Causal Influence and Controlled Experiment

Allan Gibbard
University of Michigan
Ann Arbor, Michigan 48109
U. S. A.

What is an empiricist to say about causality? Experience seems to teach us much about causal influences in the world. Statisticians warn us that correlation is not causation, but controlled experiments set their worries to rest. With a well controlled experiment, we learn how much one factor tends to influence another; and as numbers become large, we learn it beyond any serious doubt. Controlled experiments look empirical if anything does, and we need to ask: How do they reveal the causes that lie hidden behind the veil of experience?

Causation was Hume's central worry, or one of them, and notoriously he gave three distinct characterizations of a cause—characterizations that do not, on their face, look equivalent. Hume's three strategies are still the obvious contenders.[1]

First, we can try to reduce causality in the world to non-causal features of the world. Hume proposed that causality is constant conjunction.

A second strategy too is reductionistic, but in a more roundabout and surprising way—a way often called *projectivism* or *expressivism*.[2] The scheme is this: To explain causes in the world outside our thoughts, look not to other goings-on apart from our thoughts. Rather, explain causes by saying what it is to have causal beliefs. Hume analysed causality as succession, contiguity in time and place, and necessary connection; the last was the puzzling concept. What is it, Hume asked, to *believe* in a necessary connection between A and B? The belief, he answered, consists in the determination of the mind to pass from A to B. The belief, then, is not straightforwardly belief in a further feature of A and B and the world outside our thoughts connecting them. The belief consists in another feature of our thoughts.

Current writers as well have tried projectivistic strategies, often with Bayesian twists. Believing that A causes B, they propose, consists in features of one's subjective probabilities, in one's degrees of belief in propositions that are not themselves causal.

---

[1] In using these dicta of Hume's to classify strategies for explaining causation, I follow Lewis (1973, 159–61).

[2] Blackburn, Gibbard.

Instead of a determination of the mind to pass from A to B, the Bayesian can speak of subjective conditional probabilities. One's subjective conditional probability for B given A is the degree to which one stands ready to believe B on learning A and nothing else. This conditional probability is not itself a promising stand-in for causal belief; correlation, after all, is not causality. On hearing a loud screech and a bang, I might expect dented cars, but not because I think noises cause dents. Rather, I think dents likely to stem from some of the causes of the noise. Still, there are far more sophisticated Bayesian strategies for reading causal beliefs from subjective probabilities—from the subjective probabilities of non-causal propositions. I shall be using one of them as a starting point for my own discussion.[3]

I myself want to explore a third strategy—a non-Humean strategy that a few of Hume's words suggest. One object causes another, Hume says, "where, if the first object had not been, the second never had existed." That suggests causality as a relation in the world apart from our thoughts, a relation of counterfactual dependence.

How, though, could an empiricist countenance such a relation, a relation we cannot experience? The answer I'll develop is Bayesian. We learn from experience by conditionalization, modifying our prior probabilities in light of experience. Not every reasonable person will come to new experience with the same priors, but some kinds of evidence, I'll argue, will bring convergence among people with a wide range of disparate prior probabilities. Causal beliefs respond to evidence not because they are non-causal beliefs in disguise, but because *sui generis* and specially structured though they be, they respond to evidence in the same way as any other belief.[4]

The approach in this paper is closely related to that of Judea Pearl and coworkers, in his book *Causality* (2000) and subsequently.[5] I comment briefly on the relationship at the end of the paper, but only tentatively. The relationships between the two treatments need much more extensive treatment.

---

[3] Skyrms (1984) is a possible example; I discuss this work below.

[4] This is in many ways like approachs developed by David Lewis. My treatment of objective chance in this paper derives from Lewis (1980), and my talk of causal influence has close affinities to that in Lewis (1973; 1981; 1986, 179–84). But although in this paper I adopt various aspects of Lewis's broad approachs to causality and chance, I address somewhat different questions, and some differences in the philosophical morals I draw will emerge.

[5] See website http://bayes.cs.ucla.edu/jp_home.html for a list of the subsequent work by Pearl, much of it jointly with others.

1. Naive Foundations

The more or less commonsense starting point for what I'll do is this: Suppose I have a biased coin in my pocket. Then we can speak of its chance of landing heads were I to flip it. This chance is not simply my conditional credence that it lands heads given that I flip it. I may be convinced that the coin is unbiased; alternatively, though, I may think the coin biased, but give no more credence to its being biased one way than the other. In either case, my conditional credence

$$\mathcal{B}(\text{the coin lands heads / I flip it})$$

is $\frac{1}{2}$, but nevertheless the objective conditional chance that it would land heads were I to flip it is different from $\frac{1}{2}$. That, after all, is just what it is for the coin to be biased.

In saying this, of course, I have merely been recounting what we might say when we first try to think about coin-flipping in a systematic way. I have not given this notion of objective conditional chance philosophical scrutiny. Now, though, I want to explore what happens in the case of coin flips and in other such cases, and think how a Bayesian empiricist should treat such phenomena.

Take a coin that I firmly intend not to flip. Perhaps you have convinced me that if I flip it, I'll win $1000 if it lands heads and lose $2000 if it lands tails. Giving equal credence to the coin's being biased in either direction, and being unwilling to gamble at bad odds, I firmly intend not to flip it. I can speak, though, of the objective chance that it *would* land heads, *were* I to flip it.

What might this mean? It is not the chance of truth of the conditional "If I were to flip the coin, it would land heads." This flat conditional I hold untrue, since after all, I'm convinced that coin-flipping is genuinely chancy.

What, then, do I mean by "the chance of the coin's landing heads were I to flip it"? Naively, we might think this. If the coin is biased .6 toward heads, that means that with an objective conditional chance of .6 it would land heads were I to flip it. We might set up the following notation.

$$\mathcal{P}^A(C): \text{ the chance of } C \text{ were I to do } A$$

We might then go on to speak of the value that $\mathcal{P}^A(C)$ has if condition h obtains; call this $\mathcal{P}^A_h(C)$. h, for instance, might be the proposition that the coin is biased at .6 toward heads; then $\mathcal{P}^A_h(C)$ is the chance given this hypothesis of the coin's landing heads were it flipped. $\mathcal{P}^A_h(C)$ in this case would be .6.

