

# The Entropy Theory of Counterfactuals\*

Douglas N. Kutach<sup>†‡</sup>

Department of Philosophy, University of Oklahoma

---

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 method of evaluating counterfactuals that includes, as a background assumption, the low entropy of the early universe. The resulting theory attempts to vindicate 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. Although the theory has some success in evaluating a wide variety of ordinary counterfactuals, it fails as an explanation of counterfactual asymmetry.

---

**1. Introduction.** 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 some 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, 2001). In what follows, I present what I take to be the best theory of counterfactuals that makes the connection between entropy and counterfactuals explicit, hereby dubbed the Entropy Theory of Counterfactuals. While it justifies many features present in our intuitive and scientific assessments of counterfactuals, it fails in the end to count as a successful explanation of counterfactual asymmetry.

It is doubly remarkable that one can find anything approaching a good

\*Received November 2000; revised August 2001.

<sup>†</sup>Send requests for reprints to the author, Department of Philosophy, University of Oklahoma, Norman, OK 73073; e-mail: kutach@ou.edu.

<sup>‡</sup> Credit for assistance in the development of this paper goes to Tim Maudlin, David Albert, and Frank Arntzenius.

Philosophy of Science, 69 (March 2002) pp. 82–104. 0031-8248/2002/6901-0004\$10.00  
Copyright 2002 by the Philosophy of Science Association. All rights reserved.

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, 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 a 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 conception of entropy is superior to competitors like the ensemble approaches because it applies to real individual systems, and it fits into a reasonable explanation of entropy increase. The primary benefit of adopting the statistical mechanical formalization 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 experiments have confirmed the empirical adequacy of the observer-independent statistical-mechanical probability distributions, 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 Adams' puzzles, made the association with objective propensities. These perspectives are different from the standard account by Lewis (1973a, 1973b; 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 must accept that the similarity relations are constrained by the objective probability relations. While this is not problematic *per se*, the probability constraints have nothing to do with any intuitive measure of similarity, which causes trou-

ble for anyone like Lewis (1979) 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, which is the degree to which one should be willing to assert the counterfactual, or to believe it, or to assent to its truth. The notion of assertibility usually connected with indicative conditional statements, the notion that has been so successful in explaining our judgments concerning indicative conditionals, is a *subjective* conditional probability. The degree to which Jim is willing to assert, “If  $A$  is the case, then  $C$  is the case,” is equal to the conditional probability,  $\text{Pr}_{\text{Jim}}(C/A)$ , defined by Jim’s subjective degrees of belief in  $A$  and in  $C$ . One can also understand an *objective* notion of probability for subjunctive conditionals. The objective assertibility of “If  $A$  were the case, then  $C$  would be the case,” is the degree to which one *should* find that statement assertible and is equal to  $\text{Pr}_{\text{Obj}}(C/A)$ , a function of the objective likelihoods of  $A$  and of  $C$ . 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  $\text{Pr}_{\text{Obj}}(C/A)$  and discuss the corresponding theory of counterfactuals.

**2. 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  $\xi$ , then  $C$  would have been the case,” is evaluated by taking the state  $\xi$  and letting the evolution of  $\xi$  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 world 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 prob-

ability 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  $\xi$  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  $\xi$ , determine the probability for 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 obstacle appears when 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 some set of actual facts which properly count as the right background conditions without begging the question as to which counterfactuals were true. Although Goodman's project is widely understood to be a failure, that does not rule out the possibility of finding a cohesive set of tacit background conditions that people should maintain when evaluating counterfactuals. It only implies that no theory can justify 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 recog-

nizable overlap among most of the contexts where people use counterfactuals.

