

Causality and Maximum Entropy Updating

Daniel Hunter

Northrop Research and Technology Center

ABSTRACT

This paper examines an objection to maximum entropy updating and argues that the problem arises from an inadequate representation of causal information. The objection is that maximum entropy updating renders probabilistically dependent previously independent events when probabilistic information about an effect of the two events is presented. It is believed by many that such information should not render the events dependent. This paper accepts the view that independence should be preserved by maximum entropy updating, but argues that it indeed will be when the causal information is presented in an appropriate form. It is argued that presenting the causal information in the form of conditional probabilities is inappropriate. An alternative way of presenting such information, in terms of probabilities of statements known as "counterfactual conditionals," is described. It is shown that when the causal information is expressed by counterfactual conditionals, maximum entropy updating produces results that agree with intuitions shared by its critics and defenders alike about how such information should affect probabilities. An efficient algorithm is given for updating causal information in the form of probabilities of counterfactuals. Finally, the theory of probabilistic counterfactuals developed in this paper is applied to the interpretation of empirical results concerning the way in which people reason under uncertainty.

KEYWORDS: *maximum entropy inference, minimum information updating, counterfactuals, causal reasoning, uncertain reasoning*

INTRODUCTION

Maximum entropy updating (also known as "minimum information updating") is a technique for the revision of probabilistic beliefs that has received a fair amount of attention from the uncertain reasoning community in recent years. Some have argued that maximum entropy updating (known hereafter as

Address correspondence to Daniel Hunter, Northrop Research and Technology Center, One Research Park, Palos Verdes Peninsula, California 90274.

International Journal of Approximate Reasoning 1989; 3:87-114
© 1989 Elsevier Science Publishing Co., Inc.
655 Avenue of the Americas, New York, NY 10010 0888-613X/89/\$3.50

MAXENT) is unnecessary, on the grounds that standard Bayesian conditioning will always give the same result when appropriate evidence events are introduced. A stronger objection to MAXENT is that it gives the wrong answer for certain problems. Judea Pearl, for example, has posed a challenge to MAXENT of this sort. He has given what he considers a counterexample to MAXENT in which MAXENT makes events dependent that should intuitively be independent. This paper argues for three claims about this putative counterexample: (1) that the problem arises from an inadequate representation of causal information; (2) that causal information can be formulated in terms of probabilities of certain types of statements called “counterfactuals”; and (3) that MAXENT gives the intuitively correct answer when given this causal information in the form of probabilities of counterfactuals.

The next section explains MAXENT. The third section describes Pearl’s counterexample to MAXENT. In the fourth section, counterfactuals are introduced, and their connection to causal notions is explained. The fifth section puts counterfactuals into a probabilistic framework and states a theorem concerning the application of MAXENT to probabilities of counterfactuals, a theorem that shows Pearl’s objection to MAXENT to be unfounded and in addition suggests an algorithm for efficient MAXENT updating on causal information. The sixth section discusses an attempt to resurrect Pearl’s puzzle and concludes that the attempt fails, the lesson being that our intuitions about dependency or independency of events cannot always be trusted. The penultimate section applies the theory of probabilistic counterfactuals developed in earlier sections to an explanation of some psychological results on how people actually reason about probabilities. The final section summarizes the argument of this paper, deals with possible objections, and notes some limitations of the approach taken.

Although Pearl’s objection to MAXENT does not succeed, it does show the importance of causal notions in probabilistic reasoning. (Pearl himself has led the way in understanding the role of causal notions in uncertain reasoning. See Pearl [15].) One practical upshot of this point is that designers of expert systems that deal with uncertainty must be aware of the causal structure of the problem domain and must be careful how they represent this causal structure. It will be shown, for example, that representing casual relations between events by means of conditional probabilities over those events is incorrect.

MAXIMUM ENTROPY UPDATING

MAXENT allows one to make estimates of probabilities on the basis of incomplete probabilistic information. A system for medical diagnosis, for example, rarely has sufficient data to specify the probability of a disease for all possible combinations of symptoms. Typically, there is statistical evidence of

correlations between a given disease and only a small subset of the possible symptoms. Moreover, some samples may be too small to support valid statistical generalizations. Nonetheless, the available data are not completely uninformative, and some means should be found to make use of them. MAXENT is a technique for extrapolating from partial probabilistic data to a total probability function. Thus in the medical diagnosis case, MAXENT will yield an estimate of the probability of each disease on each combination of symptoms, even when direct statistical data for this probability are missing. Obviously the more statistical data available, the better MAXENT estimates the overall probability distribution. However, reasonable estimates can be made with partial data, and these estimates can be updated as additional information accrues.

The known probabilistic data act as constraints on the set of permissible probability distributions. However, because the data are incomplete there are many, perhaps an infinite number, of probability distributions that satisfy those constraints. MAXENT picks from among the distributions satisfying those constraints the one that adds the minimal amount of information (relative to the prior probability distribution), in the precise sense of "information" defined by the mathematical theory of information. The rationale is that this is the least biased of all the distributions satisfying the constraints, since it goes beyond the information present in the constraints and in the prior distribution to a minimal degree. The entropy of a probability distribution is inversely related to the degree of information in it, and hence maximizing entropy is equivalent to minimizing information.

These notions can be made more precise as follows. Let $P(\cdot)$ and $Q(\cdot)$ be probability functions over the same space $S = \{x_1, \dots, x_n\}$ (We assume throughout that the probability space is finite). The entropy of Q relative to P , written $H(Q, P)$, is defined as

$$H(Q, P) = - \sum_{i=1}^n Q(x_i) \log \left[\frac{Q(x_i)}{P(x_i)} \right]$$

The quantity $-H(Q, P)$ is known variously as the *discrimination information*, *directed divergence*, *I divergence*, *Kullback-Leibler number*, and *cross-entropy*, of $Q(\cdot)$ with respect to $P(\cdot)$.¹ $-H(Q, P)$ is often regarded as a measure of the "divergence" of $Q(\cdot)$ from $P(\cdot)$.

¹ To avoid confusion, the reader should be aware that some authors (e.g., Shore and Johnson [2]) use the term *relative entropy* synonymously with the terms just mentioned. Thus their relative entropy is the negative of relative entropy as defined in this paper. Other authors, however (e.g., Van Campenhout and Cover [3]), define the term *relative entropy* in a manner consistent with the present definition. I prefer the definition given in this paper because the term *maximum entropy inference* is well established and denotes an inference method that is equivalent to maximum relative entropy updating, as here defined, when the prior is uniform. Thus maximum relative entropy updating seems to be the natural generalization of maximum entropy inference to cases in which prior information is available.

The principle of maximum (relative) entropy, or MAXENT for short, then says that given a prior probability function $P(\cdot)$ over a space and new information in the form of constraints on the posterior for the same space, the posterior should be estimated by that probability function $Q(\cdot)$ which satisfies the constraints and whose entropy relative to $P(\cdot)$ is maximal. Shore and Johnson [2] give an axiomatic derivation of this principle and show that when the constraints are consistent there is a unique maximum entropy posterior. The use of MAXENT as a general inference principle was first proposed by Jaynes [4], and its use in expert systems explored by Cheeseman [1], Lemmer and Barth [5], Hunter [6], and others.

The constraints for MAXENT are statements that pick out a subset of the set of all possible probability distributions over S . Typically, one is concerned with constraints of the form

$$\sum_{i=1}^n a_i P(x_i) = p, \quad a_i, p \in \mathbb{R}$$

which determine a closed, convex set of probability distributions.

For our purposes, we may further restrict the constraints to those in which the coefficients are either zero or one. Hence a constraint is simply an assignment of probability to some subset of the set of primitive events.

The problem of maximizing H subject to such constraints is a familiar optimization problem: Maximize a nonlinear function subject to a set of linear constraints. Lemmer [7] gave an efficient algorithm for maximum entropy updating, and Lemmer and Barth [5] applied this algorithm to updating in expert systems.

Lemmer's algorithm is the following: Let the constraints for updating be $\text{Prob}(X_i) = p_i$, $i = 1, \dots, k$, where $\{X_1, X_2, \dots, X_k\}$ forms a partition of the probability space S (i.e., $X_i \cap X_j$ is empty for $i \neq j$ and $X_1 \cup X_2 \cup \dots \cup X_k = S$). If $P(\cdot)$ is the current probability function over S , update $P(\cdot)$ by multiplying the probability of each member of X_i by $p_i/P(X_i)$. It should be noted that this updating algorithm is essentially Jeffrey's rule, proposed by Jeffrey in Ref. 8. However, at the time Jeffrey proposed his rule, he was apparently unaware that it is a form of maximum entropy updating. That the updating algorithm just described gives the maximum entropy update was proved by Lemmer [7].