Since $\mathcal{P}^A$ gives the chances of things were I to do A, I'll call such items *chances-were*. Chances-were, then, are chances of things *were* such-and-such a condition to

obtain. They aren't chances of conditional propositions; they don't take the form $\mathcal{P}$(If A were to obtain then C would.)[6] For again, if the matter is genuinely chancy, then this conditional proposition has no chance of obtaining. Nor is this a conditional probability $\mathcal{P}(C/A)$ in the standard sense.[7] This won't be defined if $\mathcal{P}(A) = 0$, whereas we can speak of what would tend to happen *were* the coin flipped even if we know there's no chance of its being flipped. Then too, even if A is chancy in that $1 > \mathcal{P}(A) > 0$, the chance-were $\mathcal{P}^A(C)$ can diverge from $\mathcal{P}(C/A)$ if there is a common influence on A and C. Gambler Sly Pete, suppose, won't flip a coin unless he knows it is heavily weighted to favor him. He'll draw a coin from a hat and then choose whether to flip it. Which coin he'll get to flip is a matter of chance, and most of the coins he might draw are weighted against him. One coin, though, is weighted heavily for him, and he'll recognize it if he draws it. Let A be that he'll flip the coin he draws, and C that the coin will land heads. $\mathcal{P}(C/A)$ is high, since he'll flip the coin he draws only if it is one that will tend to land heads if flipped. But $\mathcal{P}^A(C)$ is low, since chances are he'll draw a coin that would tend not to land heads if flipped.

'$\mathcal{P}^A(C) = .6$' might mean $A \mathbin{\square\!\!\rightarrow} [\mathcal{P}(C) = .6]$, that were the coin flipped, then it would be the case that its chance of landing heads was .6. That's a proposal that David Lewis makes.[8] In this paper, though, I won't pronounce on whether this is a good explication—even if some of the turns of phrase I use do suggest this analysis. I'll assume that we have some informal ability to use the notion of the chance with which a coin would land heads were it flipped, and to extend this notion to other cases. Later I'll impose on it some extremely weak conditions.[9]

Suppose, now, I am convinced that the coin is biased at .6 either toward heads or toward tails, and I give equal credence to the two possibilities. Then the probability that is relevant to my decision not to gamble is $\frac{1}{2}$: It is the average of .6 and .4. For the sake of brevity, we might call this quantity my *expectation* of heads *were* I to flip the coin. Where A is "I flip the coin" and C is "It lands heads," I shall write this $\mathcal{E}^A(C)$, and call it an *expectation-were.* The formula for my expectation in this simple case is as follows: Let $\mathcal{B}$ be the function giving my credences, my *subjective probability function,* Let g be the proposition that the coin is biased .6 against heads, and h the proposition that it is

---

[6] Lewis 1981, 330–1).

[7] See Lewis (1973, 178–9).

[8] Lewis (1973, esp. 178–9; 1981, 329–35).

[9] Sobel.

biased .6 toward heads. Then where subscripts g and h indicate these respective cases,

$$\begin{aligned}\mathcal{E}^A(C) &= \mathcal{P}^A_g(C)\mathcal{B}(g) + \mathcal{P}^A_h(C)\mathcal{B}(h)\\ &= .4\mathcal{B}(g) + .6\mathcal{B}(h) = .5.\end{aligned}$$

My expectation of heads were I to flip the coin is thus my expected value of an objective chance-were. When I speak here of my "expected value", I mean an expected value reckoned in terms of my credences, or subjective probabilities. It is, in effect, a weighted average of the values the objective chance-were might have, for all I know, and the weightings are given by my credences. In this sense, my expectation-were is my subjectively expected value of an objective chance-were: the chance of the coin's landing heads were I to flip it. It is this quantity that I use to decide what odds to accept and what odds to reject in my gambling.[10]

The notion of such an expectation-were generalizes like this: We take a weighted average of all the levels the subject thinks $\mathcal{P}^A(C)$ might take. Let variable x range from 0 to 1, and take an infinitesimal interval dx centered on x. The subject's degree of belief that $\mathcal{P}^A(C)$ lies in this interval dx is

$$\mathcal{B}\Big(P^A(C) \in dx\Big).$$

The average obtained with these weights, in measure-theoretic terms, is

$$\mathcal{E}^A(C) = \int_0^1 x\mathcal{B}\Big(\mathcal{P}^A(C) \in dx\Big),$$

the weighted sum of possible levels of x from 0 to 1. This will be my official definition of an expectation-were.[11]

The expectation-were is what we all agree about when we consider the coin fair. About the objective chance-were function $\mathcal{P}$ we may be in considerable disagreement. Some of us might be *counterfactual determinists,* in the sense of thinking that $\mathcal{P}^A(C)$ must be either 0 or 1. A reasonable counterfactual determinist, though, will think he has no idea in advance which of these two possibilities obtains, and so his expectation-were will be $\frac{1}{2}$. Others of us might be *extreme counterfactual indeterminists,* in the sense of being sure

[10] Skyrms (1984, 100) focuses on this quantity, though his explication of it is different from mine; as we see below. He speaks of the "subjective expectation of conditional chance" (1984, 94), and calls it the "basic assertability value" of the conditional "If A then C". Sobel xxx???.

[11] I'll often be talking about a single credence measure B over possibilities that include objective chances-were as of a time. I'll assume that everything is well-behaved enough to give the integrals I use definite values, without exploring what this requires.

the $\mathcal{P}^A(C) = 1/2$. Such a person's expectation-were is likewise $1/2$. The phenomenon an empiricist needs to explain, then, is not that evidence brings all our beliefs about objective chances-were into convergence. Everyday evidence will always leave scope for extensive disagreement about $\mathcal{P}$. The convergence that some kinds of evidence seem to force is agreement on *expectations* of objective chances-were.