The theory that will emerge does not aspire to take sides in the debate between those theorists like Goodman, Mackie, and Kwart on one side who follow in the tradition where a counterfactual is conceived as an elliptical claim that the antecedent together with some laws and background facts entail the consequent, and theorists like Lewis and Stalnaker, on the other side, who think counterfactuals express judgments about the similarity of various possibilities. While superficially, the evaluation procedure might appear similar to Goodman's approach, it is also compatible with the superior and more general possible worlds approach. When cast as a possible worlds theory, it should be seen as an attempt to compete with Lewis's (1979) account of counterfactual asymmetry where he constructs a theory of the similarity relation that tacitly guides ordinary counterfactual evaluation in terms of match of fact and size of miracles, etc. Like Lewis's account, the Entropy Theory tries in part to capture the native reasoning people use regarding counterfactual situations and in part to offer a sophisticated formalization that can make this reasoning compatible with other important epistemic commitments and to guide us in complicated situations. Judging its success is a matter of evaluating its ability to approximate key pre-theoretical commitments and to cohere with our best scientific knowledge.

One principle that seems at first to be tacit in our counterfactual reasoning is that antecedents describing a situation without explicit 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 certainly 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.

Consider a box containing a gas, where the box has a very small perforation, a circular hole with a radius a few times 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  $\Delta t$ . Consider the counterfactual, "Had the gas been one degree hotter, a gas molecule would have escaped during  $\Delta t$ ." 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 well-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 rates would be lower.” 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. Presumably in such cases this probability is just the probability one would find in an accurate assessment of the anti-cancer benefit of vegetables. In other cases where the circumstances are too varied for a single assertibility value to achieve legitimacy, one can resort to the usual arguments regarding the assessment of vague predicates. In short, one can consider the class of more precise reformulations where such probabilities do exist.

**3. The Locality Principle.** Another principle for resolving vagueness is that, except where the 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  $\xi_R$  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  $\xi_R$ , then  $C$  would have been the case.” To evaluate this counterfactual, one should take the entire actual microscopic state at time  $t$  and modify the details in  $R$  so as to make the localized state of affairs  $\xi_R$ . Because the physical state outside  $R$  was not mentioned in the antecedent, one should not, without some overriding reason, modify the physical state outside  $R$ . In this particular case, there is clearly an informal implicature, because  $R$  was mentioned explicitly, that the situation to be considered is one where the actual world has been modified in region  $R$  and nowhere else. 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. Other considerations, like a desire to preserve conservation laws, may take precedence: “If there were more mass in  $R$ , then there would be less mass outside  $R$ .”

Despite its defeasibility, the locality principle, as a rough guide to the evaluation of ordinary counterfactuals, is implicated in both everyday us-

age of counterfactuals as well as more formal usage. Some evidence for the locality principle, beyond 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 relativity.

Start with a relativistic world where some brief distant event  $e$  is occurring, and consider a 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 modifying only the local 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. Our best relativistic surrogate for this state is a space-like hypersurface, and fortunately 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 intuitively would want to say because  $e$  is too far away for the nearby events to have any dynamical influence.

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  $A$  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

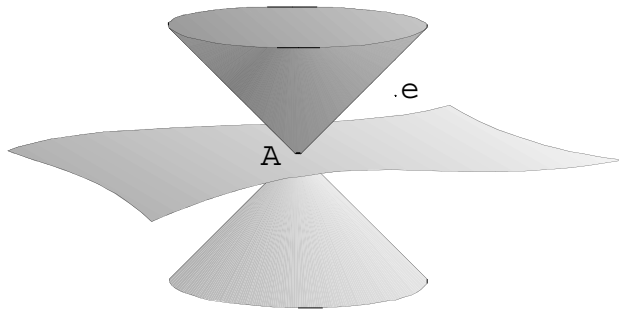


Figure 1

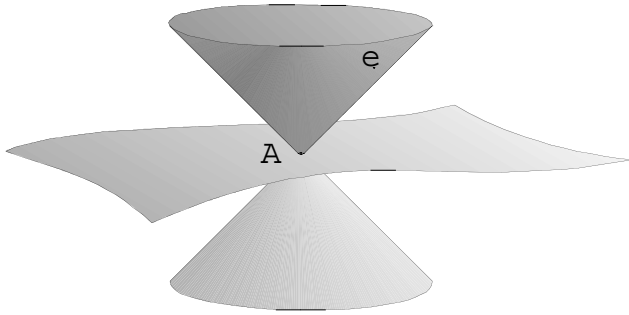


Figure 2

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 almost all, if not all, of these worlds.

