, we should strongly agree with the counterfactual

(7) Had there been an ice cube left out in this kitchen, it would have melted.

Yet, if we are sophisticated in our knowledge of physics, we will know that there is an exceedingly remote possibility that the molecules that make up an ice cube will interact with the air molecules in the room in a way that will discharge heat into the surrounding environment and preserve the frozen ice cube for an arbitrarily long period of time. If one adopts the orthodox position that the truth conditions of counterfactuals are given by a relation of comparative similarity, and if one evaluates similarity in a straightforward way, one might reasonably think that worlds containing these bizarre evolutions of matter are as close as the many normal worlds where the ice melts. After all, there is nothing to distinguish the bizarre worlds outside of the fact that *later* they evolve into

states that have very unusual macroscopic behavior. It is possible to judge the bizarre worlds as less similar than the mundane worlds in virtue of what happens after the ice is left out, but judging similarity of worlds by what happens afterwards is surely a cart-before-horse. Assume determinism and consider a rare event that actually happened, like the outcome of a raffle. Then ask whether the same ticket would have won if the barrel containing the tickets had been spun several more times before the winning ticket was drawn. On any intuitive notion of post-raffle similarity, the same ticket would win, even though one should think that it most likely wouldn't win.

The result of adopting an intuitive notion of similarity where one does not distinguish states that are merely different microscopically, is that the very plausible (7) comes out false. There is certainly room for finding *some* relation of similarity to make it such that the ice will certainly melt in the closest worlds, but because this similarity relation would have to handle an enormous and very diverse set of physical phenomena that exhibit similar behavior, a different explanation of the truth conditions is far more plausible. The reason the counterfactual is so agreeable to us is not because it is strictly speaking true, but because either it is very probably true or it is a paraphrase of a slightly different conditional that is true: Had there been an ice cube left out in this kitchen, its melting would have been very probable. (I leave open whether these two possibilities are genuinely distinct.) What I hope to do in the following sections is to explain a way of understanding these kinds of probability claims in a reasonable way that will justify our everyday assessments of counterfactuals like (7). I do so by distinguishing David Lewis's and Robert Stalnaker's (1981) positions concerning 'might' counterfactuals. By taking different positions on the analysis of the 'might' conditional, Lewis ends up with an account that fits with what I have called the strong reading of the counterfactual, and Stalnaker gets an account that fits with the weak reading. I then extend Stalnaker's treatment of 'might' conditionals to more broadly treat counterfactuals that involve chances so that the advantages of the weak reading, and later, objective assertibility become clear.

David Lewis on 'might' counterfactuals

To express the counterfactual possibility of rainfall, we may say something like,

(8) If the pressure had dropped, it might have rained.

For Lewis, such 'might' counterfactuals are often ambiguous. The operator Lewis takes to represent a counterfactual possibility connective, ' $\diamond\rightarrow$ ' is defined explicitly as $(A \diamond\rightarrow C) =_{\text{Def}} \sim(A \square\rightarrow \sim C)$. On this disambiguation, (8) is equivalent to

(9) It isn't the case that if the pressure had dropped, it wouldn't have rained.

It is also acceptable at times to translate (8) as a 'would' counterfactual with a possibility operator acting solely on the consequent:

(10) If the pressure had dropped, it would have been the case that rain was possible.

On Lewis's account of their semantics, (9) and (10) are different. Simplifying slightly but harmlessly, (9) is true iff at least one of the closest pressure-drop worlds is rainy and (10) is true iff all the closest worlds had the possibility of rain. They differ in truth value

whenever all the closest worlds have their rain-possibilities unrealized. To summarize, ‘might’ counterfactuals are ambiguous between

$$(11) A \diamondrightarrow C \text{ (The not-would-not reading)}$$

and

$$(12) A \squarerightarrow \diamond C. \text{ (The would-be-possible reading)}$$

Robert Stalnaker on ‘would’ and ‘might’ counterfactuals

The semantics of the counterfactual ‘would’ conditional, according to Stalnaker, work as follows. There is a selection function that takes a proposition and a possible world together into a possible world. The intended interpretation of the selection function $f(A, \alpha)$ is that it picks out the one world most similar to α where A is true. The counterfactual $A \squarerightarrow C$ is true iff C is true in $f(A, \alpha)$. There are some natural constraints placed on the selection function so that it fulfills to some extent its intended role as a most-similar-world selector. As just defined, $f(A, \alpha)$ must have A true. If A is true in α , then $f(A, \alpha) = \alpha$ so as to ensure for each world that the most similar world is that world itself. In addition, f must not establish inconsistent orderings for each world α . Formally, this constraint is that for any α , B and B' , if B' is true in $f(B, \alpha)$ and B is true in $f(B', \alpha)$ then $f(B, \alpha) = f(B', \alpha)$. Stalnaker also includes in his theory an impossible world, λ , so that counterfactuals with impossible antecedents come out trivially true, but we can safely omit discussion of this special case.

Although the conditions just mentioned are sufficient to establish for each base world, α , a total ordering of all accessible worlds under a similarity relation, further restriction of f is left out of the semantics. Instead, pragmatic considerations determine which selection function is appropriate. Many factors become relevant to the determination of f as it is greatly context dependent, and often no single selection function will be picked out even by pragmatic features, so that an underdetermination of the selection function by contextual factors will exist. The extent of the underdetermination is often such that even with all material facts, there will still be no determinate selection function. This can happen in such a way that some selection functions will pick out a C -world and others a $\sim C$ -world. In such cases, the counterfactual $A \squarerightarrow C$ will be indeterminate. As to whether one should treat the counterfactual as having more than two truth values or allow for degrees of truth or truth-gaps, it is open. Stalnaker opts to treat the counterfactual classically, i.e., as having two truth values with no gaps, so that the indeterminacy is pragmatic and not a part of the semantics. If one has some other preferred treatment of vagueness, Stalnaker’s theory can easily be modified appropriately.

Stalnaker (1981) argues that might conditionals are primarily epistemic. They can usually be paraphrased as, “For all I know, it is possible that: if A were true, then C would be true.” Formally, $A \diamond \rightarrow C$ for him is equivalent to $\diamond(A \Box \rightarrow C)$. This formulation is successful at accounting for the significant number of counterfactuals where the possibility is epistemic, but where the possibility is explicitly formulated as a physical possibility, problems arise. Expanding on the rainfall example, assume that rainfall is ruled out altogether in the nearby worlds where the pressure drops due to some variables of which we are ignorant. On the epistemic treatment,

(13) If the pressure had dropped, it might have rained,

should be understood as

(14) For all I know, it is possible that if the pressure had dropped, there would have been rain.

Yet, (13) is false and (14) true. In response to similar examples, Stalnaker advances a quasi-epistemic reading of the possibility. The possibility operator expresses the possibility that exists “after all the facts are in.” (1981, p. 101) The possibility in these cases exists due to indeterminacy as to which selection function is appropriate for the evaluation of the counterfactual. In this case, the unknown determining facts exist in all the worlds a relevant selection function would pick out, so the quasi-epistemic possibility of rain does not exist.

Advantages of Stalnaker’s account over Lewis’s

With such an account of ‘might’ counterfactuals, Stalnaker is in a position to explain a linguistic fact which Lewis’s account cannot. Stalnaker notes that it does not sound right to accept

(15) If A had obtained, C might have,

while also accepting

(16) It’s not true that if A had obtained, C would have.

Lewis has trouble explaining this opposition because accepting (16) on the not-would-not reading is equivalent to accepting

(17) If A had obtained, $\sim C$ might have.

Certainly, (15) and (17) are compatible. On the would-be-possible reading as well, (15) is seen to be compatible with (16). Suppose that in some of the closest A -worlds, the possibility of C is not actualized. Then (15) and (16) are both true, contrary to the observed tension. On both readings of the might counterfactual, the logic itself fails to account for the conflict between (15) and (16).

Stalnaker's own assessment of the conflict is that (15) and (16) are not inconsistent but that they form a conjunction rather like Moore's paradox: "*p* and I don't know that *p*." In both cases, the speaker asserts something, thereby presenting himself as knowing it, but then explicitly denies knowing it. One can expand on Stalnaker's explanation of the conflict by examining the relevant implicatures of (15) and (16). Understanding possibility as Stalnaker suggests, (15) is equivalent to

(18) For all one could know, it's possible that if *A* had obtained, *C* would have.

Given Stalnaker's semantics, (18) logically implies the quasi-epistemic possibility of: if *A* had obtained, *C* would have. Because (16) is an unqualified *denial* that if *A* had obtained, *C* would have, (16) seems to rule out any such possibility. Typical contexts will have (16) implying conversationally that there is no possibility that if *A* had obtained, *C* would have. This conversational implicature can be voided though, if one points out that by asserting (16) one means to deny that *necessarily*, if *A* had obtained, *C* would have, while leaving open the possibility expressed by (15). Stalnaker's explanation is good here, but it would be better if we had an explanation of why the conversational implicature exists at all in this case.

The distinction between the strong and weak reading helps to make clear why we normally take (15) and (16) to be incompatible. What makes (16) seem to imply the impossibility of "If *A* had obtained, *C* would have," is that we vacillate in our interpretation of (16) between the strong and weak readings. If we interpreted it always in the strong way, it would be clear that what is being denied is the necessity of *C* given *A*. Yet, we are tempted to think of the counterfactual in the weak way, as asserting what *would* have been the case (if *A* had been the case) as opposed to what *must* have been the case. By denying that *C* would have been the case, (16) implies that *C* would not have been the case. After making this inference, which relies on the weak interpretation, we slip back into the strong reading, and take (16) to mean "If *A* had been the case, then $\sim C$ would have to be the case," which clearly contradicts (15).

It is apparent from the semantics presented by Stalnaker and Lewis that Stalnaker's semantics model the weak reading well, and Lewis's the strong. On Stalnaker's semantics, we pick a single most similar world and let the value of *C* at that world determine the truth value of $A \Box \rightarrow C$ at our world. This follows what the weak reading proposes to tell us, that *C* would have obtained had *A* obtained. On Lewis's semantics, ' $\Box \rightarrow$ ' is a variably strict modal operator. For $A \Box \rightarrow C$ to be true, it must be that *C* holds in all the closest *A*-worlds.* This matches the implication in the strong reading that *C* must hold if *A* were to hold.

An extension of Stalnaker's position to treat chancy counterfactuals

It is a credit to Stalnaker's logic that it is able to give a better treatment of chancy counterfactuals. Chancy counterfactuals are 'would'-conditionals where the consequent takes the form of a chance claim about some event, as in

(19) Had I rolled the die, the chance of a 3 coming up would have been 1/6.