What do expectations-were have to do with causal influence? To speak of the causal influence of a factor, we need to have alternatives in mind. If we ask the degree to which my flipping the coin would tend to cause it to land heads, we mean my flipping it as opposed to not flipping it—not, perversely, my flipping it as opposed to dropping it carefully heads up. What we need to define, then, is degree to which my flipping it *as opposed to* doing nothing would tend to cause the coin to land heads. Where A is "I flip the coin" and B is "I do nothing," then this is

$$\mathcal{P}^A(C) - \mathcal{P}^B(C).$$

The expected degree of causal influence, then, is the expected value of this quantity, which is

$$\mathcal{E}^A(C) - \mathcal{E}^B(C).$$

This is an epistemic expectation of a difference in non-epistemic chances-were, a subjectively expected value of a difference in objective chances-were.

2. Admissible Partitions

Post hoc is not *propter hoc,* and even reliable correlation is not causation. The clearest cases, though, where A correlates with B without causing it have much to do with causation: they are cases in which A and B are influenced by a common cause. Hold this third cause constant, and if A has no causal influence on B, the correlation vanishes. We read off the causal influence of A on B, then, by controlling common causes and observing the correlation between A and B.

That suggests a reductionistic strategy for characterizing causation. The expected degree of causal influence just *is* the average correlation when possible common influences are held constant. Brian Skyrms is one writer who tries this strategy; his proposal goes like this: Find a suitable partition $\mathcal{K}$ of the space of possibilities. Informally, we understand each member K of the partition as a fixed determination of all the possible third causal factors. Then we define

$$\mathcal{E}^A(C) = \sum_{K \in \mathcal{K}} \mathcal{B}(C/AK)\mathcal{B}(K). \qquad (Skyrms)$$

In other words, for each K we define the expectation of C were both A and K to obtain as just $\mathcal{B}(C/AK)$. Then $\mathcal{E}^A(C)$, the expectation of C were A, is an average of these values, weighted by one's degree of belief $\mathcal{B}(K)$ in each of the various K's.[12]

Now of course a circularity looms here—a danger of which Skyrms is well aware. Our goal was to define expected degree of causal influence. We have now done so in terms of a partition $\mathcal{K}$. But I characterized the members of $\mathcal{K}$ in causal terms; each member is a full determination of the various factors that might, for all the subject knows, constitute common causal influences.

Skyrms has his own discussion of how we are to understand the partition $\mathcal{K}$ and the expectation defined in terms of it. My aim here is not to analyze this part of Skyrms's discussion, the part where he tries to complete the reduction of causal notions to non-causal features of belief. I touch on this only at the end or this paper. Rather, I want now to ask how Skyrms' treatment would look if we accepted the naïve objectivistic foundations I laid out above. What would constitute an acceptable partition $\mathcal{K}$? What properties must the partition $\mathcal{K}$ have if the expectation-were is to be given by formula (*Skyrms*)?

First some derived machinery: We can speak of a kind of epistemic conditional expectation-were, what one's expectation-were would be if one gained certain information. Let me indifferently write one's conditional credence in C given K as $\mathcal{B}(C/K)$ or as $\mathcal{B}_K(C)$. Then $\mathcal{E}^A_K(C)$ or $\mathcal{E}^A(C/K)$, one's *expectation, given K, of C were A,* is defined like this:

$$\text{Definition: } \mathcal{E}^A_K(C) = \mathcal{E}^A(C/K) = \int_0^1 x\mathcal{B}_K\Big(\mathcal{P}^A(C) \in dx\Big)$$

Note the difference between the roles of K and the role of A. The role of K is epistemic; we are asking how one's beliefs would change on learning K and nothing else. The role of A is broadly causal: $\mathcal{P}^A$ is a matter of how things would tend to be if A obtained. The *chance* of mangled cars *were* there great noises of screeching and banging is no greater then were there no such noises, whereas my reasonable *credence* in mangled cars *given* my hearing such noises would go steeply up.

Now for any partition $\mathcal{K}$, admissible or not, we have

$$\mathcal{E}^A(C) = \sum_{K\in\mathcal{K}} \mathcal{E}^A_K(C)\mathcal{B}(K) \tag{1}$$

Proof: By the probability calculus, for any proposition S we have

$$\mathcal{B}(S) = \sum_{K\in\mathcal{K}} \mathcal{B}_K(S)\mathcal{B}(K),$$

[12] Skyrms (1984, esp. Chaps. 4-5), in a free rendition that may somewhat misrepresent his intent. His formulas for what I'm calling $\mathcal{E}^A(C)$ are on p. 70 and p. 100. See discussion at the end of this paper.

and so in particular,

$$\mathcal{B}\Big(\mathcal{P}^A(C) \in dx\Big) = \sum_{K \in \mathcal{K}} \mathcal{B}_K\Big(\mathcal{P}^A(C) \in dx\Big)\mathcal{B}(K).$$

Thus we have

$$\begin{aligned}
\mathcal{E}^A(C) &= \int_0^1 x\mathcal{B}\Big(\mathcal{P}^A(C) \in dx\Big) \\
&= \int_0^1 x \sum_{K \in \mathcal{K}} \mathcal{B}_K\Big(\mathcal{P}^A(C) \in dx\Big)\mathcal{B}(K) \\
&= \sum_{K \in \mathcal{K}} \int_0^1 x\mathcal{B}_K\Big(\mathcal{P}^A(C) \in dx\Big)\mathcal{B}(K) \\
&= \sum_{K \in \mathcal{K}} \mathcal{E}^A_K(C)\mathcal{B}(K),
\end{aligned}$$

completing the proof of (1).

Now compare (1) with Skyrms' proposal. We have

$$\mathcal{E}^A(C) = \sum_{K \in \mathcal{K}} \mathcal{E}^A_K(C)\mathcal{B}(K) \tag{1}$$

$$\mathcal{E}^A(C) = \sum_{K \in \mathcal{K}} \mathcal{B}(C/AK)\mathcal{B}(K) \tag{*Skyrms*}$$

A sufficient condition for (*Skyrms*) to be equivalent to (1), then, is

Equivalence Condition:

$$\mathcal{E}^A_K(C) = \mathcal{B}(C/AK) \text{ for all } K \in \mathcal{K} \text{ such that } \mathcal{B}(K) > 0. \tag{2}$$

In saying that this is a sufficient condition for Skyrms' formulation to be equivalent to 1, I am not, of course, saying that it is a necessary condition. Still, it is hard to see how the two formulations could be equivalent otherwise except by sheer coincidence. I shall assume, then, that the two formulations are equivalent in an interesting way only if the Equivalence Condition holds.