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 instantiating  $A$  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. Let  $e$  be the selection of a lottery ticket. The winner,  $X$ , might want to say, "If I hadn't prayed for luck, I wouldn't have won." This comes out highly assertible if  $X$ 's praying had significant dynamical influences on the lottery selection, e.g., by distracting the person rotating the ticket bin. Indeed, one may reasonably think in such a case that the

assertibility is about equal to X's chance beforehand of losing the lottery. Yet, if abstaining from prayer has dynamically insignificant consequences for the lottery mechanism, then the probability of winning would remain high, and the counterfactual would be highly unassertible. Thus, the plausibility of X's assertion depends significantly on the details of the chance setup, which is a reasonable outcome for these kinds of counterfactuals. (The really hard case is when the chance outcome is genuinely indeterministic. The account here is not able to vindicate the intuition that such counterfactual outcomes should be wholly unaffected by events that are causally irrelevant.)

**4. 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 in 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*.

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. Thus, most worlds where  $A$  occurs are worlds that possess 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 very different.

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 so-

phisticated 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. 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, the  $A$ -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 a mixture of the usual statistical-mechanical distribution over the micro- $A$ -states along with any probability distribution arising from stochastic evolution. 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 surprising result of measuring counterfactuals by objective assertibility 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 perhaps even more surprising is that the Low Entropy Hypothesis will have a significant effect on the evaluation of many forward-looking counterfactuals.

**5. The Entropy Theory of Counterfactuals.** It is safe to say that typically when one considers how the world could have been otherwise, one has in mind situations that are not the result of an immensely improbable thermodynamic evolution. So in order for the theory to have even a chance

of according with 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  $C$  and the range of possibilities permitted by  $A$ .

**6. Two Successes of the Entropy Theory.** 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 historians, 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 19<sup>th</sup> 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 Napoleon's rule 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 at least microscopic records of the faking in the current physical state.

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, by 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 far more unlikely is that the match would have been struck through some thermodynamic fluctuation or quantum jump, and miraculous strikes are entirely ruled out.

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 perhaps 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.

**7. 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 and in the Napoleon example, but on the other hand, constraining the appropriate micro- $A$ -states to match actuality perfectly outside  $R$ , is likely too great a constraint if the world is deterministic. A brief look at the sizes of the relevant phase spaces should arouse some suspicion about this constraint: the universe starts somewhere in a very small region of phase-space that corresponds to extreme low entropy. Then, in about 15 billion years or so, the universe needs to evolve into some  $A$ -state. But the class of  $A$ -states 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  $R$ , which is dwarfed by the total dimension of the low entropy macrostate, which itself is dwarfed by the total phase space. This great disparity in sizes is good reason, in the absence of some countervailing argument, to doubt that there are dynamical paths of evolution that connect the two regions of phase space.

Looking at the same problem another way, consider how fragile the current microstate is regarding its ability to evolve backward in time to reach the low entropy past. Often, even the slightest microscopic changes to a state will make it such that that state, when retrodicted back to the early universe, will have very high entropy. The changes to the actual state needed to accommodate an antecedent are typically macroscopic, so consider the actual state modified in some region  $R$  to accommodate  $A$ , which is macroscopically different from actuality. Then examine the backward time-evolution of a typical corresponding micro- $A$ -state. As soon as we start the backward evolution physical interaction between the matter inside  $R$  and the matter outside  $R$  will typically occur. There are long-range gravitation effects as well as shorter-range effects that ripple outwards from  $R$ . The effect of such small differences is often sufficient to change the collision angle of a particle pairs, which in turn makes still larger changes, and so on for 15 billion years. Because of the highly sensitive nature of an evolution towards low entropy, such small changes seem to be enough to prevent any micro- $A$ -state from evolving towards low entropy. In any case, we have no reason yet to think that for any given  $A$ , there must be some micro- $A$ -state that is compatible with the Low Entropy Hypothesis. And even if one could develop an argument that makes some likelihood of acceptable states plausible in general, one would also need to demonstrate that the corresponding history would bear some resemblance to our own history, another strong constraint.