* Here I am simplifying harmlessly by ignoring cases where there are no closest worlds.

As noted previously, Stalnaker advocates understanding the possibility operator as a quasi-epistemic operator of wide scope. The obvious extension is to model chancy counterfactuals in terms of a quasi-epistemic *probability* operator. Following this strategy, the above sentence is equivalent to

(20) Given all that can be known, it is $1/6$ likely that had I rolled the die, 3 would have come up.

We are in a position now to see how to make sense of such a claim. Pre-theoretically, when we consider what would have happened had I rolled the die, we leave open which outcome would have eventuated, and we do so even if we know every fact about the actual world. (If you think that there are such things as conditional facts that determine the truth values of counterfactuals, you will disagree. In that case, take the quasi-epistemic operator not to quantify over such facts.) If the context involves a typical chance setup, we don't assume that there exists one particular outcome such that had I rolled the die, that outcome definitely would have resulted. In Stalnaker's framework, this is to say that there is a range of selection functions, divisible into 6 classes of die-outcomes, consistent with the hypothetical assumption that I rolled the die. What's more, we do assume that had I rolled the die, some or other outcome would have resulted (without assuming that there is definitely one outcome such that it would have resulted). Because there are no facts to distinguish among the various possible die-casting worlds, which one would have been actualized, the reasonable way to handle such indefiniteness is to treat it as we do ignorance about actual chancy processes: assign a uniform probability distribution over the outcomes. Consistent with everything we can know, there are many different worlds that can be picked out as the closest world, $1/6$ of which are worlds where 3 turns up as the outcome of the roll.

The appropriate probability distribution to associate with a given antecedent, where it exists, is given by the context. In context of sufficiently idealized games of chance, the appropriate probability is the obvious one given by the chance setup. In realistic situations, the chance is the chance given by the physical laws, i.e. the objective chances. There are two important and different kinds of objective chance worth distinguishing. The first, which I call *objective chance* proper, is the kind of chance that exists in indeterministic worlds where the fundamental laws plus obtaining conditions entail some probability distribution over future outcomes. The second, which I call *quasi-objective chance*, is the chance that exists when the macroscopic parameters defining some object together with the physical laws, imply some probability distribution over possible outcomes. An example of a quasi-objective chance is the chance associated with statistical mechanical probabilities in a deterministic world. Both kinds of chance are objective in the sense that there are matters of fact about what the right probability distribution is. Genuine objective chances differ from quasi-objective chances in deterministic worlds because the genuine chances can entail non-trivial probabilities given a full microscopic specification of the physical system at some previous time, but the statistical mechanical probability is not invariant when conditionalizing on certain kinds of information about the microscopic state. It gives the wrong probability distribution for predicting outcomes when special information about the microscopic variables is known.

Truth for a chancy counterfactual amounts to an equality between the degree of chance asserted to exist and the probability assigned by the context, which is often the

objective or quasi-objective chance. For a determinate assignment of truth value to be made in realistic cases, the context must be one that is sufficiently resolved to the point where it makes sense to say that there is a chance. Vague counterfactuals like

(21) If you had drunk fluoridated water your whole life, you would have had a 4% higher chance of incurring brain damage,

are troublesome because there are a multitude of ways you could have drunk fluoridated water and no good way to summarize them all. If the laws of our world are deterministic, there is no obvious quasi-objective probability distribution that one can say is the right one picked out by the context, and if even they are indeterministic, it isn't clear which worlds to put a probability distribution over. These problems are due not to any deficiency in the semantics of counterfactuals but to the vagueness inherent in the content of the claim. The way to remedy this trouble is to consider all the various classes of ways the antecedent could be fulfilled for which it does make sense to put an objective or quasi-objective probability distribution on the members. Then, there will be a truth value for each partial resolution. One might hope that the laws and circumstances are such that there is some consistency in the probabilities assigned to getting cancer over a wide range of partial resolutions, so that it makes sense to speak of *the* chance of getting cancer, but if not, the problem is with this particular nebulous counterfactual and not the logic or semantics. In practice, the fine details of the semantics usually do not stand in our way. We usually reason about such counterfactuals with the implicit assumption that we should use statistical frequencies in the actual world to tell us about chances in the counterfactual worlds. The use of such vague counterfactuals borrows its logical structure from contexts where the probabilities are quite clear, so that we may discuss important subjects that would otherwise be difficult to verbalize.

For counterfactuals generally, we should say that $A \Box \rightarrow C$ on the weak reading is true iff the closest A -world is a C -world. If it is unclear what the closest world is, but it does make sense to attribute a physical probability distribution over the relevant A -worlds, then the physical probability gives the likelihood that $A \Box \rightarrow C$ is true. We can express this more formally as follows. Given a base world α and some antecedent A that is entertainable at α , the probability function P assigns an objective or quasi-objective probability for any outcome. Then, $P_{\alpha,A}(C)$ is the probability that $A \Box \rightarrow C$. $P_{\alpha,A}(C)$ is similar to a conditional probability—it is the probability that C would have obtained, had A obtained—but it is not really a conditional probability. For example, our world might be a world where C is impossible. In that case conditionalizing on A would not raise the probability of C , but for some such C , A 's being true makes C possible. For the strong reading, $A \Box \rightarrow C$ is true iff all the relevant A -worlds are C -worlds, so the strong reading is just the weak reading with a necessitation operator, just as a naïve interpretation of natural language would have it.

It is a virtue of looking at the counterfactual in terms of the function $P_{\alpha,A}(C)$ that non-classical semantic theories can be accommodated. Those who believe in degrees of truth or non-bivalent truth values can hold that the counterfactual $A \Box \rightarrow C$ is true in α iff all selection functions $f(A, \alpha)$ pick out C -worlds and false iff all of them pick out $\sim C$ -worlds. In the intermediate cases, they can say either that $A \Box \rightarrow C$ has a degree of truth that matches the relevant probability or that $A \Box \rightarrow C$ has no truth value and take the

probability-value to be an additional semantic primitive, an idea suggested in Edgington (1995). The benefit of holding to a classic bivalent logic is that it permits us to say simply that the probability of C , given the counterfactual truth of A , is the probability of $A \Box \rightarrow C$ whereas on other accounts of the semantics, the connection between the truth of counterfactuals and $P_{\alpha,A}(C)$ is less clear.

The struggle of Lewis's position to treat chancy counterfactuals adequately

Lewis models chancy counterfactuals as straightforward 'would' conditionals. For example, (19) is modeled as

(22) I rolled the die $\Box \rightarrow \text{chance}(3 \text{ came up}) = 1/6$.

There are two types of problem for this model of chancy counterfactuals. First, it lacks the resources to explain why (22) seems incompatible with

(23) I rolled the die $\Box \rightarrow \sim(3 \text{ came up})$.

Second, it does not handle quasi-objective chances, and thus fails to adequately account for counterfactuals like "If the ice had been left out, it would have melted," in a deterministic context.

Counterfactuals (22) and (23) intuitively have some degree of incompatibility on the strong reading of the counterfactual that Lewis accepts. One idea (from Peter Menzies) is that they must be incompatible if one holds certain other principles that Lewis advances, specifically the Principle of Recombination (Lewis, 1986a): anything can coexist with anything else if they occupy distinct regions of space-time. The intuition behind this is that no more and no less of a miracle is required to fulfill the antecedent and fulfill the chance than to fulfill the antecedent and not fulfill the chance. It would follow that the 3-world and ~ 3 -world are equally close, making (23) false when (22) is true. The intuition behind this argument however, does not hold in general.

In fact, Lewis presents two sound arguments for the compatibility of (22) and (23). One of his arguments is a *reductio* on their incompatibility:

Let C be any proposition that might obtain as a matter of chance; let u and v be a C -world and $\sim C$ -world, respectively, but let them both be worlds that have a chance of going either way; let A be the proposition that holds at these two worlds and no others; and let w be any third world. It is true at w that if A , there would be some chance that C ; so by the supposed incompatibility, it is false at w that if A , it would not be that $\sim C$; so u must be at least as close to w as v is. Likewise, putting $\sim C$ in place of C , v must be as close to w as u is. That is, worlds u and v are tied in closeness to any third world. But u and v are any two worlds that differ in respect of the outcome of a matter of chance—no matter how much they differ in other respects as well! This completes the *reductio*. (1979, 1986a, p. 65)

The other relies only on an innocuous assumption that it makes sense to talk about chances in the antecedent.

What would be the case if there were some unfulfilled chance of C ? If so, then there would be some chance that C . But if so, then also it would not be that C . (1986, p. 65)

Moreover, I have a third argument that works where the chance is mentioned only in the consequent, so long as the antecedent is true: Assume A is true, and that there was an unfulfilled chance of C . Then, $A \Box \rightarrow \text{chance}(C) > 0$, but also $A \Box \rightarrow \sim C$.

While Lewis is correct that in general, sentences like (22) and (23) are not incompatible, it remains that in many normal contexts they are manifestly incompatible. The failing of Lewis's treatment of chances is not that it is inconsistent with some independent principle, but that it lacks any explanation of why the two seem incompatible.

The other advantage to treating chances with the weak reading is that it treats both objective and quasi-objective chances equally. Remember that we have been seeking some structure that would allow us to say of counterfactuals like

(24) Had there been an ice cube left out in this kitchen, it would have melted,

that they are very probably true or have a high degree of truth or something like that. It is manifest that in reasoning about such counterfactual possibilities, it is unimportant whether the laws are deterministic or indeterministic. We know that there is some probability that the cube will not melt even if the laws are deterministic. The only difference the laws make will be in which probability measure is appropriate, but for almost any reasonable antecedent, the differences are small. If we try to treat chancy counterfactuals as Lewis suggests, we run into problems when the world is deterministic. In such cases, for any world where the ice cube falls out, the chance is either 0 or 1 that it melts. So, we cannot say that the chance of it melting is very high but not 1, which is what we should want to say.

The problem might be resolved by modifying Lewis's account of chances to allow quasi-objective chances, but this would require significant tinkering with the Principal Principle. Quasi-objective chances do not remain constant under all admissible evidence because information about the past microstate is admissible, and this information will change a rational agent's beliefs so that they do not match the statistical mechanical probabilities. To the credit of this strategy, there is some independent motivation for allowing quasi-objective chances for those like Lewis who defend a Best-Systems-Analysis of the laws of nature. One of our actual laws may be a probabilistic constraint on microstates of the kind present in statistical mechanics, for it may provide the best combination of informativeness and simplicity, and these quasi-objective probabilities might best be understood as chances.