Lemmer's algorithm can be extended to the case of updates on multiple partitions. The procedure is to apply the Lemmer algorithm to the constraints for each partition in succession, iterating this process until all the constraints are simultaneously satisfied. Results of Csiszar [9] guarantee that if the maximum entropy distribution for the given set of constraints exists, then this iterative procedure will converge to it.

PEARL’S PUZZLE

Judea Pearl has given what he considers a counterexample to MAXENT.² I consider it a puzzle, rather than a counterexample, and so will refer to it by the title “Pearl’s puzzle.” There is more than one version of Pearl’s puzzle, but the version given in this section captures all the essential ingredients.

The puzzle is this: Suppose you are told that three individuals, Albert, Bill, and Clyde, have been invited to a party. You know nothing about the propensity of any of these individuals to go to the party nor about any possible correlations among their actions. Using the obvious abbreviations, consider the eight-point space consisting of the events ABC , $AB\bar{C}$, $A\bar{B}C$, etc. (conjunction of events is indicated by concatenation). With no constraints whatsoever on this space, MAXENT yields equal probabilities for the elements of the space. Thus $\text{Prob}(A) = \text{Prob}(B) = 0.5$ and $\text{Prob}(AB) = 0.25$, so A and B are independent. It is reasonable that A and B turn out to be independent, since there is no information that would cause one to revise one’s probability for A upon learning what B does. However, suppose that the following information is presented: Clyde will call the host before the party to find out whether Al or Bill or both have accepted the invitation, and his decision on whether to go to the party will be based on what he learns. Al and Bill, however, will have no information about whether or not Clyde will go to the part. Suppose, further, that we are told the probability that Clyde will go conditional on each combination of Al and Bill’s going or not going. For the sake of specificity, suppose these conditional probabilities are the ones given in Table 1.

When MAXENT is given these constraints, the result is the new probability assignment

Event	ABC	$AB\bar{C}$	$A\bar{B}C$	$A\bar{B}\bar{C}$	$\bar{A}BC$	$\bar{A}B\bar{C}$	$\bar{A}\bar{B}C$	$\bar{A}\bar{B}\bar{C}$
Probability	0.0197	0.177	0.1422	0.1422	0.1422	0.1422	0.1876	0.0469

from this table, we may calculate that $\text{Prob}(A) = 0.4811$ and $\text{Prob}(A|B) = 0.4089$. A and B are no longer independent! But this seems wrong: The information about Clyde should not make A ’s and B ’s actions dependent.

The idea that the information about Clyde is irrelevant to whether or not A and B are dependent should not be confused with the claim that conditionalizing on what Clyde does should leave A and B independent. Clearly it should not, because if we know that Clyde went to the party, then the additional information that Bill went should substantially lower the probability that Al went. The

² The example was given by personal communication and has been floating around the uncertain reasoning community for some time. Pearl informs me that the example was discovered by Norman Dalkey but was first taken as a counterexample to MAXENT by Pearl.

Table 1.

	$P(C \cdots)$
AB	0.1
$A\bar{B}$	0.5
$\bar{A}B$	0.5
$\bar{A}\bar{B}$	0.8

intuition is rather that simply presenting information about how Clyde's behavior is dependent upon Al's and Bill's behavior, without giving any information about what Clyde actually does, should not cause us to change our probabilities for what Al and Bill do.

The intuition that independence of A and B should be preserved when the information about Clyde is given can be strengthened by supposing that A and B are two different coins that Clyde flips and that Clyde uses the outcome of the two flips to pick a third coin with a particular bias, the outcome of whose flipping determines whether or not Clyde goes to the party. For example, if both A and B come up heads, this tells Clyde to pick a coin biased 9:1 in favor of tails, and Clyde will go to the party if and only if this third coin comes up heads, and so on. Here it seems clear that the information about Clyde's method of determining whether or not to go to the party should not make the flips of A and B dependent. This example differs in one important respect from the first one, however: In this example, independence is built in—we know that the outcomes of flipping two distinct coins are independent. In the first example, though, it was lack of information about any dependence between Al's and Bill's actions that led to probabilistic independence. Thus it would be consistent with the first example to later discover that Al and Bill are roommates whose actions are probabilistically dependent. In what follows, the focus will be on probabilistic independence stemming from lack of information about dependence.

Pearl's concern is over the disappearance of independence when information about an effect of two previously independent events is presented. However, the puzzle is really wider than that: A stronger claim is that information about an effect of two events should not change the marginal prior over those two events. But as can be seen from the tabulated probabilities, MAXENT applied to the constraints of Table 1 does change the prior over the events A and B .

COUNTERFACTUALS

Pearl presents his puzzle as a counterexample to MAXENT; his conclusion is that MAXENT gives the wrong answer. However, it is important to realize that a similar puzzle can be generated for other methods of updating. Bayesian

updating, for example, is subject to the same puzzle.³ Modify the example so that our information about Clyde is that he will certainly not go to the party if both Al and Bill go. This information can be represented by the formula

$$\text{Prob}(\bar{C}|AB) = 1$$

which is equivalent to the formula

$$\text{Prob}(\overline{ABC}) = 1$$

This allows Bayesian updating to be performed by conditioning on the statement \overline{ABC} . The result is the following probability assignment:

Event	ABC	$ABC\bar{C}$	$A\bar{B}C$	$A\bar{B}\bar{C}$	$\bar{A}BC$	$\bar{A}\bar{B}C$	$\bar{A}\bar{B}\bar{C}$	$\bar{A}BC\bar{C}$
Probability	0	1/7	1/7	1/7	1/7	1/7	1/7	1/7

In this assignment, $\text{Prob}(A) = 3/7$, but $\text{Prob}(A|B) = 1/3$, so A and B are dependent in the posterior.

The fact that Bayesian updating is subject to the same paradox should temper the inclination of some to put the blame on MAXENT. We need to stand back and take a fresh look at what's going on here. Maybe the problem is not in the method of updating used but resides somewhere else.

One place to start looking for the problem is in the representation of the information. Do the conditional probabilities capture the information present in the examples? The example just given would suggest not. For the statement " $\text{Prob}(\bar{C}|AB) = 1$," being equivalent to " $\text{Prob}(\overline{ABC}) = 1$," merely says that not all of A , B , and C came to the party. The information that C 's behavior is dependent upon that of A and B is missing. In other words, there is important causal information that is not captured by the conditional probability.

That conditional probabilities do not capture the direction of causality can also be seen by supposing that \bar{C} stands for a causal factor in a disease (e.g., "absence of calcium") and A and B stand for symptoms of the disease (e.g., "anemia" and "brittleness of the bones"). Then " $\text{Prob}(\bar{C}|AB) = p$ " would represent the probability of a cause given its effects. In such a case, it is no longer obvious that independence of A and B should be preserved.

Thus it is not surprising that MAXENT gives the wrong answer; information essential to getting the right answer has not been provided. Now the problem is how to provide that information. We have seen that conditional probabilities do not capture causal information. I suggest that one way to capture the required information is to employ counterfactual conditionals. A *counterfactual conditional* is a statement of the form "If A were the case, then C would be the case." Theories of counterfactuals have been extensively debated and discussed

³ This is not intended as a criticism of Bayesian updating. The only point is that a naive application of Bayesian updating produces the same counterintuitive result as does a naive application of MAXENT.

in the philosophical logic literature, and counterfactuals have recently come to the attention of the AI community (e.g., see Ginsberg [10]), as it has become apparent that they have important connections to such issues as nonmonotonic reasoning, belief revision, and, most important for the topic of this paper, causal reasoning.

Stalnaker's Theory of Counterfactuals

This section describes a standard formal theory of counterfactuals developed by Stalnaker [11]. Stalnaker's theory is later extended to a probabilistic framework.

Stalnaker's theory is posed in terms of the notion of possible worlds. A possible world is a maximally specific way the world might be. By "maximally specific" is meant that for any proposition and possible world, either the proposition is true in the world or it is false in that world, that is, the world has to specify everything. For our purpose we may take the notion of possible worlds to be relative to a language. If our language contains the atomic propositions P_0, P_1, P_2, \dots , a possible world may be thought of as a specification of the truth value of each of the P_i . If the number of atomic propositions is finite—of size n , for n a positive integer—then the number of worlds is also finite of size 2^n . I make the simplifying assumption in what follows that the number of atomic propositions is finite. This is not an unreasonable assumption in the context of a computer implementation of possible worlds semantics. Another assumption that will be made, in keeping with traditional possible worlds semantics, is that propositions can be represented by sets of possible worlds; a proposition is identified with the set of worlds in which it is true. Note that there is a distinction between propositions and sentences: propositions are nonlinguistic entities; sentences are linguistic entities that express propositions. I will sometimes blur this distinction in what follows when nothing of importance turns on it (e.g., I will sometimes speak of the "antecedent" of a counterfactual when I mean "the proposition expressed by the antecedent").