3. Guarantees of Equivalence

When, then, should we expect the Equivalence Condition to hold? I continue my excursion into a non-reductionistic strategy for giving empiricistic sense to expectations-were. Think again of the coin: Suppose I am certain that the coin will be flipped. Then my expectation of its landing heads were it flipped is just my degree of belief that it will land heads.

To say this is to deny that I fancy myself clairvoyant. I have no special foreknowledge, hunch, or premonition as to how the coin will land. That is to say, whatever credences I have on the matter have could be gleaned from my credences as to objective chances as of the time the coin might be flipped. Call such credences *non-clairvoyant*.

A general consequence of non-clairvoyance is due to Sobel. In the case of the coin, let A be the proposition that the coin is flipped at time t^*, and let C be the proposition that it lands heads. Let $\mathcal{P}$ measure the objective chances that propositions have as of time t^*. My credence measure $\mathcal{B}$ is non-clairvoyant, and I am certain that A will obtain—that is to say, $\mathcal{B}(A) = 1$. My expectation-were $\mathcal{E}^A(C)$, then, will simply be my credence $\mathcal{B}(C)$: We will have $\mathcal{E}^A(C) = \mathcal{B}(C)$. This requires that $\mathcal{B}(A) = 1$, but we can now drop this requirement. Put the matter in terms not of credence measure $\mathcal{B}$, but of $\mathcal{B}_A$, of $\mathcal{B}$ as it would be updated by knowledge of A. $\mathcal{B}_A$ is still non-clairvoyant, since A is an event that is believed settled at time t^*. we have $\mathcal{B}_A(A)$. Thus by the same reasoning as before,

Sobel Principle: If $\mathcal{B}(A) > 0$, then $\mathcal{E}^A_A(C) = \mathcal{B}_A(C)$.

In other notation, $\mathcal{E}^A(C/A) = \mathcal{B}(C/A)$.

The Sobel Principle is not meant to hold for arbitrary $\mathcal{B}$, A, and $\mathcal{P}^A$. (I speak of $\mathcal{P}^A$, here, because it figures implicitly in the Sobel Principle. $\mathcal{E}^A$ was defined in terms of $\mathcal{P}^A$.) Rather, $\mathcal{B}$, A, and $\mathcal{P}^A$ must be related as follows: Proposition A concerns a particular time t^*—in this case, the time the coin might be flipped. Measure $\mathcal{P}^A$ gives chances-were as of time t^*, as they would be were A to obtain at that time. Credence measure $\mathcal{B}$ must be non-clairvoyant as of time t^*. The Sobel Principle, then, relates $\mathcal{P}^A$ to $\mathcal{B}$: it characterizes non-clairvoyant credences in, among other things, propositions concerning objective chances-were. There is a single time t^* such that (i) A recounts an event at time t^*, (ii) $\mathcal{P}^A$ measures objective chances-were as of time t^*, and (iii) $\mathcal{B}$ measures credences that are non-clairvoyant as of time t^*.

Return, now, to the problem of characterizing suitable partitions $\mathcal{K}$ for applying formula (*Skyrms*)—partitions for which the Equivalence Condition will hold. The partition $\mathcal{K}$ should consist of propositions that one might—in principle, at least—have learned without fancying oneself clairvoyant. In other words, each member K of partition $\mathcal{K}$

must satisfy this condition: If a credence measure $\mathcal{B}_A$ is non-clairvoyant, then so is $\mathcal{B}_A K$. (Again, of course, this talk of non-clairvoyance makes implicit reference to a time, the time of A and $\mathcal{P}^A$.) Since the Sobel principle characterizes non-clairvoyant credences in chances-were, this amounts to saying that the Sobel principle continues to obtain when we conditionalize on a member of partition $\mathcal{K}$. This gives us a condition on the partition $\mathcal{K}$.

Sobel Condition: If $\mathcal{B}(AK) > 0$, then $\mathcal{E}^A_{AK}(C) = \mathcal{B}_{AK}(C)$.

Note the distinction between the Sobel Principle and the Sobel Condition. The Sobel Principle, I've been assuming, characterizes reasonable credences in objective chances-were. Against the background of the Sobel Principle, the Sobel Condition places a requirement on a partition $\mathcal{K}$. It rules out, for instance, the partition {The coin lands heads | The coin doesn't land heads}. For if the coin will be flipped, for all one believes, and how it would land if flipped is objectively chancy, then one could believe one of these things with certainty only if one fancied oneself clairvoyant.

I return in the next section to defenses of the Sobel Principle and the Sobel Condition. First, though, return to formula (*Skyrms*). A sufficient condition for (*Skyrms*) to hold, we noted at the outset, is the Equivalence Condition,

$$\mathcal{E}^A_K(C) = \mathcal{B}(C/AK) \text{ for all } K \in \mathcal{K}(A) \text{ such that } \mathcal{B}(K) > 0 \tag{2}$$

From this and the Sobel Condition, it follows that a sufficient condition for principle (*Skyrms*) to hold is this:

New Equivalence Condition: $\mathcal{E}^A_{AK}(C) = \mathcal{E}^A_K(C)$.

The Sobel Condition and the New Equivalence Condition together, then, are sufficient for formula (*Skyrms*) to hold. When a partition satisfies these two conditions, one's expectations-were given a member K are just conditional credences involving cause-free propositions.

Beliefs about causal influence are packed into both these conditions—as I'll be discussing. What our results tell us is how to go from these prior beliefs to new beliefs about causal influence. When a partition satisfies both these conditions, then we can refine our beliefs about causal influence in the same ways we can update credences in non-causal propositions. The picture that emerges, then, is non-reductionistic: We start out accepting certain truisms concerning causal influence. We then use experience to develop causal beliefs that we lacked in advance of this new experience.

It remains to be shown that the Sobel Condition and the New Equivalence Condition are relevant to this program. When will they obtain? I've already discussed the Sobel Condition: It says that the partition can't involve matters that could be known only with clairvoyance: With hunches as to how objective chances will play out. I discuss this interpretation further in the next Section.