Objective assertibility

Stalnaker's treatment of counterfactual possibility can be extended in such a way that the probability of the truth of $A \Box \rightarrow C$ at world α is the probability of C over all the relevant worlds where A obtains. If one accepts that assertibility for counterfactuals goes by probability of truth, as it does for most any other kind of proposition, then the assertibility of the counterfactual is the probability of C given A , namely $P_{\alpha,A}(C)$. One should think of $P_{\alpha,A}(C)$ as measuring the *objective assertibility* of the counterfactual. Objective assertibility is the degree to which a person should (ideally) be willing to assert a given counterfactual (again, setting aside meddling contextual factors). The sense of idealization involved in the concept of objective assertibility is exactly the sense present in Stalnaker's quasi-epistemic interpretation of 'might' quantifiers. That is, the

objective assertibility of a counterfactual is the degree to which a person with a complete knowledge of history should be willing to assert the counterfactual.

Evidence for the objective assertibility of $A \Box \rightarrow C$ being $P_{\alpha,A}(C)$ comes from assessing particular examples. How willing are we to believe or assert that if the die had been rolled, it would have turned up 3? It should be $1/6$ if we read the counterfactual in the weak sense. Why do we agree with “The ice cube would have melted, had it been left out”? We think it highly likely that among all the ways of leaving the ice out, that the ice cube melts. Additional benefits of generality arise from thinking of counterfactuals as possessing an objective assertibility. People who accept a theory of vagueness where vague propositions have no truth values and people who do not accept that conditionals have truth values cannot think of $P_{\alpha,A}(C)$ as the likelihood that $A \Box \rightarrow C$ is true, yet that person can still make sense of counterfactual chanciness by taking $P_{\alpha,A}(C)$ to be an independent element of the semantics.

The particular way counterfactuals are assigned objective assertibility is critical to understanding the unity of counterfactuals. The way to judge assertibility is to take the actual state of the world at some time, t_A , to which the antecedent A pertains and to modify it so that A is true. For any ordinary counterfactual where A specifies some macroscopic events or processes, the result of this modification is a state defined macroscopically, what we can call the A -state. For these ordinary antecedents, there are many microscopic states compatible with the A -state. To determine the probability of the consequent, C , one should consider the objective or quasi-objective probability distribution over the compatible microstates and apply this probability distribution over the corresponding worlds, i.e. the histories that are determined by the dynamical evolution of these microstates. The objective assertibility $P_{\alpha,A}(C)$ is the probability of C among these worlds.

It is critical that the objective probability distribution respect the context of the assertion of the counterfactual. As mentioned previously, some contexts are not sufficiently restricted to where a clearly correct probability distribution is available so that one can only evaluate the counterfactual with respect to instances of the counterfactual where the vagueness is better resolved. One crucial principle that serves tacitly to resolve vagueness is that one considers only worlds whose histories closely, if not perfectly, resemble the actual world’s past history for most of the time before t_A . At some time not too long before t_A , the evolution of the counterfactual worlds diverges from the actual history. This could occur through a lawful indeterministic dynamical evolution or by a deterministic evolution from states whose differences from the actual world are so slight as to be previously insignificant. The divergences in these worlds lead, by the fundamental dynamics, to the obtaining of A . The principle that the relevant worlds are entirely or are at least overwhelmingly worlds that evolve into A -worlds by probable processes, what we can call the mundaneness principle, is essential to the evaluation of counterfactuals. Without it or some deeper principle that justifies it, there is no reason to believe that the microscopic states compatible with the A -state have a history that resembles actuality. And such a similar history is a tacit component of the assumptions people typically hold when imagining how the world would be if it were different with regard to some particular matter of fact.

The Unity of Conditionals

Explaining the apparent unity of conditionals is a matter of explaining why the subjunctive and indicative agree where they do. The synonymy of the sentence pairs (3) and (4) can be understood as follows. The objective assertibility for (4) is equal to the likelihood that the closest world where Smith wins the award is also a world where he retires. We evaluate this probability by starting with the current state of the world and imagining all the ways that the world could lawfully evolve into Smith's winning next year, holding fixed all known macroscopic facts about the present and past. Our estimate of the objective assertibility of (4) is our estimate of the proportion of these ways that continue to evolve into Smith's retirement. Thus, our subjective probability for (4) is equal to the subjective conditional probability of Smith's retirement, given everything we know and given Smith's future winning. In the circumstances where Smith's winning does not conflict with anything we know about the present, this conditional probability is equal to the assertibility of (3).

In the example involving Booth and Lincoln, the assertibilities do not match. Our estimate of the probability that someone would have killed Lincoln, given that Booth didn't is determined by our imagining history being mostly similar up until some time before Booth attempted the assassination. We imagine that some cause leads Booth not to kill Lincoln—perhaps he doesn't summon the courage, perhaps his gun fails, or perhaps he is arrested by government agents—disregarding all the contingent facts of history from that time on. Then, we try to estimate the objective likelihood, given some such circumstance, that someone else kills Lincoln, and this is our estimate of $P_{\alpha,A}(C)$. This will not at all match our assertibility for (1) because that assertibility is determined by holding fixed as much of what we know as we can while still entertaining the possibility that Booth didn't kill Lincoln. Because this includes all the historical facts we know, e.g. that Lincoln was shot, there is no reason to think the two assertibilities should match.

The interesting cases are borderline cases like Edgington's example involving the prisoners Jones and Smith. To determine the assertibility of the counterfactual, we not only try to picture how the world would evolve if Jones were not trying to escape, but we picture how the world would have to be like beforehand so that he will not try to escape. We don't hold fixed all the known facts of the current and past state of the world and examine the many ways the world could evolve into Jones not trying the escape. We imagine that his not escaping comes from some small difference that is possibly in our recent past. We allow that whatever differences constitute the history where Jones doesn't escape are quite possibly differences that evolve into our never receiving the snitch's news. So the fact that antecedents regarding the future sometimes imply counterfactual differences in the present and past, namely the kinds of differences that give rise to the antecedent, allows there to be a difference in the facts that we hold fixed when judging the probability of C given A .

There are two ingredients that combine to give this successful explanation of the circumstances in which indicatives and subjunctives match. First, the concept of objective assertibility provides the needed conceptual middle ground between the possible worlds logic of subjunctives and the subjective assertibility logic of indicatives. Second, there is the principle of mundaneness, which restricts the context of counterfactuals to worlds where the antecedent obtains as a result of a lawful, dynamical

evolution of the physical world from states that closely match our actual history. This principle accounts for the fact that some subjunctives concerning the future can entail differences about the past, which in turn explains the odd cases where future subjunctives do not match future indicatives.

CHAPTER VI

ENTROPY AND COUNTERFACTUAL ASYMMETRY

Were the present somewhat different in matters of fact, the past would have been almost the same as it actually was, but the future could be significantly different. This principle, the counterfactual asymmetry, plays a significant role in ordinary discourse and to a significant degree in scientific discourse. Yet, its connection to theories of counterfactuals and to physical asymmetries is still unclear. An idea that has come under consideration recently is that the counterfactual asymmetry can be explained by an asymmetry in the universe's entropy (Albert, 2000). In what follows, I present what I take to be the best theory of counterfactuals that makes the connection between entropy and counterfactuals explicit. While it justifies many features present in our intuitive and scientific assessments of counterfactuals and is generally a good theory of counterfactuals, it fails in the end, to count as a fully successful explanation of counterfactual asymmetry.

It is doubly remarkable that one can find anything approaching a good theory to explain counterfactual assertions by way of a thermodynamic property because (1) *prima facie*, thermodynamics and counterfactuals have very little in common, and (2) in the standard way of understanding them, they have such different theoretical underpinnings. Counterfactuals, if we stick to the orthodox view, are propositions obeying a logic whose semantics is given in terms of a comparative similarity relation among possible worlds. Determining the truth-values of counterfactual conditionals is a matter of finding the most similar worlds where the antecedent holds and evaluating the truth of the consequent in these worlds. Yet classically, entropy is defined in terms of mechanical work and temperature in the context of steam engines and similar mechanical systems. To make a connection between the two requires some theory.

On the thermodynamic side, there are familiar stories about how entropy is to be understood as statistical feature of physical systems. The most plausible story associates the entropy of a system with the volume in phase space occupied by all those microstates that are macroscopically indistinguishable from the actual system. This understanding of entropy has the virtues that it applies to individual, real systems, in contrast to ensemble approaches and that it fits into a reasonable explanation of entropy increase. The primary benefit of adopting this statistical mechanical understanding of entropy for the purposes here is that one can associate with any possible situation that is sufficiently well defined, an objective probability distribution for all future situations. This allows one to make objective claims about how physically possible states of affairs would likely evolve and thereby to evaluate the relative likelihood of various counterfactual possibilities. These statistical mechanical probabilities are objective in the sense that scientific experiments have confirmed the empirical adequacy of the statistical mechanical probability distributions, which have no observer dependence even though they are not necessarily objective in the sense of being invariant under conditionalization on certain admissible facts, e.g. microscopic facts about the past.

To connect this conception of entropy with counterfactuals, one needs to adopt some theory in which counterfactual assertions are associated with probabilities. Ernest Adams (1975) attempted to associate counterfactuals with subjective probabilities, and Brian Skyrms (1994), in resolving one of Adam's puzzles, made the association with objective propensities. These perspectives are different from the standard account by Lewis (1973a, 1973c; see also Stalnaker, 1968) of counterfactuals in terms of a relation of comparative similarity, but they are also consistent with the Stalnaker-Lewis models. To hold them consistently, one needs to accept that the similarity relations are constrained by the objective probability relations. While this acceptance *per se* is unproblematic, the probability constraints have nothing to do with any intuitive measure of similarity, which would be a problem for anyone who hopes that the similarity metric governing the logic of counterfactuals is a function of relations that each count as intuitive respects of similarity. The Adams and Skyrms treatments of counterfactuals associate with each counterfactual some assertibility. The degree to which one should believe a counterfactual is the degree to which the consequent would be true if the antecedent were to hold. The notion of assertibility usually connected with conditional statements is a subjective conditional probability. For example, Jim's assertibility for "If A, then C," namely the degree to which Jim in fact finds that statement assertible, is $\text{Pr}_{\text{Jim}}(C/A)$. One can also understand an objective notion of probability, so that the objective assertibility of "If A were the case, then C would be the case," being the degree to which one *should* find that statement assertible, is $\text{Pr}_{\text{Obj}}(C/A)$. In what follows, the assertibility of a counterfactual is not the usual subjective variety but is associated with the objective conditional probability defined by stochastic laws and statistical mechanical probability distributions.