The intuition behind Stalnaker's is that the counterfactual $A \Box \rightarrow C$ is adjudged true if and only if the most similar world in which A is true is a world in which C is true. For example, to determine whether or not the counterfactual "If the match were struck, it would light" is true, we consider the world most similar to the real world in which the match is struck and see whether or not in that world it lights. If it does, we judge the counterfactual true; otherwise, we say it is false. The notion of similarity here is deliberately vague; different notions of similarity lead to different ways of evaluating counterfactuals. The notion of similarity to a given world is often linked to the notion of a "minimal revision" of a world. In this sense of similarity, the most similar world in which A is true is the one that results from adding A to the truths about the real world and then making a minimal revision of this set of statements so that A can be

consistently maintained. Again, there may be different views as to what counts as a “minimal” revision, and different views produce different ways of evaluating counterfactuals.

What matters for our purposes is that the *formal* theory of counterfactuals, the *logic* of counterfactuals, may remain the same regardless of how we flesh out the notion of similarity or minimal change. Stalnaker’s formal theory begins by assuming that we are given a set W of possible worlds and a function f , called a *selection function*, that takes as arguments a proposition and a possible world and yields as value a possible world (intuitively, the most similar world to the argument world in which the proposition is true). For all worlds w and propositions A and B , f is stipulated to satisfy:

- (i) A is true in $f(A, w)$, provided A is logically consistent.
 - (ii) If A is true in w , then $f(A, w) = w$.
 - (iii) If A is true in $f(B, w)$ and B is true in $f(A, w)$, then $f(A, w) = f(B, w)$.
- Stalnaker’s definition of the counterfactual is:

$A \Box \rightarrow C$ is true at world $w =_{df}$. If A is consistent,
then C is true at $f(A, w)$.

Intuitively, the selection function selects, for each proposition and world, the most similar world to the given world in which the proposition is true. The above definition therefore says that a counterfactual is true if its consequent is true at the most similar world at which its antecedent is true, if there is such a world. If the antecedent is inconsistent (so there is no most similar world at which it is true), then the counterfactual is vacuously true.

Stalnaker’s theory of counterfactuals has enjoyed wide acceptance. Alternative theories have been developed, but the basic framework is still the one laid out by Stalnaker. One subject of dispute concerns the existence and uniqueness of a most similar world in which a given proposition is true. Stalnaker assumes that there is a unique most similar world, provided the proposition is consistent; others do not. Lewis ([12], pp. 19–21 and 77–83), for example, argues that there may be more than one maximally similar world in which a given proposition is true or there may be no such world even though the proposition is consistent.

Modifications can be made to Stalnaker’s theory to accommodate these varying intuitions. However, these modifications would not materially change the conclusions reached below and would make the analysis more difficult. For this reason, we will stick with Stalnaker’s theory in its original form.

Counterfactuals and Causality

Many philosophers have recognized the connection between counterfactuals and causality (see the articles in Sosa [13], especially the one by Lewis [14]). The truth of the counterfactual “If A were to occur, then B would occur,” when neither A nor B in fact occurs, suggests a causal connection of some sort or

another between A and B ; exactly what sort of causal connection is a subject of spirited debate. The indicative conditional “If A occurs, B occurs” does not seem to capture the causal connection between A and B : A might be the effect and B the cause, as in “If there’s smoke, there’s fire.” Indicative conditionals often express *evidential*, rather than *causal*, relations between propositions; for example, a detective in a murder case might say “If the butler didn’t do it, the maid did” without implying a causal connection between what the butler did (or did not do) and what the maid did. Note that if the butler really did commit the murder, it would probably be false to say “If the butler had not done it, the maid would have” (a counterfactual), but, in the absence of knowledge as to the murderer’s identity, the indicative “If the butler didn’t do it, the maid did” is a perfectly reasonable statement to make.

Nor do conditional probabilities capture causal connections between events in the way counterfactuals do. The probability that the maid committed the murder, conditional on the butler’s not having committed the murder, might be high without the butler’s having failed to commit the murder being a cause of the maid’s committing the murder. However, if the counterfactual “If the butler had not done it, the maid would have” has a high probability, this would indicate some kind of causal connection between the butler’s lack of action and the maid’s action, such as a prior arrangement between butler and maid that if one of them is unable to pull off the murder the other is to do the deed.

Returning to the original problem, the suggestion is that the relations between Al’s and Bill’s actions on the one hand and Clyde’s on the other are expressible as counterfactual conditionals, that there is a certain probability that if Al and Bill *were* to go to the party, then Clyde *would not* go, and so on. The information to MAXENT should be probabilities of counterfactuals rather than conditional probabilities. However, this raises the question of how probabilities of counterfactuals are to be represented, which is the topic of the next section.

PROBABILITY MEASURES OVER COUNTERFACTUALS

One way to attach probabilities to counterfactuals would be to consider a probability function $P(\)$ over the set of worlds W and take the probability of the counterfactual $A \Box \rightarrow C$ to be $\sum_{w \in W} P(w) \chi(A \Box \rightarrow C, w)$, where $\chi(B, w)$, for B a proposition and w a world, is one if B is true in w and zero otherwise. This method has the drawback that it does not allow uncertainty about counterfactuals when uncertainty about worlds has been removed. For example, let S stand for “The match is struck” and L for “The match lights.” Consider the set W of worlds $\{SL, S\bar{L}, \bar{S}L, \bar{S}\bar{L}\}$ and a particular selection function f . Suppose it is known that the match is not struck and that it is not lit. Then $P(S\bar{L}) = 1$. What about $P(S \Box \rightarrow L)$? By the method for computing probabilities of counterfactuals above, this probability will be either zero or one. Yet certainty about the

actual state of affairs should be compatible with uncertainty about what would happen if the match were struck.

It should be noted that this problem arises because we have chosen to identify possible worlds with possible combinations of truth values of atomic formulas. In a more abstract formulation, it could be left open whether or not worlds that agree on all the atomic formulas are the same world. If we distinguish worlds not just in terms of which atomic formulas are true in them, but in terms of which counterfactuals are true in them, then the problem mentioned above does not arise. However, such a method would exponentially increase the number of worlds, and so for reasons of computational efficiency we will stick to identifying worlds with combinations of truth values of atomic formulas.

The problem is that we are working with a fixed selection function f . We need some way of considering different possibilities for the selection function so as to reflect the uncertainty about counterfactuals that does *not* stem from uncertainty about worlds.

The most direct way to allow variability in the selection function is to take the probability space to be $W \times F$, where F is the set of all possible selection functions over W . However, we choose the more indirect route of considering the set of all linear orderings of members of W . Call this set Ω . Thus an arbitrary member of Ω is given by a sequence $\langle w_1, w_2, \dots, w_k \rangle$ of members of W such that there are no repetitions in the sequence and each member of W occurs in the sequence. We will show how both W and F can be extracted from Ω .

For each s in Ω define:

$A \Box \rightarrow C$ is true at s iff (i) A is logically impossible or (ii) if w is the first element of s such that A is true at w , then C is true at w .

Note that if A is a tautology then the above definition implies that $A \Box \rightarrow C$ is true at s if and only if C is true at the first member of s . Since $A \Box \rightarrow C$ is equivalent to C when A is a tautology, this means that truth of a noncounterfactual sentence at a sequence is truth at the first member of that sequence. Therefore a probability measure over Ω induces a probability measure over W by taking the probability of a world w to be the sum of the probabilities of all sequences in which w is the first member.

Each sequence represents an ordering of the possible worlds in terms of their similarity to the first world in the sequence. Since the same world may be the initial member of different sequences, this provides a means for representing different selection functions. The well-known fact that a selection function can be represented by a linear ordering of worlds follows from the properties of the selection function. Clearly a linear ordering of worlds determines $f(A, w)$ for w the first member of the ordering—just take $f(A, w)$ to be the first world in the ordering at which A is true (or undefined if A is inconsistent). Conversely, a selection function $f(\)$ determines for each world w a linear ordering of worlds with w as first member of the ordering, as follows: For $u, v, w \in W$, define u

$\leq v$ by $f(\{u, v\}, w) = u$. $\{u, v\}$ is the proposition true just at the worlds u and v . The properties of $f(\)$ ensure the truth of the following theorem:

THEOREM 1. $u \leq_w v$ is a linear order with w as least element.

See the appendix for a proof of Theorem 1.

It can easily be shown that if selection function f induces \leq_w , then $A \Box \rightarrow C$ is true at w relative to f if and only if either A is impossible or else the first u under the ordering \leq_w at which A is true is a world at which C is true. Hence the definition above is consistent with the original Stalnaker definition of the counterfactual.