What, though, of the New Equivalence Condition? This condition will turn out to be crucial to gaining reliable causal information from experience. What it requires of each K in the partition is this: that K, in effect, screen off any information that knowledge of A might give about A's own causal influence. Once one knows K, then whether or not A obtains will not be further diagnostic of A's causal influence. Each K must be such that, once one knew it, learning that A was to obtain would give one no further information about the causal influence of A. Let K, for instance, tell the objective weighting of a coin, and A be the news that the coin gets flipped. Even if the coin's getting flipped would indicate to you that it is biased toward heads, once you learned its weighting for sure, learning that it gets flipped would tell you nothing further about its tendency to land heads if flipped. I show an application of such a requirement in the final Section, where it figures in our gleaning causal information from a simple controlled experiment.

4. Non-Clairvoyance

The Sobel Principle, I have claimed, requires simply that one regard oneself as non-claivoyant. One's credences in chances and chances-were yield everything one believes as to how these chances would play out. One's credences as to how the coin is weighted yield, in a canonical way, all one's credences as to how the coin would tend to land were it flipped. In this section, I try to elucidate these interpretations and derive them from simpler principles.[13]

Perhaps we should take the Sobel Principle as an axiom. It characterizes, jointly, the probability measures $\mathcal{P}$ and $\mathcal{B}$ and those propositions A for which we'd be seeking suitable partitions $\mathcal{K}$. It would be more illuminating, though, to analyze see how the Sobel Principle accomplishes this. What, in it's purest form, is the assumption of non-clairvoyance, and what other assumptions would have to be built into a justification of the Sobel principle? One might reject the assumptions I'll make and still accept the Sobel principle, but the following appear to me to be the most elementary assumptions underlying the Sobel Principle's plausibility.

First, we've specified that proposition A concerns an event that takes place at or before the time t^* as of which objective chances and chances-were are being considered.

[13] I have modified this derivation of the Sobel Condition to meet an important objection by James Joyce.

A, say, is that a certain coin is flipped at time t^*. If A obtains, this means, its obtaining is settled at time t^*.

A-Settledness: A implies $\mathcal{P}(A) = 1$

The implication in question here is epistemic: A-settlednss is the condition that $\mathcal{B}_A(\mathcal{P}(A) = 1) = 1$. This is a consequence of one's certainty that by time t^*, A's obtaining or not will have been settled, and not be a matter of how objective chances left open at t^* will subsequently play out.

So far, we've specified very little of the nature of chances-were. We've said informally that measure $\mathcal{P}^A$ measures the chances with which things would happen were A to obtain. We've observed that this makes implicit reference to a time t^*, a time as of when these objective chances are taken, and at which A's obtaining is settled. What assumptions, then, are needed for a plausible derivation of the Sobel Principle? I'll continue to keep the time t^* implicit rather then explicit in my statement of principles: We're considering only a single time that plays it's role, and incorporating a time parameter into our formal apparatus would produce clutter we can avoid for our purposes here. Now we can make do, I'll claim, with weak assumptions. The first two have been implicit in the discussion all along; only the last of is really new to the discussion. The following are axioms characterizing $\mathcal{P}$ and $\mathcal{P}^A$:

Axiom 1. $\mathcal{P}$ is a probability measure.

Axiom 2. $\mathcal{P}^A$ is defined for A, and is a probability measure.

Axiom 3. If $\mathcal{P}^A$ is defined for A and if $\mathcal{P}(A) = 1$, then $\mathcal{P}^A(C) = \mathcal{P}(C)$

Much is still left implicit in these formulations, but they say enough to allow us to derive the Sobel principle.[14] Assume these axioms are believed with certainty. From belief in these axioms, we get

$$\mathcal{P}(A) = 1 \text{ implies } \mathcal{P}^A(C) = \mathcal{P}(C) \tag{3}$$

As with the statement of A-settlednss, the implication in question here is epistemic.[15]

[14] All talk of "probability measures" means on the same Boolean ring $\mathcal{R}$ of propositions, and variables for propositions have propositions of $\mathcal{R}$ as their domain. $\mathcal{R}$ determines a set of possible worlds or atoms; we can think of these as given by the ultrafilters. For each atom w there is are probability measures $\mathcal{P}_w$ and $\mathcal{P}^A_w$. Credences in statements about chances-were are interpreted on this pattern: $\mathcal{B}(\mathcal{P}(C) < x)$ would be the measure of all w such that $\mathcal{P}_w(C) < x$. In a careful treatment, we would have to specify that all the sets in question are measurable, and that all the relevant Lebesgue integrals exist.

[15] I owe (3) to James Joyce.

Now from (3) and A-settlednss, we get the following centering condition.

A-CENTERING: A implies $\mathcal{P}^A(C) = \mathcal{P}(C)$

If the coin will in fact be flipped by time t^*, then the chance (as of t^*) of its landing heads were it flipped will be just its chance (as of t^*) of landing heads. A-Centering says in effect that one is certain of this. Expressed more explicitly, it says that

$$\mathcal{B}_A\Big(\mathcal{P}^A(C) = \mathcal{P}(C)\Big) = 1 \tag{4}$$

So far, we've assumed certain things as believed with certainty: That whether A obtains will be causally settled by time t^*, and that Axioms 1–3 characterize $\mathcal{P}$ and $\mathcal{P}^A$. None of this speaks to non-clairvoyance. The essence of the non-clairvoyance requirement, I think, is contained in Lewis's "Principal Principle" (19xx), which I'll briefly expound in the interpretation I'll be giving it.

This minimal content of a non-clairvoyance requirement on my beliefs is this: Suppose I am certain what the objective chance is, as of now, of a proposition C. Then the degree to which I expect C will just be the objective chance I believe C to have. We have, then,

$$\mathcal{B}_{AK}\Big(C \;\Big/\; \mathcal{P}(C) = x\Big) = x \tag{5}$$

Indeed (5) would hold even if I gained new information, so long as I gained it without clairvoyance. It will hold under conditionalization on any proposition I believe settled. To say that a proposition S is *believed settled,* recall, is to say that S epistemically implies $\mathcal{P}(S) = 1$. In other words, $\mathcal{B}_S(\mathcal{P}(S) = 1) = 1$. Here, then, is a version of Lewis's Principal Principle:

LPP: Suppose proposition S is believed settled. Then

$$\mathcal{B}_S\Big(C \;\Big/\; \mathcal{P}(C) = x\Big) = x$$

Turn now to deriving the Sobel Principle, and to finding a condition sufficient for a partition $\mathcal{K}$ to satisfy the Sobel Condition. By A-settlednss, A is believed settled, and so LPP applies when S is A. Require, then, that all members of the partition $\mathcal{K}$ be believed settled.