The interesting consequences of this theoretical connection between statistical mechanics and counterfactual conditionals are apparent only when one examines the detailed way in which counterfactual statements are interpreted consistent with some antecedently accepted dynamical laws. After developing some of the tacit assumptions implicit in the evaluation of counterfactual statements, I will provide an evaluation procedure for determining $\text{Pr}_{\text{Obj}}(C/A)$.

Straightforward Cases

There are some kinds of counterfactual claims the evaluation of which is relatively unproblematic. Those counterfactuals involving antecedents and consequents that are easily translatable into the language of fundamental physics, often have their assertibility determined by the laws of nature. A counterfactual of the form "Had the full physical state at time t been ξ , then C would have been the case," is evaluated by taking the state ξ and letting the evolution of ξ under the fundamental laws indicate what the states are at all other times. If the laws are deterministic, then it follows that the counterfactual has assertibility one if the consequent is true in this evolution and zero otherwise.

If the laws are indeterministic, the picture is complicated somewhat, and the particular variety of indeterminism at work becomes relevant. If the laws are indeterministic but still stochastic, the laws give a probability measure over many worlds that share the modified physical state at t . In such cases, the assertibility of the counterfactual is just the physical probability of the consequent, given the modified

physical state. For forward-looking counterfactuals, counterfactuals where the antecedent pertains to happenings before those of the consequent, the state ξ is itself often sufficient to determine the probabilities for all future C 's. Stochastic theories that have been advanced as credible theories of nature determine forward transition probabilities: they assign to any possible state at one time, probabilities for possible later states. The stochastic laws, together with ξ , determine the probability for all future states, which determines a probability for C . This probability is the assertibility of the counterfactual. However, no serious, known theory gives backward transition probabilities, probabilities for past states, given the current state. Hence, backward-looking counterfactuals will need some other theoretical treatment.

This rather simple recipe, while sufficient for determining these counterfactuals, needs significant modification to be general enough to apply to the kinds of counterfactuals people typically use. The major source of trouble is the difficulty posed by trying to extend the simple method to counterfactuals where the modified physical state is underdetermined by the antecedent. For any antecedent that under-describes the total physical state, there are usually an infinite number of total physical states where the antecedent obtains with no objective probability measure over them. As a result, the assertibility is not immediately determinable.

In order to make some determination of the assertibility, the vagueness of counterfactual antecedents needs to be better resolved. One needs some way of characterizing the background assumptions inherent in counterfactual evaluation that will clear up what one means when one postulates a certain counterfactual situation. Nelson Goodman (1947) famously struggled with this problem to no avail. He had hoped to find a principled method of determining which actual facts are properly background conditions without begging the question as to which counterfactuals were true. Although his project is very unlikely to succeed, it *is* possible to find a cohesive set of tacit background conditions that people should maintain when evaluating counterfactuals without justifying these conditions in terms independent of our judgment of counterfactuals. The set of background conditions can be justified in some weaker sense just in virtue of its being a simple, cohesive collection of principles that generate truth conditions or assertibility conditions for counterfactuals that match refined assessments of counterfactuals. The goal in what follows is to find some general principles that people should hold fixed when evaluating mundane counterfactuals, i.e., ordinary counterfactuals involving physical objects, processes, and events. These principles are not fully general, and their applicability is context dependent. Yet, they hold in a wide range of typical cases just because there is some sizable and recognizable overlap among most of the contexts where people use counterfactuals.

One principle that seems reasonable at first is that antecedents that describe a situation without reference to microscopic detail should typically avoid distinguishing some specific microscopic instance as being special. When considering what would have happened had the match been struck, one does not consider one microscopic configuration of matter corresponding to the match and its environment. In most cases, as one actually thinks about such matters, one places no attention at all on the possible underlying microphysics, and a moment of thought should lead one to recognize that there is nothing in the *meaning* of the antecedent, nor in the implicit meaning of the counterfactual conditional itself that picks out some microstate as being the microstate

that would have obtained had the antecedent been true. Rather, one should consider a range of possibilities of how the match might have been struck.

For example, consider a box containing a gas with a very small perforation, a circular hole in the box with a radius larger than the size of the gas molecules. A device is set to measure whether a gas molecule escapes during a short, specified period of time Δt . Consider the counterfactual, “Had the gas been one degree hotter, a gas molecule would have escaped during Δt .” For the gas to have been hotter, the molecules must have been in a more energetic microscopic configuration. Lacking any overriding reason to distinguish such configurations, we consider all the configurations consistent with the higher temperature constraint. Given these microstates and the statistical mechanical probability measure over them, there is a defined value for the likelihood that a gas molecule escapes, and the assertibility of the counterfactual is equal to this value.

There are still many situations where the antecedent or context is too vague for the application of such probability measures: “If people ate more vegetables, cancer wouldn’t be as big of a problem.” There is no objective probability measure over all the ways people could eat more vegetables, so the best we can do is to evaluate the vague antecedent by evaluating its less vague precisifications. We can assign a probability measure over some significantly precise way in which people eat more vegetables, ways that are described not in microscopic detail, but are described with many general macroscopic constraints. In some cases, it will turn out that almost all such precisifications will assign almost the same probability, so that one can reasonably assign an assertibility value to the vague counterfactual. In other cases, while no assertibility is appropriate for the counterfactual, it is easy to understand the failure to have such a value. It fails to have an assertibility value just as ordinary vague predicates fail to have determinate truth values when applied to borderline cases. If the antecedent is too vague to admit of an objective probability measure, then for evaluation it needs to be broken down into sub-cases where such a probability does exist.

The Locality Principle

Another principle for resolving vagueness is that, except where antecedent implies otherwise, the counterfactual situation should match the actual world. Call this the locality principle for counterfactuals. For an example, let r be some spatial region and ξ_r be a precise (microscopic) physical state of region r , and consider a counterfactual of the form, “Had the physical state in r at time t been ξ_r , then C would have been the case.” To evaluate the counterfactual, it seems, one should take the entire actual microscopic state at time t and modify the details in the region r so as to make the localized state of affairs, ξ_r . Because the physical state outside r was not mentioned in the antecedent, one should not, without some overriding reason, modify the physical state outside r . Clearly, there is an informal implicature that the situation to be considered is one where the actual world has been modified in region r *and nowhere else*. The fact that r was mentioned as opposed to some other region gives some reason, albeit a defeasible one, to think that in the hypothetical situation being considered, areas outside r are no different from the way they actually are.

As with any informal implicature, the strength varies by context, so this principle should not be taken strictly, even with counterfactuals that are physically quite precise. Evaluating a conditional starting with, “If the Earth had a denser atmosphere...” one

should start with the actual state of the universe and construct a state consistent with the antecedent, which we can call the *A*-state. The *A*-state is created by replacing the relevant region, a volume occupied by the Earth's atmosphere with a physical state that has more nitrogen and ozone molecules, etc. Then, one lets the fundamental physical laws acting on the *A*-state imply everything else about the counterfactual situation, including the probability of *C*. Taking the locality principle strictly means not fiddling with the state outside the atmospheric region, but there are contexts where this may not seem correct. After all, atmospheric particles don't come from nowhere. One may reasonably think that one needs to subtract out some atoms from thin gas far away from Earth in order to build up the denser Earth atmosphere. Just on the grounds that keeping the physical laws the same in the counterfactual world may require conserving certain physical quantities like mass-energy, it may be necessary to balance changes in the region with some changes outside the region. Regardless of whether this is necessary, the mere plausibility of a context where such balancing needs to be performed is enough to demonstrate the non-universality of the locality principle.

Despite its defeasibility, the locality principle, as a rough guide to the evaluation of ordinary counterfactuals, is implicated in both everyday usage of counterfactuals as well as more formal usage. Some evidence for the locality principle, in addition to its intuitive appeal, comes from its role in addressing quasi-chancy phenomena, chanciness that arises in deterministic systems from objective probabilities of the statistical mechanical sort, and becomes manifest in worlds with a deterministic and local dynamics. In order to explore this possibility, one can consider how counterfactuals should be evaluated in the context of the General Theory of Relativity (GTR). Take a world where GTR is a law and where some brief distant event *e* is occurring, and consider some counterfactual possibility, *A*, that is only about some local state of affairs a very short time ago (see Figure 1).

We want to explore in what circumstances and to what degree one should believe the proposition "Had *A* been true, *e* would not have happened." As illustrated in figure 1, the antecedent *A* involves only modifying the recent physics so that the region corresponding to *A* is not in the backward light cone of *e*. If we were treating the situation classically, we would modify the actual state at the time to which *A* pertains. Because in general, there is no global temporal structure present in GTR worlds, the implicit time references in the counterfactual need to be interpreted relativistically. In GTR, one has something similar to instantaneous states, namely space-like hypersurfaces of the full space-time. Fortunately, the dynamics of GTR are such that we can select any hypersurface as the 'time to which *A* pertains' with the result that the dynamical consequences everywhere will be the same. In particular, with *e* outside the light-cone of *A*, any hypersurface cutting through the happenings relevant to *A* is such that *e*'s existence will be entailed. Thus, if *A* had happened, *e* still would have occurred, which is exactly what one would intuitively want to say since *e* is too far away for the nearby events to have any dynamical influence.

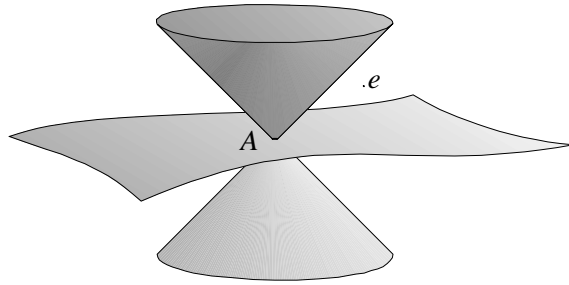


Figure 1

In figure 2, e is presented as far away but occurring far enough in the future of A that some influence from electromagnetic radiation associated with A 's being true, can have dynamical influence on the physical evolution in the neighborhood of e . In such a case, the particular details of e and A are relevant to the determination of the counterfactual's assertibility. If e is the kind of event whose existence was improbable a minute ago in the past, but whose existence was probable given the state of the world a few seconds ago, then intuitively we should find the counterfactual highly assertible. An example of such an e is the announcement of some lottery winner a minute after the lottery drawing. Before the drawing, the announcement of X as the winner was improbable, but after the drawing, it was highly likely that the lottery spokesperson could successfully read the name from the winning ticket. In such a case, whatever radiation interference might have come from A 's having been the case is unlikely, if not impossible, to have disturbed the stable post-lottery announcement process. This intuitive result again matches the process that one uses to evaluate counterfactuals by considering all the various space-like hypersurfaces corresponding to all the microscopic ways that A could be instantiated, calculating their deterministic evolutions into the future, and having the announcement of X being entailed in all or almost all of these worlds.