To develop a solution, let us temporarily set aside 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 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. 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.

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 about 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 boarded the morning train and traveled 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:

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. One fact that would bear on his likelihood of taking the train back home is the existence of a return ticket. If we fail to make adjustments in the current state to accommodate a mundane evolution of the counterfactual situation, we have no reason to think he has a return ticket with him, but making such adjustments makes his having a ticket more likely.

**8. The Revised Entropy Theory.** The only plausible emendation to the Entropy Theory is to relax the condition that the microstate outside  $R$  should be fixed exactly. I have no precise proposal for a replacement, but one could adopt the position that the constraint is some best fit of mundaneness and locality that is to be determined on a case by case basis. If some such replacement is suitable, one would not necessarily need to worry about the many disparate microscopic records spreading out into the future from a counterfactual event. Small, localizable counterfactual suppositions might 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 one the freedom to go back and make 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 localized in the man's brain.

The first worry about such a solution is whether it really exists, whether in general there are micro-*A*-states having low entropy origins as well as sufficiently insignificant differences outside *R*. If the counterfactual differences outside *R* one hopes are small are in fact significant, spurious counterfactual dependence between distant events threatens. For such a counterexample, take some historical event *e* for which there are no macroscopic records. If *e* was improbable given physical circumstances shortly before *e* occurred, then the following in most cases turns out incorrectly as highly assertible: "If the world were presently a little different, *e* would not have occurred." Its high assertibility follows from the fact that *e* is improbable to begin with and the likely 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 example vitiates any hope that the *A*-state constrained by the Low Entropy Hypothesis is sufficient to keep all distant, unrelated past facts fixed. The best result one can achieve is that most facts about the past are kept fixed.

A second worry is that the positive achievement of the Entropy Theory in accounting for quasi-chancy counterfactuals would be undermined to some extent. If the *A*-state is not fixed microscopically outside *R*, then even with a deterministic evolution, the outcomes of distant quasi-chancy processes are not held fixed. This problem is not too severe because the under constraining of distant events lends itself to explanation: If events were different now, that would be because events were somewhat different in the past. Because these past differences would have their usual dynamical consequences spreading out in the future, subtle effects may change the outcomes of processes that are very sensitive to microscopic conditions.

A third worry is that by giving up on a strict interpretation of the locality principle, we are 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 ante-

cedent's obtaining, and there may be none. We have no objective way of sorting out where to start modifying the physical state. By relaxing the locality constraint, we lose the main ingredient in our recipe for objective assertibility, the objective probability distribution given by statistical mechanics.

To an optimist, this is something of a success of the Revised Entropy Theory. Backward-looking counterfactuals are hard to evaluate because our intuitions are strained by the piecemeal utilization of conflicting principles. 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 previous quarrel, and if there were no quarrel, Jack would oblige. In other cases, we sometimes struggle to evaluate counterfactuals where there is no clear small region where counterfactual changes first start becoming macroscopic. We then have a match between unclear intuitions about how backward-looking counterfactuals should be evaluated and unclear pronouncements from the Revised Entropy Theory.

Yet, to a sober observer, 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 poses a credible threat of circularity to the Revised Entropy Theory's explanation of the counterfactual asymmetry. To examine this, one needs an account of the closely related influence asymmetry.