We may now define a probability measure over Ω . Since counterfactuals are defined over Ω , we can now meaningfully speak of the probability of counterfactuals.

Returning to our original problem, the causal information in Pearl's counterexample to MAXENT can be expressed in terms of counterfactuals: Instead of the conditional probability $\text{Prob}(\bar{C}|\bar{A}\bar{B}) = p$ we use the counterfactual probability $\text{Prob}(AB \Box \rightarrow \bar{C}) = p$. In the most general case, we would have a number of such counterfactual probabilities, one for each combination of causal factors on which C depends.

The issue is whether or not giving MAXENT the information in the form of counterfactual probabilities results in an answer that agrees with the intuitions previously expressed about Pearl's example. Happily, the next theorem says that it does. Theorem 2 shows that if MAXENT is given constraints in the form of probabilities for counterfactuals saying what would happen were a certain combination of causal factors to obtain, then it will leave the probability distribution over the causal factors unchanged, changing only the conditional probabilities of the effect given the causes. Hence if the causal factors were independent before the application of MAXENT, they will still be independent after its application.

To state Theorem 2, we need the following definitions. Let A_1, \dots, A_k be causal factors for C . Let W be the set of all combinations of truth values of $\{A_1, \dots, A_k, C\}$, and let Ω be the set of all linear orderings of elements of W . For each possible combination of truth values of the A_i , there is a formula that is true if and only if the A_i have those truth values, namely the conjunction in which A_i occurs as a conjunct if A_i is true and \bar{A}_i occurs if A_i is false (e.g., if all the A_i are true, the corresponding formula would be " $A_1 A_2 \dots A_k$," the conjunction of all the A_i). Call such a formula an *atom* over $\{A_1, \dots, A_k\}$. Let S_1, \dots, S_r be all the atoms over A_1, \dots, A_k . The S_i stand for the various combinations of causal factors. Let $P(\)$ be a probability measure over $\{S_1, \dots, S_r\}$ and extend $P(\)$ to Ω by defining, for each $s \in \Omega$, $P(s)$ to be $P(S_i)/N$, where S_i is true in s and N is the number of members of Ω in which S_i is true. Finally, where $R(\)$ is a probability measure and Γ is a set of probability constraints, let $\text{MAXENT}(R, \Gamma)$ be the posterior probability measure that results from applying MAXENT to

$R()$ with the members of Γ as constraints. Then the following theorem holds:

THEOREM 2. *Let $\Gamma = \{ \text{Prob}(S_i \square \rightarrow C) = p_i : i = 1, \dots, r \}$. Then if $Q = \text{MAXENT}(P, \Gamma)$, Q satisfies (i) $Q(S_i) = P(S_i)$ and (ii) $Q(C|S_i) = p_i$.*

Clause (i) of Theorem 2 says that applying MAXENT to the counterfactuals in Γ leaves unchanged the marginal prior over the S_i —that is, over the causal factors. Clause (ii) says that the probability of the effect, conditional on a particular combination of causal factors, is equal to the probability of the counterfactual that says that if that combination of causal factors were to obtain then the effect would obtain.

These two facts completely determine the posterior distribution over W . Thus they imply that the updating could be performed entirely within W without invoking the sequence space Ω . This is a relief from a computational point of view, since the size of the sequence space is 2^n , where n is the number of atomic propositions. This number is ridiculously large when there are more than a few atomic propositions (the exponential size of W in the number of atomic propositions is bad enough). Fortunately, a simple modification of the Lemmer algorithm allows updating to be done in the small space W : To update on the constraint $\text{Prob}(S_i \square \rightarrow C) = p_i$ with respect to probability function Q_{i-1} over W , multiply the probability of each $S_i C$ world by $p_i/Q_{i-1}(C|S_i)$ and the probability of each $S_i \bar{C}$ world by $(1 - p_i)/Q_{i-1}(\bar{C}|S_i)$. Leave the probability of all other worlds unchanged. This updating method reaches the same posterior over W without going outside of W as would be achieved by applying MAXENT to Ω and taking the marginal over $\{A_1, \dots, A_k, C\}$.

To illustrate the modified Lemmer algorithm, suppose we replace the conditional probabilities in Table 1 with the corresponding counterfactual probabilities. Then the constraints become

$$\text{Prob}(AB \square \rightarrow C) = 0.1$$

$$\text{Prob}(A\bar{B} \square \rightarrow C) = 0.5$$

$$\text{Prob}(\bar{A}B \square \rightarrow C) = 0.5$$

$$\text{Prob}(\bar{A}\bar{B} \square \rightarrow C) = 0.8$$

Assume that the prior $P()$ on the space of atoms over $\{A, B, C\}$ is uniform (thus each of the eight points in this space has probability 0.125). To update on the first constraint, we multiply the probability of the ABC world by

$$\text{Prob}(AB \square \rightarrow C)/P(C|AB) = 0.1/0.5 = 0.2$$

to get 0.025 and then multiply the probability of the ABC world by 0.9/0.5 to get 0.225. The remaining constraints are handled in the same fashion. The resulting posterior distribution is shown.

Event	ABC	$ABC\bar{C}$	$A\bar{B}C$	$A\bar{B}\bar{C}$	$\bar{A}BC$	$\bar{A}\bar{B}C$	$\bar{A}\bar{B}\bar{C}$	$\bar{A}\bar{B}\bar{C}$
Probability	0.025	0.225	0.125	0.125	0.125	0.125	0.2	0.05

Note that with only the counterfactual information as constraints and no prior information about the causal factors A and B , MAXENT yields a uniform distribution for the marginal over A, B , thereby making A and B independent. Thus MAXENT satisfies our intuitions about this case.

Generating Dependence

It is interesting to consider what would happen were we to learn of some common causal factor for A and B . Imagine that we learn that Al and Bill are both enamored of a certain lady whom we will call Diane. If Diane goes to the party, there is a high probability that Al will go and a high probability that Bill will go; if she does not go, then the probability of either of them going is low. Let's say that our information about Diane's influence over Al and Bill is captured by the following probabilistic counterfactuals:

$$\text{Prob}(D \Box \rightarrow A) = \text{Prob}(D \Box \rightarrow B) = 0.9$$

and

$$\text{Prob}(\bar{D} \Box \rightarrow A) = \text{Prob}(\bar{D} \Box \rightarrow B) = 0.1$$

Then applying MAXENT to the uniform prior with the above counterfactuals as constraints results in the following distribution over the atoms of $\{A, B, D\}$

Event	ABD	$AB\bar{D}$	$A\bar{B}D$	$A\bar{B}\bar{D}$	$\bar{A}BD$	$\bar{A}\bar{B}D$	$\bar{A}\bar{B}\bar{D}$	$\bar{A}\bar{B}\bar{D}$
Probability	0.405	0.005	0.045	0.045	0.045	0.045	0.005	0.405

Note that these probabilities result from the application of the modified Lemmer algorithm first to the uniform distribution over the atoms of $\{A, B, C\}$ with constraints $\text{Prob}(D \Box \rightarrow A) = 0.9$ and $\text{Prob}(\bar{D} \Box \rightarrow A) = 0.1$ and then to the new distribution with constraints $\text{Prob}(D \Box \rightarrow B) = 0.9$ and $\text{Prob}(\bar{D} \Box \rightarrow B) = 0.1$.

From the probabilities tabulated above, we can calculate that $\text{Prob}(A) = \text{Prob}(B) = 0.5$ and $\text{Prob}(B|A) = 0.82$, so that A and B are rather strongly dependent in the MAXENT distribution. This is as it should be, for learning of the existence of B is very good evidence that the cause D obtains and the obtaining of D is good evidence for the obtaining of A . However, A and B become independent when the cause is known with certainty. That is, A and B are *conditionally* independent under both D and \bar{D} , for from the above values we have

$$\text{Prob}(B|AD) = \text{Prob}(B|D) = 0.9$$

and

$$\text{Prob}(B|A\bar{D}) = \text{Prob}(B|\bar{D}) = 0.1$$

It is often taken as a mark of one event's being a cause of others that knowledge of the occurrence or nonoccurrence of that event makes the other events probabilistically independent. When the links in Pearl's Bayesian networks [15] are interpreted causally, the algorithm for computing probabilities for such a network embodies the assumption that effects are conditionally independent with respect to their causes. Whether this assumption is in general true is doubtful. However, in the special case in which our knowledge concerns only the probabilistic relation between the cause and each of its effects singly, the above result shows this assumption to be justified.