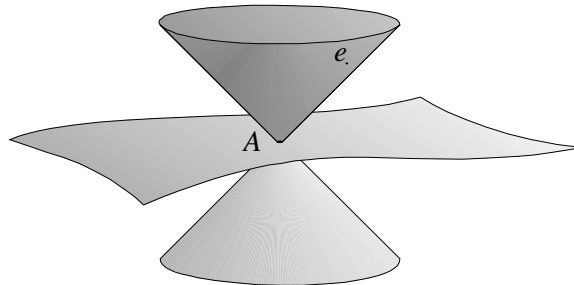


Figure 2

If the process by which e occurs involves a process that is unstable in the very recent past of e , the probability of e 's occurrence given A can be significantly affected. If most hypersurfaces that instantiate A evolve into full histories where e does not occur, then the assertibility of the counterfactual is low. This again agrees with our intuitive assessment of counterfactuals where we recognize that if the physical influences have enough time to reach the neighborhood of e and the physical influences are of the kind that in part determines whether e occurs, then e 's existence is no longer necessarily likely. In fact, some kinds of e 's, like the outcome of a coin toss, are such that the right

kind of A can entail that the probability of e is equal to the subjective likelihood that we typically have of events like e . One example regards a bet on a fair coin toss. I bet heads, but it came up tails. I might want to say, “If I had bet tails, I would have won.” On the method of evaluating counterfactuals this comes out highly assertible if my having said “Tails” would have had no significant dynamical influence on the coin’s trajectory, as would likely have happened if I had spoken as the coin was nearly finished falling. The assertibility would come out around 0.5 if I had said “Tails” a few seconds before the toss and you are the kind of person who modifies the strength of a coin toss based on sensitive details of your mental or physical state, e.g. on what you have heard recently. Thus, the plausibility of my assertion depends significantly on the details of the chance setup, which accords with the controversial status of such counterfactuals**

Backward-Looking Counterfactuals

The method for evaluating vague counterfactuals applies well to forward-looking counterfactuals, but an important problem regarding its applicability to backward-looking counterfactuals needs to be addressed. Previously, the problem with backward-looking counterfactuals was that in the indeterministic case, the laws of nature as we know them do not give any stochastic dynamics by which we can assign probabilities to propositions about the past. A new problem is that in the deterministic case we can assign a probability, but it’s the wrong one. Without adjusting our statistical mechanical probability measure, for almost any antecedent A , the most probable worlds consistent with A are worlds that have a long history of decreasing entropy up until t , worlds that are not at all like those we typically postulate. To see how, start with the presumption that there is some macroscopic counterfactual difference at the state of the world at time t , which can be described macroscopically. The statistical mechanical probability measure for such a situation is such that the overwhelming majority of physical states compatible with the description will evolve into higher entropy states in the future *and have evolved from higher entropy states in the past*. Because we take it for granted that the actual history of the world at least approximately matches what our memories and historical documents purportedly report, we take for granted that the actual world evolved from a low entropy past. Furthermore, the set of worlds that evolved from the same low entropy past as our world forms a connected but very thin and extremely fibrillated set in the space of all possible states. This has as a consequence that a small alteration in the actual physical state will likely take one from a point in the set of having-evolved-from-low-entropy states to some point outside the set. What is very reasonable, is that the kind of counterfactual differences in the present that people typically consider involves just this kind of difference: a change that with high probability goes to a history with a high entropy past. Whether the counterfactual changes are shifts in the position of a single atom or large-scale reshaping of the present physical state, the overwhelming majority of such changes are changes that would entail the past being very different from what it was.

** For people trying to make sense of ordinary claims of causation, the hard case is when the dynamics is indeterministic. The application of this method of evaluating counterfactuals will be of only partial aid. The partial solution is that by appeal to the locality principle, one could count as fixed any event outside the future light cone bounding R . This solution is only partial because it will not help in cases where the influences are within the future light cone but are nevertheless intuitively causally independent of the antecedents’ obtaining.

The consequences of having such a history of decreasing entropy are significant. These worlds are going to look exactly like the actual world immediately before t outside the region R , but any differences in R , so long as they are not completely physically isolated from outside R , are almost always sufficient to disturb the microstate in such a way that the past state outside of R is also an entropy decreasing history. As shown in figure 3, the way such worlds evolve in a GTR world is that the past light-cone of R , is filled with entropy decreasing physics while the space-time outside the cone has entropy-increasing physics. Remember that historical intervals where entropy decreases involve bizarre events like pottery shards assembling themselves into vases, pond ripples converging, etc. Thus, most worlds where A occurs are worlds that possess bizarre anti-thermodynamic behavior in the past light cone of A . Compared to the kind of counterfactual influence the present has on the future, this influence over the past is of a much greater scale. Were the present slightly different, the overwhelming majority of our past history would have been much different.

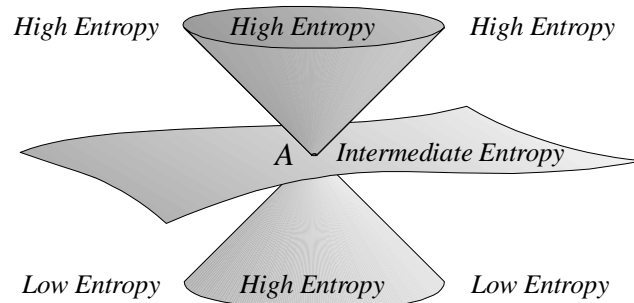


Figure 3

The source of the trouble is the identification of the objective probability distribution with an unsophisticated application of statistical mechanical probabilities, and the appropriate response is to become sophisticated by refining what probability distribution is the appropriate distribution for the evaluation of mundane counterfactuals. The easiest and best solution to this problem is simply to rule out of hand the worlds that evolve from high entropy by assuming the existence of a low entropy macroscopic state occurring in the very early universe. Let the low entropy hypothesis be the name for the claim that the universe at one temporal end, is constrained to be at very low entropy. Furthermore, assume what appears to be true so far as we can tell: that the other temporal end is not so constrained. Then identify what we call the past with the temporal end that is constrained. When we do so, the account of the assertibility conditions of counterfactuals as it applies to a wide range of ordinary counterfactuals may be paraphrased as follows:

$A \square \rightarrow C$ is assertible to degree p , where p is the objective probability of C given A and the low entropy hypothesis.

After doing so, we have the following procedure for evaluating a vague counterfactual $A \square \rightarrow C$ that is about some localizable states of affairs in a deterministic world. One should select a region R that pertains to A at the relevant time, t_A . Then modify the physics in R so that A obtains. Call the state thus obtained, the A -state. This A -state is macroscopically specific in R , and microscopically specific outside R . For all the many ways that A can be instantiated consistent with the context, alter the physics in this

region by putting in some microstate corresponding to A . Call these microstates the micro- A -states. For every micro- A -state, there are some possible worlds that are the worlds one gets by letting the fundamental laws of nature determine what happens at all other times. Let the objective probability distribution over the A -worlds be the usual statistical mechanical distribution over the micro- A -states. Then, conditionalize this probability distribution on the occurrence of the low entropy big bang, and the assertibility of $A \square \rightarrow C$ is the probability of C among the A -worlds.

The important and perhaps surprising result of measuring counterfactuals by assertibility in the way described is that in order for us to accurately evaluate backward-looking counterfactuals, we need to assume that the early universe was in a low entropy state. What is even more surprising is that the low entropy hypothesis will have a significant effect in the evaluation of many forward-looking counterfactuals.

The Entropy Theory of Counterfactuals

In order to have even a chance of having the theory accord with something like our typical assumptions about what we mean when we utter counterfactuals, we are virtually forced to accept the low entropy hypothesis as a tacit background condition. Given this, it would be nice to reap some benefit beyond dissolving an apparent contradiction between our standards for evaluating counterfactuals and our pre-theoretical intuitions about the counterfactual past. Fortunately, we can gain a significant increase in the simplicity of the account by using the low entropy hypothesis. Instead of using a probability measure for each counterfactual that is tailored to each antecedent, we can take a single probability measure over all the worlds consistent with the low entropy hypothesis, i.e. the probability measure that is uniform on the early macrostate. Then, by conditionalizing properly on each antecedent, one gets an objective assertibility rating for $A \square \rightarrow C$.

By thus slightly revising the previous procedure, we gain a simpler way of evaluating $A \square \rightarrow C$. The theory that this procedure gives the assertibility conditions for mundane counterfactuals is what we can call the Entropy Theory of counterfactuals:

1. Take the actual world and select a region R that pertains to the obtaining of A at some time, t_A .
2. Modify the physics in the region so that A obtains. Call the state thus obtained, the A -state. The A -state is macroscopically specific in R , and microscopically specific outside R .
3. Consider the single probability distribution over the initial low entropy microstates that evolve into the A -state.
4. Conditionalize this probability distribution on the occurrence of the A -state.
5. The assertibility of $A \square \rightarrow C$ is the probability of C among these worlds.

The Entropy Theory of counterfactuals is notable in that it takes the macrostate of the early universe and the objective probability distribution over the microstates compatible with this macrostate to be the objective probability distribution for evaluating *any* mundane counterfactual. The particular differences among assertibility ratings for particular counterfactuals are then determined by the range of possibilities permitted by A and C .

How the Entropy Theory Explains the Counterfactual Asymmetry

The following pair of examples demonstrates two remarkable successes of the Entropy Theory. Consider what would have been the case if the match had been struck and ask about some significant historical event in the distant past, e , whether e would still have occurred. Intuitively, we should want to say that e would still have occurred had the match been struck. In using the evaluation method, one important feature is that the current time-slice that has been modified to make the match strike includes physical features of the kind we commonly call records of event e having occurred. If e is Napoleon's ruling France, then we have lots of pieces of the modified physical state at t_A of relevance to Napoleon's ruling France. We have accounts written in books, information in the heads of history professors, etc. Given the macrostate of the early universe, what are the most likely ways that the universe would have evolved to reach the state t_A ? Although there is no sure evidence, it is very plausible that the most likely ways are ways that involve Napoleon's actually ruling France, ways that are, at most, microscopically different from the actual world in the early 19th century. What is far more unlikely is that the textbook accounts of Napoleon would have arrived at their form through independent random processes. What is also unlikely is that the rule of Napoleon would have been a hoax. This kind of situation is ruled out because the typical ways such events are faked are large enough to leave records of the faking in the current physical state. It doesn't matter if these records are scattered and microscopic as their mere existence in the current microstate makes Napoleon's rule highly likely.