**9. The Asymmetry of Influence.** A simple analysis of influence in terms of counterfactual dependence is fairly straightforward. Let  $A$  and  $S$  be about some state of affairs.  $A$  influences  $S$  to the degree that

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

is assertible. This analysis applies to cases where  $S$  is about 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, there may be 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 to 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:

- (1) 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 had better turn out that (1) 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 the change to the guard's brain state 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 (1) 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 almost all be worlds where there is no explosion. The Revised Entropy Theory can account for his not pressing the button in this mundane way, by having 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 (1) 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. 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.

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, in which the counterfactual differences are initially microscopic and localized, form the core that the Revised Entropy Theory successfully explains. So, this argument would go, the influence asymmetry is grounded in these special cases and then 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 the core counterfactuals.

While this strategy may be 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 most actual macroscopic past facts high, but would reduce the probability of some actual macroscopic future facts.

But in order to save the Entropy Theory from the untoward result that it conflicted with the mundaneness principle, 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 influence asymmetry.

Thus, the Revised Entropy Theory at best succeeds in explaining the counterfactual asymmetry in terms of a global entropy asymmetry when the antecedent involves only small changes at a time. 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 and more generally the counterfactual asymmetry, but in effect assumes the asymmetry as a principle of vagueness resolution.

**10. Conclusion.** Counterfactual reasoning about physical affairs contains an intrinsic tension. On one side, *our* interests as reasoning beings establish what hypothetical situations are relevant for consideration, and on the other side, *nature* settles questions about the consequences of our speculation. We need both aspects because we demand both relevance and objectivity. One important goal in theorizing about counterfactuals is to illuminate the contours of this divide. The failure of the Revised Entropy Theory to sustain an adequate explanation of the influence asymmetry demonstrates that even when we allow ourselves the resources of science to refine the context of assertion, the division remains analogous to political boundaries—in places following natural features of the landscape and at other times gerrymandered to serve human interests.

The failure to find a formula to evaluate mundane counterfactuals in a natural way is born from an imperfect marriage of certain key folk intuitions about the way the world works with the physicist's knowledge. Our acceptance of counterfactual asymmetries is to some significant extent guided by a simple model of the world in which nature progresses as a result of individual events. The individuation of events, objects, and processes, is of such obvious utility that it is unsurprising for a folk worldview to presuppose things that can be named and distinguished from one another, and which can be related by resemblance and cause-effect relations, etc. Among the events of most concern to humans are human actions themselves, and we find it easy to think of our own actions as localized causes of changes in the external world. I speculate that because these kinds of relations are so important to us, we take them as central to the working of the world in many contexts and reason counterfactually by trying to picture how collections of localizable events could conspire

to bring about the circumstances we hypothesize. In doing so, we look to resolve the vagueness inherent in our counterfactual speculations, if necessary, by dismissing physically possible scenarios that do not fit this scheme.

The fundamental physics shows only modest respect for such naïve preconceptions of how the world operates, as there is no reasonable candidate for a physical principle that will guarantee that our most reasonable and ordinary counterfactual speculations are physically realizable. The Low Entropy Hypothesis offers some degree of hope that counterfactual affairs can arise via mundane physical evolution in circumstances where those affairs are highly localized, but its prospects for success remain limited and unsubstantiated. The most commonly considered alternative, the postulation of miracles, provides no safe refuge either, for no plausible account yet exists that divulges the crucial details of how miracles can occur in a way that vindicates the critical intuitions.