PARADOX REGAINED? OR A LESSON IN CAUTION

The previous section showed that MAXENT agrees with our intuitions about independence of causes when the information is presented in the form of counterfactuals whose antecedents are mutually exclusive ways in which the causal factors might be combined. However, there are ways of formulating the causal information that render the causes dependent. In particular, if the causal factors mentioned in the antecedents of the counterfactuals are not mutually exclusive, then they may become dependent in the posterior resulting from updating on those counterfactuals. For example, consider the probability statements

$$\text{Prob}(A \Box \rightarrow C) = 0.9 \qquad \text{and} \qquad \text{Prob}(B \Box \rightarrow C) = 0.8$$

We might learn, for example, that it is 90% certain that if Al were to go to the party then Clyde would go too and 80% certain that Clyde would go were Bill to go. If we apply MAXENT to these probability statements (starting with a uniform prior), we end up with the following marginal over *A*, *B*, and *C*.

Event	<i>ABC</i>	<i>ABC̄</i>	<i>ĀBC</i>	<i>ĀB̄C̄</i>	<i>̄ABC</i>	<i>̄ĀB̄C̄</i>	<i>̄ĀB̄C̄</i>	<i>̄ĀB̄C̄</i>
Probability	0.295	0.015	0.202	0.028	0.168	0.062	0.115	0.115

From this table we may calculate that the probability of *A* is approximately 0.54, but the probability of *A* conditional on *B* is about 0.57—a small difference, but a difference nonetheless. It might be thought that this result resurrects Pearl's puzzle; for why should the information given produce even a small correlation between *A* and *B*?

As before, however, the suspicion that MAXENT is to blame can be laid to rest by considering what would happen if Bayesian conditioning were used

instead. Suppose the constraints this time are

$$\text{Prob}(A \Box \rightarrow C) = 1 \quad (1)$$

and

$$\text{Prob}(B \Box \rightarrow C) = 1 \quad (2)$$

That is, we learn that C would certainly go if A were to go and C would certainly go were B to go. We can therefore use Bayesian conditioning to update on these two counterfactuals, and the resulting marginal over A , B , and C is

Event	ABC	$ABC\bar{C}$	$A\bar{B}C$	$A\bar{B}\bar{C}$	$\bar{A}BC$	$\bar{A}B\bar{C}$	$\bar{A}\bar{B}C$	$\bar{A}\bar{B}\bar{C}$
Probability	0.4	0	0.2	0	0.2	0	0.1	0.1

here $\text{Prob}(A) = \text{Prob}(B) = 0.6$, but $\text{Prob}(A|B) = \text{Prob}(B|A) = 0.666 \dots$. So in this case also there is a slight correlation between A and B .

We face a conflict between intuition and the results of an analysis in terms of counterfactuals. As in the case of the previous analysis in terms of conditional probabilities, we have the escape route of saying that the analysis in question somehow misrepresents the given information. But in this case it is not at all clear what is an alternative to the counterfactual analysis. Perhaps a different escape route can be found; intuitions can be mistaken. Is there any reason to think our intuitions have misled us in this case?

Yes, there is. Consider the following pair of counterfactuals:

$$A \Box \rightarrow C \quad \text{and} \quad B \Box \rightarrow \bar{C}$$

The first says that if Al were to go, then Clyde would go, and the second says that if Bill were to go then Clyde would *not* go. If we knew these two counterfactuals to be true, then surely A and B would be dependent: If A is true, then B could not be true, and vice versa. But now generalize this observation to the probabilistic case. Consider these probabilities of counterfactuals:

$$\text{Prob}(A \Box \rightarrow C) = 0.9$$

and

$$\text{Prob}(B \Box \rightarrow C) = 0.1$$

In this case also, it is plausible that A and B turn out to be dependent. For learning A means that C is highly probable, which in turn implies that B is highly *improbable*. Thus if one of these counterfactuals has a high probability and the other a low probability, it is natural that their antecedents become dependent.

The only issue, then, concerns that case in which the two counterfactuals either both have a high probability or both have a low probability. Even here an

argument can be made that dependence is the right result. Consider the case already discussed in which both counterfactuals have probability one—we know both $A \Box \rightarrow C$ and $B \Box \rightarrow C$ with certainty. Why should A and B end up dependent in this case?

One way to understand the matter is to consider a strong and a weak sense of counterfactual implication. Both counterfactuals say that C is counterfactually implied by some antecedent causal factor. There is a strong sense of counterfactual implication in which to say that C is counterfactually implied by X is to say that if X were to happen, then *no matter what other antecedent events occurred*, C would happen. One thing that may be producing the paradox in this case is that we take the counterfactuals to express the strong sense of counterfactual implication. If the strong sense is meant, then we are really asserting the counterfactuals

$$AB \Box \rightarrow C \quad A\bar{B} \Box \rightarrow C \quad \text{and} \quad \bar{A}B \Box \rightarrow C$$

But then the antecedents of the counterfactuals are mutually exclusive, so Theorem 2 guarantees that independence of A and B will be preserved.

Suppose the strong sense of counterfactual implication is not what is meant. Then what we assert is compatible with these two counterfactuals:

$$A\bar{B} \Box \rightarrow \bar{C} \tag{3}$$

and

$$\bar{A}B \Box \rightarrow \bar{C} \tag{4}$$

But if (1)–(4) are jointly true, then the following two counterfactuals will also be true:

$$A \Box \rightarrow B \tag{5}$$

and

$$B \Box \rightarrow A \tag{6}$$

This follows from the valid rule of inference “ $\phi \Box \rightarrow \chi$, $\phi\chi \Box \rightarrow \psi$; therefore $\phi \Box \rightarrow \psi$.” For either A is inconsistent or A is consistent. If A is inconsistent, then it counterfactually implies everything, including B . If A is consistent, then it cannot counterfactually imply \bar{B} , since if it did, $A \Box \rightarrow \bar{B}$ and $A\bar{B} \Box \rightarrow \bar{C}$ would entail $A \Box \rightarrow \bar{C}$, contradicting (1). But either $A \Box \rightarrow B$ or $A \Box \rightarrow \bar{B}$, and since $A \Box \rightarrow \bar{B}$ is false, $A \Box \rightarrow B$ must be true. A similar argument shows that (6) follows from (2) and (4).

But that means that given (1) and (2), (3) and (4) imply that A and B are positively correlated, that is, they are no longer independent. Now if you consider other possibilities for how the various combinations of truth values of A

and B counterfactually imply a truth value for C , you will find that some imply no correlation, positive or negative, between A and B ; some imply a positive correlation; while others imply a negative correlation (e.g., if we have $AB \Box \rightarrow \bar{C}$, then the truth of $A \Box \rightarrow C$ and $B \Box \rightarrow C$ requires the truth of both $A \Box \rightarrow \bar{B}$ and $B \Box \rightarrow \bar{A}$). However, it turns out that the cases in which there is a positive correlation “outweigh” the cases in which there is a negative correlation between A and B . That is, if we condition on $A \Box \rightarrow C$ and $B \Box \rightarrow C$, the former cases will have greater probability than the latter. Therefore, on balance, A and B will have a slight positive correlation after $A \Box \rightarrow C$ and $B \Box \rightarrow C$ are conditioned upon.

Thus a good argument can be made that Al’s and Bill’s actions should *not* remain independent after conditionalization on $A \Box \rightarrow C$ and $B \Box \rightarrow C$. A similar argument can be made for the case in which it is learned that these counterfactuals have a high probability.

The above discussion shows that there is danger in one suggested strategy for dealing with causal information. It has been suggested that one way to get the right result in Pearl’s puzzle, without appealing to counterfactuals, is to *assume* independence when there is no information about dependence. The problem with this strategy is that it is not always obvious when there is a lack of information about dependence. Our intuitions can mislead us here. We have seen that although it is very easy to assume that the two counterfactuals $A \Box \rightarrow C$ and $B \Box \rightarrow C$ provide no information about any dependency between A and B , this assumption is wrong. What is needed is a systematic theory of how causal information should be assimilated, a theory that allows us to compute what dependencies are implied rather than relying on our sometimes faulty intuitions. The theory of probabilistic counterfactuals presented in this paper is such a theory.

CAUSAL VS. DIAGNOSTIC REASONING

An important distinction in uncertain reasoning is the one between causal and diagnostic reasoning. The distinction is roughly the same as between reasoning from causes to effects as opposed to reasoning from effects to causes. The difference between these two directions of reasoning is sharply brought out in Pearl’s Bayesian networks [15], where the parameter π is a measure of causal support and the parameter λ is a measure of diagnostic or evidential support. Psychological studies (Tversky and Kahneman [16]) of these two methods of reasoning purport to show that people are biased in favor of causal reasoning that is, that people find it easier to reason from a cause to an effect than from an effect to a cause, even when the degree of informativeness of the cause about the effect is the same as the degree of informativeness of the effect about the cause.