In a second example, assume that in reality the match is in the matchbox and consider what would have been the case if the match had been struck. Ask whether the match would have been taken out of the matchbox in the recent past. Intuitively, we should want to say this is likely. Given the Entropy Theory, it turns out to be highly likely that the match was taken out of the matchbox because somehow the match has to be struck and being taken out of the matchbox is the only probable means, given the obvious context in which it can be struck. Generally, it is highly likely that the match-striking situation evolved from some previous situation in ways that are physically likely *just because* they are the likely paths of physical evolution. What is unlikely is that the match was struck through some thermodynamic fluctuation or quantum jump, and what is ruled out altogether is that some miracle needed to occur for the match to be struck.

In these two cases, the procedure for evaluating counterfactuals accords with common sense and with intuitions about the physics underlying the counterfactual events. The picture of how a world evolves into counterfactual situations given by the Entropy Theory is that the universe begins macroscopically just like the actual universe, differing at most only very slightly in microscopic detail. The universe then evolves according to the fundamental physical dynamics. The small differences in microscopic detail make no significant difference for most of the history of the world (or no difference at all if the right kind of indeterminism holds) until finally the small differences become macroscopic and bring about the antecedent in just the ways the antecedent typically comes about.

A Problem with the Entropy Theory

The Entropy Theory, as formulated is likely not correct. The problem is that, on the one hand, the region outside R needs to be defined microscopically in order to

generate the positive results illustrated in figures 1 and 2, but on the other hand, constraining the appropriate micro- A -states to match actuality perfectly outside R , is quite likely too great a constraint. Looked at from the perspective of evaluating sizes in phase space, the problem is this: the universe starts somewhere in a very small region of phase-space that corresponds to extreme low entropy. Then, in about 10 billion years or so, the universe needs to evolve into the A -state. But the A -state is far, far smaller than the already miniscule initial state. Because the A -state has the microstate outside R held fixed *microscopically*, its dimension in phase space is about the same order as the number of particles inside R , maybe somewhere near 10^{30} . Because this number is dwarfed by the total dimension of the universe's classical phase space, it is virtually a single point in phase space, and thus we have some reason to think it very improbable that even a single microstate will evolve into the A -state within about 10 billion years. Even if one could develop an argument that such a possibility exists, one would also need to demonstrate that the corresponding history would bear some resemblance to our own history, another strong constraint.

To address a solution, let us temporarily disregard the Entropy Theory's goal of explaining the counterfactual asymmetry in order to examine how we can create a better match between the way we think counterfactual histories would likely go and the Entropy Theory.

One issue left unclear under the treatment proposed by the Entropy Theory is how to determine the proper extent of R . On the one hand, before becoming sophisticated about such things, we think that counterfactual suppositions about striking matches do not involve twiddling with the physical details in distant places. On the other hand, we think that counterfactual suppositions involve a past history that comes about rather naturally. While the Entropy Theory may seem to be compatible with these principles, there is a tension between these two guiding rules. If the match striking comes about through a process that involves its being taken out of the matchbox and if that history is anything like what we think it would be, that course of events will very likely leave its usual records in the counterfactual present. Images of the matchbox opening, the hand going into the box, etc. will duly ripple out of windows into deep space. If the A -state is constructed according to the principle that counterfactual changes come about through rather mundane physical evolutions, the A -state will contain microscopic records of the match coming from the matchbox and will lack records of any other process by which the match may have gotten out of the matchbox. This conflicts with the idea that the modified present state does not involve microscopic changes far away from match.

As a matter of practice, we typically do not think that the match was struck by teleporting out of the matchbox, miraculously passing through the box or exiting by any other outlandish possibility. Knowing that the matches typically leave matchboxes by human intervention and lacking any other reason, we should tend to think that a human hand took out the match. Counterfactual differences almost always arise through rather mundane physical evolutions, not through miraculous transitions or fantastically improbable quantum jumps, etc. The principle that we restrict consideration to kinds of physical evolution that have non-remote objective possibilities can be dubbed the principle of mundaneness. Adoption of the principle of mundaneness does not commit one absolutely to the impossibility of such possibilities. Certain contexts may make them

immediately relevant and other constraints may raise the probability of such possibilities to levels where they then become relevant to the evaluation of a counterfactual. The principle of mundaneness should rather be seen as another defeasible principle of vagueness resolution for mundane counterfactuals that is respected just when the assertibility ratings are low for improbable forward transition probabilities and relatively high for more probable transitions.

It is a part of this somewhat vague principle of mundaneness that the counterfactually supposed antecedents will have their counterfactual precedents, so that the counterfactual history, while not identical to actual history, is still usually assumed not to stray further from actuality in a way that departs gratuitously from the actual course of affairs. Counterfactual suppositions about events of the past week, as a matter of conversational practice, do not entail differences of historical fact about the far distant past. Were the weather more pleasant yesterday, Napoleon would still have ruled France and the Hanging Gardens of Babylon would have still been constructed. Of course, in conversation, usually no conscious assumptions at all are made concerning such historical events. Counterfactuals of this far-backward-looking sort are, to be generous, very odd. Yet, to the extent such history is considered, it is taken to be mostly fixed under counterfactual consideration. Where it is not fixed, the appropriate methods of accommodation are often obvious. If someone had found what appears to be a dinosaur fossil in some pit where none actually exists, then the supposed fossil would have been found because a dinosaur had died nearby some long time ago. What would not have happened (or rather what is exceedingly unlikely to have happened) is that only recently, particles of soil would have assembled themselves into a false record of a dinosaur's existence. The overall fixedness of the past then stands as a background assumption that we may safely take to be a component of the mundaneness principle. The mundaneness principle should be understood as the assumption that significant counterfactual differences arise mostly as the result of a past history very much like actuality that at some point starts to diverge in rather ordinary ways.

The problem with the method for evaluating counterfactuals as it stands now is that keeping the microstate fixed outside of R is much too strong a condition for the mundaneness principle to hold. For example, consider a potential traveler who considers leaving town by the morning train in order to travel to the city on the only train that goes to the city that day. He decides, in the end not to travel, but we can speculate as to what would have happened had he been in the city that afternoon. If we maintain the principle of mundaneness, this situation is presumably associated with possible worlds where he got on the morning train and traveled in as a person normally does. The difference between this sequence of affairs and the actual sequence of affairs shows up importantly outside the region R . By various radiative processes, light-images of his location, noises from his feet and mouth, etc. will travel outward and make some difference in regions remote from his town and the city.

Even though mundaneness is primarily a principle concerning backward-looking counterfactuals, its conflict with locality seems to indicate that we cannot straightforwardly evaluate even forward-looking counterfactuals like

(25) If the man were in the city this afternoon, he would take the train back home at night.

By the Entropy Theory, this conditional is evaluated by taking a time-slice of the actual world in the afternoon, delineating some small portion of the city and modifying it to include the man with all his usual proclivities, and then letting the dynamical laws indicate whether the man travels home by train that night. But, this certainly cannot be a method that represents accurately what would have happened had the man been in the city. If he were in the city, he would be there because he left town in the morning and had the usual sorts of physical interactions. If we don't make adjustments in the current state to accommodate a mundane evolution of the counterfactual situation, the situation being evaluated by the Entropy Theory does not correspond with the intended interpretation of the counterfactual, and it may result in the wrong conclusions. If the man were to be in the city this afternoon, and in the town immediately beforehand, then he must have been teleported there, which would likely arouse suspicion in others as well as in the man himself, all of which may lead to his checking himself into a hospital instead of returning home.

The most reasonable emendation to the Entropy Theory is to relax the condition that the microstate outside R be fixed exactly. The constraint could be replaced by some restriction that the macroscopic features of the world outside R be maintained or that the positions of particles outside R be maintained within some acceptable distance. If so, one would not need to worry about the many disparate microscopic records spreading out into the future from a counterfactual event. Small, localizable counterfactual suppositions as exist in human decision-making could then be evaluated in a way that respects the mundaneness principle. The A -worlds would have histories that start with extremely small differences from the actual world growing very slightly throughout history until they differ macroscopically from actuality in R . There would certainly be some differences outside R necessitated by the interaction of the particles throughout their history with other elements of the universe, but one could hope that these differences remain small.

This Revised Entropy Theory would allow the freedom to go back and make some minimal changes at some time before the antecedent that would lead naturally to the antecedent's obtaining. In the example above, what seems intuitively correct is that the appropriate time at which to start making modifications is in the morning as the man is deciding whether to travel. At that time, the only sizable changes needed are highly localized in the man's brain.

There are two potential problems with applying this strategy. First, this method causes problems with affecting unrelated chancy outcomes. Suppose that in actuality some rare chancy outcome occurred. To be specific, say that some woman in the city whom our man would never meet even if he went to the city won a huge sum at the roulette wheel at around noon. If we choose to evaluate what would have happened if the man had been in the city in the afternoon by modifying the state in the morning and calculating what was likely to have happened, it will turn out that it was likely that the woman lost her gamble. Thus, the method seems to license the inference

(26) If the man were in the city this afternoon, the woman likely would not have won her gamble.

This implies some kind of counterintuitive dependence of the woman's income on the man's decision to stay home even in circumstances where they never come near one another and would not have come near one another if he had traveled to the city. For any

theorist who proposes that causation is a matter of increasing the probability of certain counterfactual events occurring, our formulation so far might imply that the man's staying in the town had caused the woman to win: had he been in town (as he was), she would have won; had he not been in the town, he would have been in the city and she likely would not have won. Admittedly, any potential problems with counterfactual accounts of causation that involve probability raising, need to be addressed to specific theories of causation. Such an account might evade the counterintuitive results for instance, by considering the chances of the woman winning at the time the decision is made by the man to stay in town. How such a problem bears on accounts of causation is unclear, but there are potential pitfalls of unsystematically varying the time at which counterfactual changes are made.