PARTITION SETTLEDNESS: Every member K of partion $\mathcal{K}$ is believed settled.

LPP thus applies when S is AK, and so we have

$$\mathcal{B}_{AK}\Big(C \;\Big/\; \mathcal{P}(C) = x\Big) = x \tag{6}$$

From this and A-Centering, the Sobel Condition on $\mathcal{K}$ in turn follows. *Proof:*

$$\begin{aligned}
\mathcal{E}^A_{AK}(C) &= \int_0^1 x\mathcal{B}_{AK}\Big(\mathcal{P}^A(C) \in dx\Big) \\
&= \int_0^1 x\mathcal{B}_{AK}\Big(\mathcal{P}(C) \in dx\Big) && \text{(from } A\text{-Centering)} \\
&= \int_0^1 \mathcal{B}_{AK}\Big(C \;\Big/\; \mathcal{P}(C) = x\Big)\mathcal{B}_{AK}\Big(\mathcal{P}(C) \in dx\Big) && \text{(from (6))} \\
&= \mathcal{B}_{AK}(C) && \text{(probability calculus)}
\end{aligned}$$

Review the assumptions from which the Sobel Principle and the Sobel Condition are derived. We assumed that Axioms 1–3 are believed with certainty, and that proposition A is believed settled. We also assumed LPP. From these assumptions, the Sobel Principle follows. (For this in the above derivation, we let K be the logically true proposition T.) Against this background of assumptions, we have shown, if partition $\mathcal{K}$ satisfies Partition Settledness, then $\mathcal{K}$ satisfies the Sobel Condition.

Hence from the last section, we see the following. Assume:

(a) Axioms 1–3 are believed with certainty.

(b) A is believed settled.

(c) LPP.

Then suppose partition $\mathcal{K}$ satisfies Partition Settledness. Then $\mathcal{K}$ satisfies the Equivalence Condition (2) if and only if $\mathcal{K}$ satisfies the NEC.

The Sobel Condition requires that partition $\mathcal{K}$ consist of propositions that, one believes, one wouldn't have to be clairvoyant to come to know. This means, we can now say, that it consists of proposition for which (6) obtains—partitions which, in this sense, satisfy the conditions of Lewis's Principal Principle.

5. Converging On Expectations-Were: A Controlled Experiment

The treatment I'm deriving from Skyrms is not entirely reductionistic. True, given a partition each member of which satisfies the New Equivalence Condition (NEC), we can indeed dispense with irreducible chance-were, and use (*Skyrms*) to calculate expectations-were. NEC itself, though, is stated in terms of expectations-were, and expectations-were are defined in terms of objective chances-were. We can, to be sure, offer a paraphrase of NEC that takes some of the mystery out of it. Still, it should be emphasized that the paraphrase will have to involve talk about evidence concerning "causal influence". No

total reduction of the metaphysical to the epistemic has been achieved. This I consider no criticism, since I doubt that any such reduction can be achieved; I simply observe that as things have turned out, the feat I suspect impossible hasn't indeed been performed.

It would, of course, be gratifying if we could stick to plain conditional credences, purged of any causal notions. We'd carry so much the less metaphysical baggage, and our empiricism could then be straightforward. Credences and conditional credences, after all, are knowable, in that they approximate observed frequencies. When we observe frequencies for long enough, our conditional credences come to approximate ratios of frequencies we have observed.

I'm proposing, though, that causal notions can't be reduced away. Can I then still be an empiricist? As a non-reductionist, I'll need a story of how evidence can be brought to bear on the metaphysical propositions I'm countenencing. I've given part of such a story: We have background beliefs about expectations-were, I say, and given these, frequency information can yield new expectations-were. Learning about causes is Bayesian, like everything else: With new experience, we update old degrees of credence.

This broad answer, I think, is the right one. We bring evidence to bear on questions of causality against a deep background of causal beliefs. Those causal beliefs themselves can presumably claim the backing of frequency information; how they do this is a question that merits careful investigation. At least in the short run, though, we come to new experiences with causal beliefs already in hand. It is in virtue of these prior convictions that new frequency information can lead to new causal beliefs.

How does this work? Let me explore a crucial kind of case. Controlled experiments are explicitly designed to force agreement on causal questions; how can they do this?

In a controlled experiment, *experimental* and *control* groups are selected in some what that is scrupulously arbitrary—say, by a chance device. Now one firmly entrenched antecedent belief of ours is this: If I settle on what to do by chance, doing so insulates my action from being diagnostic of the causal pattern in which I act. Consider a stock example: We want to know whether providing children with free milk in school will add to their growth rates. We do this by providing free milk to some children and not to others. Now if we let teachers choose on their own which children get the free milk, this may contaminate the results: teachers may direct the the milk goes to the children who most need it. Who gets milk will be diagnostic of who needs it. Comparative rates of growth in the two groups may then reflect not only the causal influence of the milk, but other differences between the two groups such as poverty.[16]

[16] Ref to experiment, Seidenfeld.

We decide, then, to let who gets free milk be determined by a random device, by the flip of a coin. Take a particular child Eliza, then, and define M, G, and C as follows:

M: Eliza gets free milk in school.

G: Eliza grows satisfactorily.

S: Whether M is settled by the fall of a coin.