A more general result of this same sort is that people are more comfortable reasoning from a strong indicator of a cause to a weak indicator of a cause even though probabilistically such an asymmetry is unjustified. Some researchers have concluded from these results that people reason irrationally in the presence of causal information.

This section applies the theory of counterfactuals and causality developed in the previous sections to the question of whether or not people are reasoning irrationally when they treat causal and diagnostic information asymmetrically. It will be argued that depending on what question is being asked, such an asymmetrical treatment may not be irrational and, furthermore, that instances where such asymmetrical treatment is irrational can be explained by supposing that people are confusing a noncausal question with a causal or counterfactual one. Tversky and Kahneman [16] report on an experiment in which subjects were asked to state which, if either, of the following events is the more probable:

- (a) That a girl has blue eyes if her mother has blue eyes.
- (b) That the mother has blue eyes if her daughter has blue eyes.

Sixty-nine subjects said that (a) is more probable, only 21 felt that (b) is more probable, and 75 said that the two events are equally probable.

Tversky and Kahneman point out that since the a priori probability of the daughter's having blue eyes is equal to the a priori probability of the mother's having blue eyes, the conditional probability of the daughter's having blue eyes given that the mother has blue eyes must equal the conditional probability of the mother's having blue eyes given that the daughter has blue eyes. This follows from the definition of the conditional probability $P(X|Y)$ as $P(XY)/P(Y)$, which implies that if $P(X) = P(Y)$, $P(X|Y) = P(XY)/P(Y) = P(XY)/P(X) = P(Y|X)$.

One problem in evaluating this study is that (a) and (b) were phrased in ordinary English and, at least in Ref. 16, no information was presented as to whether or not the subjects understood the probabilities of (a) and (b) to be conditional probabilities as standardly defined in probability theory. Perhaps Kahneman and Tversky thought there was no other way to understand (a) and (b), but if so they were mistaken. One alternative possibility is that at least some of the subjects interpreted (a) and (b) as counterfactuals, so that their probabilities were probabilities of counterfactuals, not conditional probabilities.

Although (a) and (b) are phrased in the indicative mode and counterfactuals are normally expressed in the subjunctive, it is plausible to think that conditionals in the indicative mode are sometimes given a counterfactual interpretation. For example, the conditional "This dissolves if it is put in water" seems to convey the same information as "If this were put into water, then it would dissolve." Therefore one possible interpretation of (a) and (b) is as the following counterfactuals:

- (a') If the mother were to have blue eyes, the daughter would too.
 (b') If the daughter were to have blue eyes, then so would the mother.

The question is whether or not the probabilities of (a') and (b') must be equal if the probabilities of their antecedents are equal. More generally, is

$$P(A) = P(C) \Rightarrow P(A \Box \rightarrow C) = P(C \Box \rightarrow A) \quad (7)$$

a valid implication?

Now many students of counterfactuals have noted that a conditional probability need not equal the probability of the corresponding counterfactual conditional (e.g., Lewis [12], pp. 71–72]). So the fact about conditional probabilities cited by Tversky and Kahneman provides no support for (7).

In fact, counterexamples to (7) are easy to generate. For example, within the framework for probabilistic counterfactuals developed in this paper, suppose we update on the constraints $P(A \Box \rightarrow C) = 1$ and $P(\bar{A} \Box \rightarrow \bar{C}) = 1$. Then no world in which A is true could be a world in which C is false, and no world in which A is false could be a world in which C is true. Therefore, $P(A) = P(C)$. Does it follow that $P(C \Box \rightarrow A) = 1$? Not at all. For updating on the previous two counterfactuals essentially involves eliminating all sequences in which these two counterfactuals are false and renormalizing over the remaining sequences. But sequences in which $C \Box \rightarrow A$ is *false* are among the remaining sequences. For example, the sequence $\langle \bar{A}\bar{C}, \bar{A}C, AC, A\bar{C} \rangle$ is not eliminated (since both $A \Box \rightarrow C$ and $\bar{A} \Box \rightarrow \bar{C}$ are true in it), yet $C \Box \rightarrow A$ is false in this sequence. Hence $P(C \Box \rightarrow A) < 1$.

For a more intuitive example, consider the following urn model: There are two urns, U_1 and U_2 . U_1 contains 90 black balls and 10 white balls, while U_2 contains 10 black balls and 90 white balls. A ball is to be randomly drawn from one of the urns, which urn it is drawn from being determined by the flip of a fair coin. Let B stand for "A black ball is drawn." Then $P(B)$, the probability of a black ball being drawn, is equal to $P(U_1)P(B|U_1) + P(U_2)P(B|U_2) = 0.5(0.9) + 0.5(0.1) = 0.5$. Since $P(U_1)$ is also 0.5, we have $P(U_1) = P(B)$. However, the probability that a black ball would be drawn were U_1 chosen is not equal to the probability that U_1 would be chosen were a black ball to be drawn. The first probability is clearly 0.9. But the second probability is equal to the probability of U_1 , since the assumptions of the example entail that $B \Box \rightarrow U_1$ is equivalent to U_1 , as the following argument shows: Suppose U_1 is in fact chosen. Then it would certainly be true to say that if a black ball were drawn, U_1 would (still) have been chosen. If, however, U_1 is not chosen, then it would be false to say that if a black ball were drawn then U_1 would have been chosen. All this can be known from the description of the case. Hence $P((B \Box \rightarrow U_1) \equiv U_1) = 1$, so $P(B \Box \rightarrow U_1) = P(U_1) = 0.5$.

The above urn model not only invalidates (7) but also illustrates an asymmetry between cause and effect in the evaluation of counterfactuals. Choosing a

specific urn has a causal influence on what the color of the drawn ball will be,⁴ whereas the color of the drawn ball in no way affects which urn is chosen. As the urn model shows, when the probability of the cause is less than the probability of the effect given the cause, as will typically be the case, the probability of the counterfactual saying that if the cause were to occur then the effect would occur, is higher than the probability of the counterfactual saying that if the effect were to occur then the cause would have occurred, even when cause and effect have equal prior probabilities. This asymmetry between cause and effect perhaps also accounts for the responses of the 69 subjects in the experiment who thought statement (a) more probable than statement (b). For the subjects would presumably believe that the mother's eye color has a causal influence on the eye color of the daughter, but not vice versa.

The above considerations suggest that what have been characterized as "errors of reasoning" in judging the probabilities of statements such as (a) and (b) may not be errors at all. The putatively erroneous judgments may very well be correct if they are judgments about the probabilities of counterfactuals rather than judgments of conditional probability. And even if the judgments are clearly erroneous, counterfactuals can still shed light on what is going wrong. Some of Tversky and Kahneman's examples are phrased directly in terms of conditional probabilities, using standard mathematical notation for conditional probability, or they are phrased in terms of *predicting* one event on the basis of information about another event, where it seems clear that the basis for prediction should be the conditional probability of the first event given the second. Tversky and Kahneman report finding the same biases in judgment (in favor of reasoning from cause to effect or from a strong indicator of cause to a weak indicator of cause) in these sorts of examples as in the examples discussed above. But this bias in judgment can be explained if we suppose that the subjects mistakenly translate conditional probabilities into probabilities of counterfactuals. Such a mistaken translation is plausible because evaluating the counterfactual $A \Box \rightarrow C$ and evaluating the conditional probability $P(C|A)$ both can be done by hypothetically adding proposition A to one's set of beliefs and then making some kind of minimal revision of beliefs to restore consistency. The difference is that different types of revisions are performed (in the one case, conditionalization; in the other, shifting the probability of a world to its most similar A world). The close similarity between the two processes of evaluation makes it plausible that one process could be wrongly substituted for the other.

Of course, much more needs to be done to develop a theory based on

⁴ More precisely, the event of choosing a specific urn has a causal influence on the events described by "Whichever ball is drawn is white (black)." The event of choosing a specific urn does not have a causal influence on an event such as "Ball number 38 is white," where the balls are first numbered and then randomly assigned to the urns.

probabilistic counterfactuals that explains people's specific judgments about the probability of such statements as (a) and (b). This section was meant merely to open up discussion on this topic by pointing out that whether or not such judgments are correct often depends on whether the probability in question is interpreted as the probability of a conditional or as a conditional probability.

DISCUSSION

It has been shown that MAXENT can be applied to causal information and that the resulting posterior leaves the prior over the causes unchanged, affecting only the conditional probability of the effect given a particular combination of causes. This fact answers Pearl's objection to MAXENT that it eliminates independence of events when information about an effect of those events is given. The problem was seen to lie, not with MAXENT, but with the attempt to express causal information in terms of conditional probabilities, which, as was shown, are inadequate to express certain causal notions that are better expressed in terms of counterfactual statements.