It is not so counterintuitive that the man's being in the city has *some* influence on the woman's winning the gamble. Subtle influences from the man's presence on the train may pass from him to other passengers and objects, which further transmit these influences to the roulette wheel operator. If the wheel system is a sensitive enough device, the outcome of the gamble may be affected. The same goes for quasi-chancy outcomes at very great distances from *R*. For the reasons mentioned earlier (p. 59), the Entropy Theory successfully treats quasi-chancy phenomena outside *R*, a success we would like not to undermine. What we had earlier is that counterfactual suppositions about localized events did not affect the outcomes of quasi-chancy events outside the light-cone centered around *R*. Now, we are allowing that counterfactual events in *R* make it likely that certain non-actual events preceding *A*'s obtaining would be likely, and that these new events would have an influence outside the light-cone, possibly affecting quasi-chancy outcomes. This is not problematic as an exercise of intuitions about counterfactuals because we have a story to explain the distant influences. Were *A* to be true, the past must have been different, and if the past were different, that difference would have made a broad impact on the present. Nevertheless, any application of the Entropy Theory to causation will face a challenge in explaining away these kinds of "super-luminal" influences.

Second, by giving up on a strict interpretation of the locality principle, we appear to be giving up on an objective evaluation of the counterfactual. It is perhaps plausible to pick some definite, reasonable time in the example of the potential traveler, but that is in part because the example was tailored to make such a selection seem reasonable. In general, there may be many small local changes that could bring about the antecedent's obtaining, and there may be none. We have no objective way of sorting out where to start modifying the physical state. It also may be that the counterfactual differences outside *R* that one hopes are small are not at all negligible, so that spurious counterfactual dependence between distant events threatens.

On the one hand, this is a success of the Entropy Theory. Backward-looking counterfactuals are hard to evaluate because our intuitions are strained by the piecemeal utilization of principle that conflict with one another. This explains the difficulty we have with examples like Downing's (1959; see also Bennett, 1984, and Lewis, 1979). In Downing's example, Jim and Jack quarreled yesterday, so it seems if Jim asked Jack for a favor, Jack would refuse. Yet, Jim is proud, so if he asked for a favor there would have been no quarrel, and if there were no quarrel, Jack would oblige. Even without the conflicting facts illustrated in Downing's example, it is unclear how to evaluate

counterfactuals unless there clearly is a single small region where counterfactual changes first start becoming macroscopic. We thus have a match between unclear intuitions about how backward-looking counterfactuals should be evaluated and unclear pronouncements from the Entropy Theory.

On the other hand, the inability of the theory to select objectively either an appropriate time or an appropriate local region to use as an input into the evaluation procedure is a major deficiency in the Entropy Theory's explanation of counterfactual asymmetry. The Entropy Theory's explanation of counterfactual asymmetry is that the assertibility of truly counter-to-fact conditionals of the form $A \square \rightarrow C$ when C describes a macroscopic fact about the past is typically very near 1, whereas it is not typically near 1 when C describes a macroscopic fact about the future. This is just the formal rendering of the claim that if the world were somehow different, most past facts would still obtain, but not necessarily most future facts. Critical to making this explanation work is that significant past events leave sufficient records in the present state. One counterexample to the Entropy Theory is to take some improbable event e for which there are no macroscopic records. Exotic examples exist, such as a macroscopic happening on a planet that later falls into a black hole, but there is no shortage of everyday examples because there are many processes in our world that remove records, processes that reduce macroscopic differences to microscopic differences. We can imagine a situation where pure water has been spilled in the front half of a very clean room, has evaporated, and has diffused uniformly. Had the water been spilled in the back half, the water in time would have evaporated and spread uniformly. In any such case, differences between the actual world and the counterfactual world would all be microscopic. If our event e for which there is no microscopic record was improbable given the physical circumstances shortly before e occurred, then the following would incorrectly turn out as highly assertible: If the world were presently a little different, e would not have occurred. This follows from the fact that e is physically improbable to begin with and the fact that there is nothing in the A -state to make it more probable. The occurrence of e , then, greatly depends on the way the world is now, even when its existence should be independent of the present because it is spatially remote or in the distant past. This problem successfully vitiates any hope that the A -state constrained by the low entropy hypothesis is sufficient to keep all distant, unrelated past facts fixed.

In the end, even the Revised Entropy Theory is a flawed account of counterfactual conditionals because in general, there is no systematic way to evaluate mundane counterfactuals except in the special case when one the counterfactual differences between the A -state and actuality at time t_A are small. In light of this failure it may be tempting to dismiss the theory entirely. This would be a mistake because one can salvage a positive story about how counterfactuals should be evaluated. There are many processes that obey the counterfactual asymmetry properly, processes where small counterfactual differences at one time are able to generate worlds in which the antecedent is true and past facts are mainly fixed, and in which the past differences are entirely due to the past needing to be different in order for the physical evolution to give rise to the antecedent. These are cases where the counterfactual differences can be traced to very small spatial regions at some appropriately short time before t_A .

Furthermore, it seems plausible that the decisions of rational creatures are the kinds of processes that we can easily imagine as arising out of very localized differences,

so that rational decision making might very well be a success case for the Entropy Theory. One good measure of success for the Revised Entropy Theory is that it fit into an explanation of free will and influence among other things. It may be reasonable to say that our intuitions about the fixedness of the past and openness of the future are attributable to our intuitions about special cases like human decisions where the antecedents involve only small changes or can be unproblematically thought of as arising from small changes. The asymmetry in these cases can be explained by the low entropy at the beginning of the universe in the way proscribed by the Revised Entropy Theory. Then, we extend the principles founded on these special cases, to larger-scale counterfactual differences where they don't apply so well. Mainly, the differences between highly localizable antecedents and larger-scale antecedents is not troublesome and not even noticed because typical counterfactual assertions are both vague and forward-looking, two characteristics that tend to hide the difficulties brought out in a close evaluation of the Entropy Theory.

While the Revised Entropy Theory does not enjoy full success, it has several promising features. First, unlike alternative accounts, it uses an assertibility condition that successfully treats counterfactuals that involve chancy and quasi-chancy phenomena. Second, in evaluating counterfactuals by way of modifying an actual state of the world to accommodate the antecedent in a context sensitive manner, it treats counterfactuals in a way that is more consistent with the way people imagine alternatives developing. Third, unlike other theories, the Revised Entropy Theory in many cases makes definitive assessments of backward-looking counterfactuals that accord with naïve intuitions about such cases. While not a perfect theory for the objective evaluation of all ordinary counterfactuals, the Revised Entropy Theory does provide a good procedure for evaluating a wide range of mundane counterfactuals.

The Asymmetry of Influence

While the adequacy of the Revised Entropy Theory may be acceptable as a useful theory of the proper evaluation of ordinary counterfactuals, its ability to play a significant role in the explanation of the influence asymmetry is questionable at best. The best explanation of the influence asymmetry from the Entropy Theory goes as follows:

A simple analysis of influence in terms of counterfactual dependence is fairly straightforward. Let S be a rigid designator of some state of affairs in the actual world. P 's bringing about the obtaining of A influences S to the degree that

$$\sim A \square \rightarrow \sim S$$

is assertible. This analysis applies to cases where S is the obtaining of some event(s) as well as to cases where the potential influence is over the exact nature of some occurrence. For example, the salinity of water influenced the rusting of iron to the degree that: had the water not been as salty as it actually was, the iron wouldn't have rusted the amount it actually did. The asymmetry of influence just follows from the counterfactual asymmetry. Because macroscopic facts about the past are mostly fixed under mundane counterfactual supposition, mundane actions are such that they don't influence past events.

The worrisome part about such a treatment of influence, together with the Revised Entropy Theory, is that it implies some counterintuitive results. First, because there is some microscopic counterfactual dependence of the past on the future for almost any counterfactual when the laws are deterministic, we may have some microscopic influence on the past. Second, since there are some nearby recent facts about the past that counterfactually depend on some present facts by way of the mundaneness principle, we may have some macroscopic influence over the immediate nearby past.

The following represents a situation someone might worry about. Suppose there is a reliable guard whose job it is watch a certain field and if he sees an explosion, to press a certain button. Otherwise, he is not to press it. An explosion occurs, and he dutifully presses the button. Consequently, the influence of the button pressing on the explosion is given by the assertibility of

(27) If he had not pressed the button, there would have been no explosion.

Intuitively, he is not influencing or causing the explosion, but merely reporting it. So it better turn out that (27) has low assertibility.

The Revised Entropy Theory's evaluation of this counterfactual illustrates the problem of balancing the principles of locality and mundaneness. By placing a lot of weight on locality, one modifies the state after the explosion to put the guard's brain state in a configuration where he will press the button. One might try to presume that there is some microstate of the initial universe that will eventuate in a state where that one small change is the only significant change. If one then takes the region R just to include the guard's brain, all the corresponding micro- A -states will contain extensive records of the explosion. Hence, almost all the A -worlds will be worlds with an actual explosion, and the assertibility of (27) will be very low.

Yet, one might weigh mundaneness more heavily. After all, if the guard is truly reliable, the probability is low that he would have failed to press the button after seeing the explosion. Hence, the worlds where he doesn't press the button will be almost all worlds where there is no explosion. The Revised Entropy Theory can account for his not pressing the button in this mundane way, by having a microscopic changes in the early universe that eventuate in significant differences in the world some time before the explosion to make the explosion not occur. Most of these worlds are worlds where the guard doesn't press the button. Thus, using this procedure, the assertibility of (27) is quite high.

In a defense of the Revised Entropy Theory's ability to explain the asymmetry of influence, one might just staunchly defend the first evaluation, where locality counts greatly. (David Lewis's treatment of the problem of effects for his earlier theory of causation follows this strategy.) The problem with doing so is that one can enhance the example to make this interpretation as implausible as one wants. Replace the guard with extremely reliable detecting equipment with as many independent backup systems as are needed. To the extent that one tries to hold on to the first interpretation against such modifications, one is denying the intrinsic appeal of the mundaneness principle beyond credulity. Rather, one should accept that there is a counterfactual dependence of the explosion on the button pressing, but identify it not as an instance of backwards influence, but as ordinary detection and record making.

An alternative defense is to argue that the asymmetry present in our concept of influence does not track the counterfactual asymmetry of individual cases but rather in

the collection of cases that constitute the basis for our intuitions about counterfactuals. Using the Revised Entropy Theory's procedure, we have many examples of counterfactual differences at one time entailing significant differences in the future but not in the past. These examples, like human decisions, form the core that the Revised Entropy Theory successfully explains. And using the procedure in these cases, we also lack any realistic cases where small counterfactual differences entail significant differences in the past. So, this argument would go, the influence asymmetry has its ground in the special cases where the counterfactual differences are initially microscopic and localized. Then, the influence asymmetry takes on a life of its own, so that in cases where the conditions lead one to think that the past would have been different if the present were different, one identifies such cases as examples where the counterfactual asymmetry does not match up with the influence asymmetry. That is, the direction of influence is equal to the predominant direction of counterfactual dependence among mundane counterfactuals.