Information about how the coin lands would not affect our expectation of the causal effect of giving Eliza milk. Antecedently we accept S: say, that Eliza gets milk if and only if the coin lands heads. Learning that she'll get milk, then, is equivalent to learning that the coin will land heads—and this makes no difference to our expectation of satisfactory growth were she to get milk. In other words,

$$\mathcal{E}^M(G/MS) = \mathcal{E}^M(G/S). \tag{7}$$

We think antecedently too that making Eliza's getting free milk depend on the fall of a coin does not alter the causal influence of her getting the milk. That is to say,

$$\mathcal{E}^M(G/S) = \mathcal{E}^M(G). \tag{8}$$

From these two formulas (7) and (8) we obtain

$$\mathcal{E}^M(G/MS) = \mathcal{E}^M(G). \tag{9}$$

Now (9) is an instance of the New Equivalence Condition

$$\mathcal{E}^A(C/KA) = \mathcal{E}^A(C/K). \tag{NEC}$$

with K in NEC becoming the trivial proposition T in (9), and A in NEC becoming MS in (9). Now NEC, we've demonstrated, is a sufficient condition for the Old Equivalence Condition OEC, which for this case is

$$\mathcal{E}^M(G) = \mathcal{B}(G/MS).$$

(In saying that NEC is sufficient for OEC, I am, of course, taking on assumptions (a)–(c) at the end of the previous Section. I don't need specially to assume Partition Settledness, since the partition $\mathcal{K}$, in this case, is the trivial partition, to which the condition applies trivially.)

$\mathcal{B}(G/MS)$ is a straight credence in non-causal propositions, and we can tell the usual story of how convergence on this credence is forced by a long enough string of observations:

eventually the quantity comes to approximate the observed frequency $\mathcal{F}(G/MS)$ of G cases among cases where MS obtains. It approximates the observed frequency—where getting free milk or not is decided by the flip of the coin—of satisfactory growth among children receiving milk.

We have seen, then, how observers who start out agreeing on causal truisms can be forced by shared observations to causal agreement—forced as quickly as they are forced to agreement in their non-causal credences. (7) and (8) are entrenched causal opinions that we bring to the milk/growth experiment. A controlled experiment like this one exploits shared causal truisms to force convergence in other causal matters, in matters where causal beliefs aren't initially shared.

6. Commentary

What are we to say, then, about the "metaphysical baggage" of causality in my account, this talk of evidence concerning causal influence? Are we worse off talking of expectations-were than we would be if we could talk in epistemic terms alone—in terms of subjective probabilities of non-causal, non-counterfactual propositions? On one argument, I might say no. A chief argument for the innocence of talk of credence is that we can, by ingenious decision-theoretic procedures, read credence from a person's choice dispositions. I think that this claim needs to be qualified in certain ways, but that is not to my point here, which is this: If such an argument was supposed to recommend credences, a like argument should recommend expectations-were. Expectations-were, after all, are the weights to use in making decisions. They should have as good behavioristic recommendations as credences.[17]

Still, all direct talk about objective chances-were has dropped out of the discussion. Everything we have learned about it is now filtered through subjective expectations of its value. We started out talking about objective chances-were $\mathcal{P}^A(C)$, but by now, talk about $\mathcal{P}$ has receded into the background, and the things I am saying can be put in terms of conditional expectations-were $\mathcal{E}^A_K(C)$. The same expectations-were conditional on anything we can easily learn can be built on quite different underlying credences in objective chance-were functions. For all practical purposes, then, certain expectations-were can do the work I was calling on objective chances-were to do at the outset.

[17] There has been heated debate about whether expectations-were do give the right weights to use in decision, or whether the weights to use are subjective conditional probabilities of cause-free propositions. Psychological experiments indicate that where the two diverge, people in fact don't use pure expectations-were; see Tversky and Quattrone (19xx). If, somehow, we knew that an agent used expectations-were as decision weights, then aspects of his expectations-were would be revealed in action. That leaves the question of how we could tell, as a matter of radical interpretation, whether the person uses expetations-were $\mathcal{E}^A(C)$ or conditional credences $\mathcal{B}(C/A)$ as decision weights for an act A. This may need investigating.

Expectations-were will be practically objective if they stem from a partition that meets the conditions we have examined and that are resiliant under all finer suitable subpartitions that it is practicable to investigate.

An empiricist should be happy with expectations-were so long as they respond suitably to evidence.[18] That does not mean, though, that an empiricist need be happy with my explication of $\mathcal{E}^A$. That explication was in terms of credences in objective chances-were $\mathcal{P}^A(C)$, and though our subjective expectations of these values converge to common levels with suitable common experience, our credences in particular values need not. Or at least ordinary controlled experiments do not force such a convergence, even among people with reasonable priors.[19]

What is the upshot? A Bayesian empiricist can well accept that causal belief are not reducible to degrees of belief in non-causal propositions.[20] On the Bayesian idealization, a

---

[18] That is the standard that self-avowed empiricist and pragmatist Skyrms uses to determine whether a statement is "metaphysical" for a person or group of people: to be *metaphysical* for a person is to be insensitive, in that person's credences, to all possible evidence (112). Skyrms, of course, does not claim that chance and what I am calling chances-were are metaphysical in this sense—on the contrary. He does claim that they are eliminable; see below.

[19] One might also object that the entire objectivistic framework I have given is lamentably metaphysical. According to Skyrms, a claim, say, that for a given coin the chance of heads is $1/2$ is empirically significant, whereas such claims as "that chances are really 'out there' and not supervenient on manifest properties," or that chances can disagree with limiting relative frequency, are metaphysical. "It is, I think, the metaphysical admixture in 'chance' that de Finetti really objects to." He looks to analytical tools "for separating the empirically meaningful component of chance from the metaphysical admixture" 18. However, he does not demand that all metaphysics be eliminated: a group's framework principles, he says, may legitimately be metaphysical for them. They then don't constitute knowledge or candidates for knowledge, but they may still be sensible (114).