One way to accomplish the same result as MAXENT applied to counterfactuals, without actually using counterfactuals, would be to apply MAXENT to the given conditional probabilities together with constraints that fix the probabilities of the combinations of causes to their prior probabilities. It might be suggested that for this reason the whole excursion into counterfactuals was unnecessary. This is misguided, not just for the reason given earlier that our intuitions about dependence or independence are sometimes wrong, but also for the following more fundamental reason: The alternative just described neglects the distinction between the case in which the prior is known and the case in which it is simply a best estimate, based on incomplete information about the prior. For example, in the version of the puzzle involving coin tossings, the prior over the causes (the tosses of the two coins *A* and *B*) may very well be known to us. If we know that the two coins are fair and physically independent of one another, this forces the prior over the coin tosses to be the uniform prior, and it is therefore reasonable, upon receiving information about an effect of the coin tosses, to take as a constraint that the marginal probability over the coin tosses be uniform. However, it is different when the prior over the causes is based on a lack of information, as in the case of the invitees to the party, Al and Bill. In this case, we make Al's and Bill's actions probabilistically independent because we lack information about any connection between them, not because we know that their actions are not connected in any way. In such a case, it is ad hoc to take independence or uniformity as a *constraint*, and we should want our updating rule to preserve independence without such a constraint. It is not clear how a strict Bayesian could preserve independence in such cases without making independence an assumption.

Another advantage of MAXENT over the Bayesian approach to this problem is that MAXENT can work with *incomplete* information about causal connections. It is not necessary that the constraints cover all the possible combinations of causes. If S stands for some combination of causes such that the probability of $S \Box \rightarrow C$ is unknown, then applying MAXENT to a set of constraints that does not include a constraint for that counterfactual is equivalent to applying MAXENT to the same set of constraints plus the constraint that the probability of that counterfactual is $1/2$. Hence the results of Theorem 2 apply also to the case in which the S_i are not exhaustive, with $Q(C|S) = 1/2$ for each missing combination of causes S .

Another misgiving that some might have about my analysis is the issue of whether or not counterfactuals provide a *noncircular* analysis of causality. As noted in the section on counterfactuals and causality, the exact nature of the connection between counterfactuals and causal notions is a matter of controversy. Ginsberg [10], for example, presents counterfactuals in which the relation between the antecedent and the consequent seems to be the same as between effect and cause or in which there seems to be *no* causal relation between antecedent and consequent.

One of Ginsberg's examples of a counterfactual going in the "wrong" direction is this one:

"If the result of the test had been positive, then the organism would have been rodlike."

Clearly, the test's coming out positive would not cause the organism to be rodlike nor would the test's not coming out positive cause it not to be rodlike.

One response to Ginsberg is to deny that the above counterfactual is true. If the organism is not rodlike, instead of asserting the above counterfactual, one might instead assert:

"If the result of the test had been positive, then the test would have been faulty."

What is important for our purposes, however, is not that Ginsberg's reading be ruled out, but that the causal reading, the one in which the counterfactual is false, be an allowable one. Counterfactuals are irremediably vague, but a logic of counterfactuals is still useful if it remains correct under different ways of resolving the vagueness of counterfactuals.

This last point connects with the issue of whether or not counterfactuals can be used to give an analysis of causality. The worry is that for counterfactuals to give an analysis of causality, the similarity relation between worlds has to be understood in a certain way, but explaining how the similarity relation is to be understood would require recourse to causal notions.

This worry need not bother us. For it is not important for the purposes of this paper that counterfactuals be able to provide a *noncircular* analysis of causality.

All that is important is that there be the right sort of connection between counterfactuals, for some choice of the similarity relation, and causality. Then the formal properties of counterfactuals can be used to prove the sort of results that have been proved in this paper.

Finally, we note some limitations of the modified MAXENT updating rule. That applying this rule in the small space W gives the same result as doing straight MAXENT in the large space Ω depends, if one examines the proof of Theorem 2, upon the fact that each member of W is neutral with respect to any counterfactual not entailed by that member. That is, we start off with a probability measure $P(\cdot)$ such that if $i \neq j$, $P(S_i \Box \rightarrow C | S_j C) = P(S_i \Box \rightarrow C | S_j \bar{C}) = 1/2$. If this neutrality did not hold, then the posterior marginal over the causes would differ from the prior marginal over the causes, and the difference could not be determined by looking only at the space W . However, it is not unreasonable to suppose that this type of neutrality holds if the counterfactual constraints are all that is known about the causal connections between the A_i and C .

APPENDIX

The well-known result that a Stalnaker selection function determines a linear ordering of the worlds with respect to each world (assuming the selection function to be defined on propositions, i.e., sets of worlds) can be proved as follows. Let $f(\cdot)$ be a selection function for the set of worlds W , and let u, v , and w be members of W . Define $u \leq_w v$ by:

$f(\{u, v\}, w) = u$, if $\{u, v\}$ is a proposition true just in the worlds u and v .

Then we have the following theorem:

THEOREM 1. $u \leq_w v$ is a linear order with w as least element.

Proof The selection function $f(\cdot)$ is stipulated to have the following properties:

- (i) A is true in $f(A, w)$, provided A is logically consistent.
- (ii) If A is true in w , then $f(A, w) = w$.
- (iii) If A is true in $f(B, w)$ and B is true in $f(A, w)$, then $f(A, w) = f(B, w)$.

That w is a least element follows from the fact that by clause (ii), $f(\{w, v\}, w) = w$.

That $u \leq_w v$ is reflexive follows from the fact that $f(\{u, u\}, w) = u$, by clause (i).

$u \leq_w v$ is antisymmetric: If $f(\{u, v\}, w) = u$ and $f(\{u, u\}, w) = v$, then clearly $u = v$.

$u \leq_w v$ is transitive: First note that if A is true in $f(C, w)$ and B is true in $f(D,$

w), then $A \cup B$ is true in $f(C \cup D, w)$, since clause (iii) entails that $f(C \cup D, w)$ is either $f(C, w)$ or $f(D, w)$. Therefore

$$\{u, v\} \text{ is true in } f(\{u, v\} \cup \{v, x\}, w) = f(\{u, v, x\}, w).$$

Since $\{u, v, x\}$ is true in $f(\{u, v\}, w)$, clause (iii) implies that

$$f(\{u, v, x\}, w) = f(\{u, v\}, w) = u.$$

Therefore $\{u, x\}$ is true in $f(\{u, v, x\}, 2)$. Since $\{u, v, x\}$ is true in $f(\{u, x\}, w)$, clause (iii) yields

$$f(\{u, x\}, w) = f(\{u, v, x\}, w).$$

Hence $f(\{u, x\}, w) = u$. \square

To prove Theorem 2, we need some definitions and a lemma: S_1, \dots, S_r are the atoms over $\{A_1, \dots, A_k\}$ (the set of causal factors) and $P(\cdot)$ is a probability function over $\{S_1, \dots, S_r\}$. W is the set of all atoms over $\{S_1, \dots, S_r, C\}$ (where C is a possible "consequence" for each combination of causal factors S_i), and Ω is the set of all sequences of members of W . $P(\cdot)$ is extended to Ω by defining, for each s in Ω , $P(s)$ to be $P(S_i/N)$, where S_i is true in s and N is the number of members of Ω in which S_i is true. MAXENT(p, c), where p is a probability function and c is an assignment or set of assignments of probability to a subset or subsets, of p 's space, denotes the result of applying maximum entropy updating to p with c as constraint. Now we may state the following lemma:

LEMMA 1 *Define the sequence Q_0, Q_1, \dots, Q_r of probability measures over Ω by*

$$Q_0 = P$$

$$Q_i = \text{MAXENT}(Q_{i-1}, \text{Prob}(S_i \square \rightarrow C) = p_i)$$

Then for each i , $0 \leq i \leq r$:

$$Q_i(S_j) = P(S_j), \quad 1 \leq j \leq r \quad (1)$$

$$\text{If } i > 0, \quad Q_i(S_i \square \rightarrow C) = Q_i(C | S_i) = p_i \quad (2)$$

$$\text{If } i > 0, \quad Q_i(S_i \square \rightarrow C) = Q_i(C | S_i) = Q_{i-1}(C | S_j), \quad j \neq i \quad (3)$$

Proof The proof is by induction on i .

Basis case: Equations (2) and (3) are vacuously true for $i = 0$. By definition, $Q_0 = P$, so Eq. (1) holds for $i = 0$.