Even though the Revised Entropy Theory fails to explain the influence asymmetry, there remain some positive features of the theory that recommend its procedure as a reasonable method for the evaluation of counterfactuals involving physical affairs. First, the general procedure that it formalizes has some psychological plausibility. Of course, the recourse to thermodynamic concepts and statistical mechanics in the theory itself bestows considerable psychological implausibility, but the formal procedure does make scientifically respectable a more pedestrian mode of considering counterfactual possibilities: the method of taking the antecedent, filling out the context, and seeing what nature entails or makes probable. There are several psychologically plausible features respected by the Entropy Theory:

(1) We are able to countenance two ways of thinking about the relationship between antecedent and consequent. We can evaluate the counterfactual by considering whether the antecedent's truth in some sense implies or necessitates the consequent's, and we can think of the consequent as being more or less likely to be true, given the truth of the antecedent. The Entropy Theory has the resources to address both cases. More important, in the second case, our thinking about the degree of connection between antecedent and consequent is *entirely independent* of whether the world itself is chancy, and the Entropy Theory respects this judgment.

(2) To the extent that we consider the past in evaluating counterfactuals, we typically think that if the present were different, the differences would arise in more or less ordinary ways. The changes don't arise from miraculous circumstances.

(3) In many contexts, a thinker will imagine counterfactual circumstances coming about by seeking critical events that had they been otherwise, the alternative state of affairs would have ensued. That is, the

thinker intuitively seeks a context where locality and mundaneness are respected.

Second, the Entropy Theory does not identify the truth conditions of the counterfactual with what the laws of nature entail. The most glaring deficiency of Goodman's law-based account and its cousins, are their incapacity to place counterfactuals involving physical happenings properly in the broader context that does not take laws of nature for granted: "If there were no such thing as friction, the brakes wouldn't work." Because the Entropy Theory is merely a theory for generating objective assertibility conditions for mundane counterfactuals, one can appeal to the general principles of similarity that motivate the Stalnaker-Lewis counterfactual logic. The Entropy Theory merely elaborates and constrains the relevant similarity relation in order to vindicate our ordinary practice. Finally, as detailed previously, the Entropy Theory provides the resources to treat successfully a range of counterfactuals involving quasi-chancy phenomena.

The Revised Entropy Theory, then, has several features that recommend it as a plausible way of evaluating counterfactuals. Detracting from its attractiveness is the outstanding problem of properly balancing mundaneness and locality. There are apparently no resources for objectively settling how to relax the mundaneness constraint, and it is highly questionable whether there is any plausible way of doing so. Short of providing some solution, the advocate for the Entropy Theory must attribute the difficulty to the notorious vagueness of counterfactuals and rely on some external resolution. It is unclear whether this weakness is absolutely fatal to the theory as a general procedure for clarifying the proper evaluation of counterfactuals, but regardless, the Revised Entropy Theory fails to produce an adequate explanation for counterfactual asymmetry.

#### REFERENCES

- Adams, Ernest (1975), *The Logic of Conditionals: An Application of Probability to Deductive Logic*. Dordrecht: D. Reidel.
- Albert, David Z. (2000), *Time and Chance*. Cambridge: Harvard University Press.
- Bennett, J. (1984), "Counterfactuals and Temporal Direction", *Philosophical Review* 93: 57–91.
- Downing, P. B. (1959), "Subjunctive Conditionals, Time Order, and Causation", *Proceedings of the Aristotelian Society* 59: 125–40.
- Goodman, Nelson (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.
- Lewis, David (1973a), *Counterfactuals*. Oxford: Blackwell.
- (1973b), "Counterfactuals and Comparative Possibility", *Journal of Philosophical Logic* 2: 418–46. Reprinted in W. L. Harper, R. Stalnaker, and G. Pearce (eds.) (1981), *Ifs*, Dordrecht: D. Riedel, 57–85.
- (1979), "Counterfactual Dependence and Time's Arrow", *Noûs* 13: 455–76.
- Skyrms, Brian (1994), "Adams Conditionals", in E. Eells, and B. Skyrms (eds.), *Probability and Conditionals*. Cambridge: Cambridge University Press, 13–26.

Stalnaker, Robert (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.) (1975), *Causation and Conditionals*. Oxford: Oxford University Press, 165–79, and in W. L. Harper, R. Stalnaker, and G. Pearce (eds.) (1981), *Ifs*. Dordrecht: D. Riedel, 41–55.