Examples exist where the counterfactual dependence seems to go back far in time and be systematic. For instance,

(28) If our galaxy were rotating in the opposite direction, it would have been rotating in the opposite direction a billion years ago.

This is the kind of counterfactual that the Entropy Theory would assess as having a very high assertibility because if a small change were to bring about the difference in direction, it would have to have been long before the galactic motion was set stably in place. It is obvious enough, though, that the current direction of the galaxy does not influence the past direction so much as it is a consequence of the earlier states. The concept of influence behind these intuitions is one that has a direction that is fixed and independent of the relevant counterfactuals. The alternative strategy accepts this but in addition holds to the idea that if the world were such that small counterfactual differences of the kind that are implicated in human action were systematically to entail larger differences in the past than in the future, then the direction of influence would be altered. We would say for the beings of these worlds, that they influence what we in this world call the past.

While this strategy is plausible, it cannot serve as an account of how the counterfactual asymmetry explains the influence asymmetry because the account is circular. To see how, remember what the hoped-for explanation of the counterfactual asymmetry was before the conflict with the mundaneness principle was noted and the Entropy Theory was revised by weakening the locality principle. Originally, the explanation was that the fleshed out counterfactual state, i.e. the *A*-state, together with the low entropy hypothesis would leave the probability of actual macroscopic past facts high, but would reduce the probability of some actual macroscopic future facts. This is just what it means in the Entropy Theory for the future to be malleable and past to be fixed.

But in order to save the Entropy Theory from the untoward result that it conflicted with the mundaneness principle (and hence had little relevance to the kinds of counterfactuals people actually use to explain choices, chance outcomes, etc.), the theory was modified to allow the principles of locality and mundaneness to be balanced against one another in resolving the vagueness inherent in ordinary counterfactuals. In doing so, there is an implicit reliance on the influence asymmetry. How do we select worlds to

evaluate a given counterfactual? We find some interval in the *past* where we can modify the actual state in a small way so that the antecedent obtains from a microscopic difference in the initial variables of the universe that becomes macroscopic only at this time. Remember that there is no guarantee in general for the ordinary kinds of counterfactuals under consideration that the changes can be focused in one area, nor can it be guaranteed that there will be no macroscopic changes far away from the region R . Rather, we pin the success of our theory for such counterfactuals on the hope that such a resolution is possible. Because we are setting out to find a resolution that keeps the past mainly fixed while still guaranteeing ordinary kinds of physical evolution, we are putting a commitment to the mundaneness principle before our commitment to the low entropy hypothesis. The principle of mundaneness includes essentially a restriction to worlds that mostly match our world in the distant past, which is a poorly disguised elaboration of the asymmetry influence.

The true relationship between the counterfactual asymmetry and influence asymmetry, then, is as follows. The Revised Entropy Theory succeeds in explaining the counterfactual asymmetry in terms of a global entropy asymmetry when the antecedent involves only small changes at a time, and the influence asymmetry matches the counterfactual asymmetry in these core cases. The theory, however, extends to more general counterfactual conditionals in a reasonable way only by taking for granted the generality of the influence asymmetry. Thus, the Revised Entropy Theory does not explain the influence asymmetry but assumes it as one principle of vagueness resolution.

BIBLIOGRAPHY

- Adams, E. W. (1965) "The Logic of Conditionals," *Inquiry*, 8, 166-97.
- . (1966). "Probability and the Logic of Conditionals," in J. Hintikka and P. Suppes (eds) *Aspects of Inductive Logic*. Amsterdam: North Holland, 265-316.
- . (1970). "Subjunctive and Indicative Conditionals," *Foundations of Language*, 6, 89-94.
- . (1975). *The Logic of Conditionals: An Application of Probability to Deductive Logic*. Dordrecht: D. Reidel.
- . (1976). "Prior Probabilities and Counterfactual Conditionals," in W. L. Harper and C. A. Hooker (eds), *Foundations of Probability Theory, Statistical Inference and Statistical Theories of Science*, Volume 1, 1-21.
- . (1993). "On the Rightness of Certain Counterfactuals," *Pacific Philosophical Quarterly*, 74, 1-10.
- . (1996). "Four Probability-Preserving Properties of Inferences," *Journal of Philosophical Logic*, 25, 1-26.
- Albert, D. (2000). *Time and Chance*.
- Ayers, M. R. (1965). "Counterfactuals and Subjunctive Conditionals," *Mind*, 74, 347-64.
- Bennett, J. (1974). "Counterfactuals and Possible Worlds," *Canadian Journal of Philosophy*, 4, 381-402.
- . (1984). "Counterfactuals and Temporal Direction," *Philosophical Review*, 93, 57-91.
- . (1988). "Farewell to the Phlogiston Theory of Conditionals," *Mind*, 97, 509-27.
- Bowie, G. L. (1979). "The Similarity Approach to Counterfactuals: Some Problems," *Noûs*, 13, 477-98.
- Chisholm, R. M. (1946). "The Contrary-to-Fact Conditional," *Mind*, 55, 289-307, reprinted with revisions in H. Feigl and W. Sellars (1949) *Readings in Philosophical Analysis*. New York: Appleton-Century-Crofts, 482-97.
- Downing, P. B. (1959). "Subjunctive Conditionals, Time Order, and Causation," *Proceedings of the Aristotelian Society*, 59, 125-40.
- Dudman, V. H. (1983). "Tense and Time in English Verb Clusters of the Primary Pattern," *Australian Journal of Linguistics*, 3, 25-44.
- . (1984). "Conditional Interpretations of *if*-sentences," *Australian Journal of Linguistics*, 4, 143-204.
- . (1988). "Indicative and Subjunctive," *Analysis*, 48, 113-22.
- . (1989). "Vive la Révolution!" *Mind*, 98, 591-603.
- Edgington, D. (1995). "On Conditionals," *Mind*, 104, 235-329.
- Eells, E., and Skyrms, B. (eds). (1994). *Probability and Conditionals: Belief Revision and Rational Decision*. Cambridge: Cambridge University Press.
- Ellis, B. (1984). "Two Theories of Indicative Conditionals," *Australasian Journal of Philosophy*, 62, 50-66.
- Elga, A. (2000). Conference Paper.
- Fine, K. (1975). Critical Notice of David Lewis's *Counterfactuals*, *Mind*, 84, 451-58.
- Ghirardi, G. C., Rimini, A. and Weber, T. (1986). "Unified Dynamics for Microscopic and Macroscopic Systems," *Physical Review D* 34: 470.

- Goodman, N. (1947). "The Problem of Counterfactual Conditionals," *The Journal of Philosophy*, 44, 113-28 reprinted in N. Goodman (1955) *Fact, Fiction, and Forecast*. Indianapolis: Bobbs-Merrill.
- . (1955). *Fact, Fiction, and Forecast*. Indianapolis: Bobbs-Merrill.
- Grice, H. P. (1975). "Logic and Conversation," in D. Davidson and G. Harman (eds), *The Logic of Grammar*. Encino: Dickenson, 64-75.
- Harper, W. L. and Hooker, C. A. (eds). (1976). *Foundations of Probability Theory, Statistical Inference and Statistical Theories of Science*, Volume 1. Dordrecht: D. Reidel.
- Harper, W. L., Stalnaker, R., and Pearce, G. (eds). (1981). *Iffs: Conditionals, Belief, Decision, Chance, and Time*. Dordrecht: D. Reidel.
- Jackson, F. (1977). "A Causal Theory of Counterfactuals," *Australasian Journal of Philosophy*, 55, 3-21.
- Kvart, I. (1975). *Counterfactual Conditionals*. Ph.D. Dissertation. University of Pittsburgh.
- . (1980). "Formal Semantics for Temporal Logic and Counterfactuals," *Logique et Analyse*, 23, 35-62.
- . (1986). *A Theory of Counterfactuals*. Indianapolis: Hackett.
- . (1991). "Counterfactuals and Causal Relevance," *Pacific Philosophical Quarterly*, 72, 314-37.
- . (1992). "Counterfactuals," *Erkenntnis*, 36, 139-79.
- . (1994). "Counterfactuals: Ambiguities, True Premises and Knowledge," *Synthese*, 100, 133-64.
- Lewis, D. (1971). (1973a). *Counterfactuals*. Oxford: Blackwell.
- . (1973b). "Causation," *The Journal of Philosophy*, 70, 556-67.
- . (1973c). "Counterfactuals and Comparative Possibility," *Journal of Philosophical Logic*, 2, 418-46, reprinted in W. L. Harper, R. Stalnaker, and G. Pearce (eds) *Iffs*. 57-85.
- . (1976). "Probabilities of Conditionals and Conditional Probabilities," *Philosophical Review*, 85, 297-315, with erratum in (1976), 85, 561, reprinted in W. L. Harper, R. Stalnaker, and G. Pearce (eds) *Iffs*. 129-47.
- . (1979). "Counterfactual Dependence and Time's Arrow," *Noûs*, 13, 455-76.
- . (1986a). *On the Plurality of Worlds*, Oxford: Blackwell.
- . (1986b). *Philosophical Papers, volume 2*. Oxford: Oxford University Press.
- . (1979). "Cotenability and Counterfactual Logics," *Journal of Philosophical Logic*, 8, 99-115.
- . (1980). *Topics in Conditional Logic*. Dordrecht: D. Reidel.
- Quine, W. V. (1950). *Methods of Logic*. New York: Holt, Rinehart, and Winston.
- Skyrms, B. (1981). "The Prior Propensity Account of Subjunctive Conditionals," in W. L. Harper, R. Stalnaker, and G. Pearce (eds), *Iffs*. Dordrecht: D. Reidel, 259-65.
- . (1994). "Adams Conditionals," in E. Eells, and B. Skyrms (eds), *Probability and Conditionals*. Cambridge: Cambridge University Press, 13-26.
- Slote, M. (1978). "Time in Counterfactuals," *Philosophical Review*, 87, 3-27.

- Stalnaker, R. (1968). "A Theory of Conditionals," in N. Rescher (ed), *Studies in Logical Theory*, *American Philosophical Quarterly Monograph Series*, No. 2, Oxford: Basil Blackwell, 98-112, reprinted in E. Sosa (ed) *Causation and Conditionals*, Oxford: Oxford University Press, 165-79, and in W. L. Harper, R. Stalnaker, and G. Pearce (eds). *Ifs*, 41-55.
- . (1981). "A Defense of Conditional Excluded Middle," in W. L. Harper, R. Stalnaker, and G. Pearce (eds), *Ifs*. Dordrecht: D. Reidel, 87-104.