Inductive step: Assume (1)–(3) for all $m < i$, where $i > 0$. First we show that

$$Q_{i-1}(S_i \square \rightarrow C) = Q_{i-1}(S_i \square \rightarrow C | S_j) = 1/2.$$

Partition each of the S_j by the possible combinations of truth values of $S_1 \square \rightarrow C$,

$\dots, S_{i-1} \Box \rightarrow C$. Let Δ be an arbitrary member of the partition of an arbitrary S_j . By the Lemmer algorithm, all members of Δ have the same probability in Q_{i-1} (e.g., if Δ corresponds to all of $S_1 \Box \rightarrow C, \dots, S_{i-1} \Box \rightarrow C$ being true, then for each member s of Δ ,

$$Q_{i-1}(s) = \frac{P(S_j)}{N} \prod_{k=1}^{k=i-1} \frac{p_k}{Q_{k-1}(S_k \Box \rightarrow C)}.$$

$S_i \Box \rightarrow C$ is true in exactly half the members of Δ . (since for any sequence in Δ in which S_i & C occurs before S_i & \bar{C} there is a sequence in the same partition in which the order of these two worlds is switched). Hence $Q_{i-1}(S_i \Box \rightarrow C | S_j) = 1/2$. Therefore,

$$\begin{aligned} Q_{i-1}(S_i \Box \rightarrow C) &= \sum_{h=1}^r Q_{i-1}(S_i \Box \rightarrow C | S_h) Q_{i-1}(S_h) \\ &= 1/2 \sum_{h=1}^r Q_{i-1}(S_h) = 1/2. \end{aligned}$$

By the Lemmer updating algorithm

$$\begin{aligned} Q_i(S_j) &= \frac{Q_{i-1}(S_j(S_i \Box \rightarrow C))p_i}{Q_{i-1}(S_i \Box \rightarrow C)} + \frac{Q_{i-1}(S_j(S_i \Box \rightarrow \bar{C}))(1-p_i)}{Q_{i-1}(S_i \Box \rightarrow \bar{C})} \\ &= p_i Q_{i-1}(S_j | S_i \Box \rightarrow C) + (1-p_i) Q_{i-1}(S_j | S_i \Box \rightarrow \bar{C}). \end{aligned}$$

From the fact proved above that $Q_{i-1}(S_i \Box \rightarrow C) = Q_{i-1}(S_i \Box \rightarrow C | S_j)$, it follows that

$$Q_{i-1}(S_j | S_i \Box \rightarrow C) = Q_{i-1}(S_j | S_i \Box \rightarrow \bar{C}) = Q_{i-1}(S_j).$$

Hence $Q_i(S_j) = p_i Q_{i-1}(S_j) + (1-p_i) Q_{i-1}(S_j)$ and (1) follows by induction.

In Eq. (2), the identity $Q_i(S_i \Box \rightarrow C) = p_i$ holds trivially [since Q_i results from updating on the constraint $\text{Prob}(S_i \Box \rightarrow C) = p_i$]. $Q_i(C | S_i) = Q_i(C \& S_i) / Q_i(S_i) = [\text{by Eq. (1)}] Q_i(C \& S_i) / Q_{i-1}(S_i)$. By the Lemmer updating algorithm, $Q_i(C \& S_i) = Q_{i-1}(C \& S_i) p_i / Q_{i-1}(S_i \Box \rightarrow C) = [\text{by the inductive assumption for Eq. (3)}] Q_{i-1}(C \& S_i) p_i / Q_{i-1}(C | S_i) = Q_{i-1}(S_i) p_i$. Hence

$$Q_i(C | S_i) = Q_{i-1}(S_i) p_i / Q_{i-1}(S_i) = p_i$$

To prove Eq. (3), consider again the set of all possible combinations of truth values of $S_1 \Box \rightarrow C, \dots, S_{i-1} \Box \rightarrow C$. As in the proof of (1), we use these combinations to partition a subset of Ω —in this case, the subset corresponding to the formula $S_j \Box \rightarrow C$. Some subsets in the partition may be empty; for example, if $j \leq i-1$, then every combination of truth values of the given counterfactuals in which $S_j \Box \rightarrow C$ is false will result in the empty set when applied to that

counterfactual. However, for each non-empty subset in the resulting partition, we observe, as we did in the proof of (1), that the members are equiprobable under Q_{i-1} and that $S_i \square \rightarrow C$ is true in exactly half of them (remember that j is assumed to be distinct from i). Thus

$$Q_{i-1}(S_i \square \rightarrow C | S_j \square \rightarrow C) = 1/2.$$

By parallel reasoning, we may establish that

$$Q_{i-1}(S_i \square \rightarrow C | S_j \square \rightarrow \bar{C}) = 1/2.$$

Hence $Q_{i-1}(S_i \square \rightarrow C) = 1/2$. Thus $S_i \square \rightarrow C$ is independent of $S_j \square \rightarrow C$ under Q_{i-1} , from which it follows that $Q_i(S_j \square \rightarrow C) = Q_{i-1}(S_j \square \rightarrow C)$. A similar argument establishes that $Q_i(C | S_j) = Q_{i-1}(C | S_j)$. Now if $j = i - 1$, (2) implies that $Q_{i-1}(S_j \square \rightarrow C) = Q_{i-1}(C | S_j)$, so that in this case,

$$Q_i(S_j \square \rightarrow C) = Q_{i-1}(S_j \square \rightarrow C) = Q_{i-1}(C | S_j) = Q_i(C | S_j).$$

Assume, then, that $j \neq i - 1$. Then the inductive assumption for (3) implies that $Q_{i-1}(S_j \square \rightarrow C) = Q_{i-1}(C | S_j)$, from which again we get that $Q_i(S_j \square \rightarrow C) = Q_i(C | S_j)$. \square

We can now prove Theorem 2.

THEOREM 2. *Let $\Gamma = \text{Prob}(S_i \square \rightarrow C) = p_i; i = 1, \dots, r\}$. Then if $Q = \text{MAXENT}(P, \Gamma)$, Q satisfies (i) $Q(S_i) = P(S_i)$ and (ii) $Q(C | S_i) = p_i$.*

Proof $Q_r = \text{MAXENT}(P, \Gamma)$, since by Lemma 1, Eqs. (2) and (3), the i th constraint will be satisfied in Q_i and will remain satisfied thereafter. By Eq. (1) of Lemma 1, $Q_r(S_i) = Q(S_i) = P(S_i)$. By Eqs. (2) and (3) of Lemma 1, $Q_r(C | S_i) = Q(C | S_i) = p_i$. \square

References

1. Cheeseman, P., A method for computing generalised Bayesian probability values for expert systems, *Proceedings of the 8th International Joint Conference on AI*, Karlsruhe, West Germany, 198–202, 1983.
2. Shore, J. E., and Johnson, R. W., Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy, *IEEE Trans. Inf. Theory*, IT-26(1), 26–37, 1980.
3. Van Campenhout, J. M., and Cover, T. M., Maximum entropy and conditional probability, *IEEE Trans. Inf. Theory*, IT-27(4), 483–489, July 1981.
4. Jaynes, E. T., *Papers on Probability, Statistics and Statistical Physics* (R. D. Rosenkrantz, Ed.), D. Reidel, Dordrecht, Holland, 1983.
5. Lemmer, J. F. and Barth, S. W., Efficient minimum information updating for

- Bayesian inferencing in expert systems, *Proceedings of the National Conference on AI*, Pittsburg, Penn., 424–427, 1982.
6. Hunter, D., Uncertain reasoning using maximum entropy inference, in *Uncertainty in Artificial Intelligence* (L. N. Kanal and J. F. Lemmer, Eds.), North-Holland, Amsterdam, 203–209, 1986.
 7. Lemmer, J. F., Algorithms for incompletely specified distributions in a generalized graph model for medical diagnosis, PhD Thesis, University of Maryland, College Park, MD., 1976.
 8. Jeffrey, R., *The Logic of Decision*, McGraw-Hill, New York, 1965
 9. Csiszar, I., I-divergence geometry of probability distributions and minimization problems, *Ann. Probability* 3(1), 146–158, 1975.
 10. Ginsberg, M., Counterfactuals, *AI* 30, 35–79, 1986.
 11. Stalnaker, R., A theory of conditionals, in *Causation and Conditionals* (E. Sosa, Ed.), Oxford University Press, 165–179, 1975.
 12. Lewis, D., *Counterfactuals*, Harvard University Press, 1973.
 13. Sosa, E., Ed., *Causation and Conditionals*, Oxford University Press, 1975.
 14. Lewis, D., Causation, in *Causation and Conditionals* (E. Sosa, Ed.), Oxford Univ. Press, 180–191, 1975.
 15. Pearl, J., Fusion, propagation, and structuring in Belief networks, *AI* 29, 241–288, 1986.
 16. Tversky, A., and Kahneman, D., Causal schemas in judgments under uncertainty, in *Judgment under Uncertainty* (D. Kahneman, P. Slovic, and A. Tversky, Eds.), Cambridge University Press, 117–128, 1985.