[20] Skyrms himself does claim to show how the concepts of chance and causal necessity, along with subjunctive conditional notions, can be eliminated (1984, 115, 118). I won't attempt comment here on his ingenious eliminative treatment of physical chance (Chap. 3), but I am puzzled by his claims to have shown how to "eliminate" causal necessity and subjunctive conditionals—and along with them, I take it, expectations-were. He clearly does devise an intelligible notion of conditional chance *relative to* a family of partitions. How, though, do we pass from this to the kind of a non-relative notion we need if we are to guide ourselves by expectations-were? I take this paper to offer an answer—but the answer does not eliminate cause-laden notions; it invokes them. Skyrms himself isn't claiming "that the choice of an appropriate partition will always be easy. In real life, various pragmatic factors may be relevant and all of them together may still underdetermine the correct choice" (98). What, we need to ask, are these "pragmatic factors"? Do they include matters of which events, to our settled belief, are causally uninfluenced by which other events? Such matters are cause-laden even if truistic—and so if it comes down to these, we haven't eliminated causal notions. "What is required of the Ks," he writes, "is that they form a partition which captures the *decision maker's* beliefs about conditional chance" (138, note 14). This makes it seem that to be decision makers, we'd better have beliefs about "conditional chance". If this means what I've been calling "chances-were", that's what I've been claiming. If it means conditional probabilities $\mathcal{P}(C/A)$ for some probability function $\mathcal{P}$, this isn't equal to the corresponding chance-were $\mathcal{P}^A$ for arbitrary proposition A,

rational thinker meets life with a set of credences in non-causal propositions, and uses new experience to update these prior credences. Just so, I am proposing, a rational thinker will approach life with a set of causality-laden prior expectations-were. She then uses new experience to update these prior expectations-were.

Expectations-were are explicable in terms of subjective credences the thinker might have in an array of propositions about causal influence, about chances-were. They do not reduce, I am assuming, to any property of her credences in causality-free propositions—to any property that can be described without invoking, directly or indirectly, some notion that amounts to a notion of causal influence. We do not, then, explain causal beliefs as configurations of non-causal beliefs, but we do show them to be as empirically well behaved as are non-causal beliefs.

A chief triumph of Bayesian theory lies in demonstrations that thinkers with widely divergent prior credences tend to converge as experience accumulates. These demonstrations depend, though, on the priors' of different thinkers sharing important characteristics: it is by no means a theorem that enough common experience will force convergence among thinkers with any array of prior credences whatsoever. Likewise with the convergence of expectations-were that controlled experiments force: the convergence depends on our prior expectations'-were sharing important features. We must accept together such things as that a child's need for milk does not affect how a coin lands, and that the setup of a reported experiment assures that the landing of the coin affects growth only by affecting whether the child receives milk.

I have been trying systematically to identify what these common features of causal credences must be for our expectations-were to converge as readily as do our credences in causality-free propositions. Apart from belief in a few structural features of causality (given by axioms 1–3 above), these amount to the following: Common beliefs about when certain matters are settled, such as that by the time a coin lands heads, its chance of having landed heads is one. Common attitudes of regarding oneself as non-clairvoyant, in that the credences of each of us satisfy LPP. Finally, two common features of our respective credences concerning an experimental setup. In the milk example, these are that we both regard the experimental setup as assuring (8) that whether (a) a child's getting milk or not is determined by the flip of a coin, is no indication of (b) how much the child would tend to grow were he to receive the milk; and (7) that where the child's getting milk or not is determined by a flip of the coin, whether the child gets milk is no indication of how much he would tend to grow were he to get milk.

---

and we need to pass to $P(C/AK)$ for a member of a suitable partition. We are then back to asking what makes a partition suitable. Either way, it is unclear that cause-laden notions have been eliminated from the decision-maker's thinking.

Causal truisms, then, along with shared beliefs about an experimental setup, allow certain of our expectations-were to approach the same level as evidence from such an experimental setup accumulates.[21]

7. Relations to Pearl

Judea Pearl (2000 and elsewhere) has revived and largely invented the study of our knowledge of causality. He discusses counterfactuals (2000, pp. 201–257 and elsewhere) and incorporates them into his theory, which is far more general than anything I have attempted here. He argues, however, that the basis of a theory of inferred causation is better found not in a chancy conditional of the kind I take as basic in this paper, but in the notion of an "action" or "intervention". He represents the "action" of setting value X to x as $do(x)$. His notation for the causal effect of value X on value Y is '$P(y/do(x))$' (p. 70). This makes it look as if this is a standard conditional probability, but conditioned on an action or intervention, which is a special kind of event or something of the sort. We need to examine whether there is anything that an "action" could be that would fit this conception. In *Causality* and elsewhere, though, he (sometimes jointly with a coworker) also talks in more directly causal terms. Shpitser and Pearl (2007) explain, "$do(\mathbf{x})$ stands for hypothetically forcing variables $\mathbf{X}$ to attain values $\mathbf{x}$ regardless of the factors that influence $\mathbf{X}$ in the model while leaving all other functional relations unaltered" (first page). The talk of "forcing" looks causal, though what is being explicitly described is an operation on a mathematical model. That leaves us to ask what such an operation on a model represents in the world, and what its relation is to the apparatus in this paper. In its current version, though, this paper leaves these questions unexplored. I greatly admire Pearl's work, but leave questions of the relation of his conceptual apparatus to that in this paper for further study. I also postpone questions of whether this paper contributes to our understanding of how we can learn of causal relations in any way that is independent of Pearl's treatments of the question, and whether the conceptual apparatus in this paper has advantages over Pearl's.[22]

---

[21] Notes on Lewis to go somewhere: Lewis (1980, 111-3) discusses "a broadly Humean doctrine" that causes, chances, and the like all supervene on "particular fact", which is not itself cause-laden or chance-laden. This is, he says, "something I would very much like to believe if at all possible," but trying coherently to believe it confronts him with dilemmas which he discusses. Still, says he, "Neither is it very easy to believe in features of the world that are not supervenient on particular fact" (113). I intend my discussion to make this easier: We learn about such features, I claim to be showing, in very much the same way as we learn about "particular fact".

[22] I am grateful to Jim Joyce for discussions of issues in this paper. My thinking has also been affected by discussions with Carl Hoefer, but his influence is not reflected in this version of the paper.

References

[Most references remain to be filled in.]

Blackburn

Gibbard

Lewis, David (1973), "Causation". In Lewis (1986).

Lewis, David K. (1980). "A Subjectivist's Guide to Objective Chance". In Lewis (1986).

Lewis, David K. (1981). "Causal Decision Theory". In Lewis (1986).

Lewis, David K. (1986). *Philosophical Papers*, Volume II (New York: Oxford University Press).

Pearl, Judea (2000). *Causality: Models, Reasoning, and Inference* (New York: Cambridge University Press).

Skyrms, Brian (1984). *Pragmatics and Empiricism* (New Haven: Yale University Press).

Shpitser, Ilya and Pearl, Judea (2007). "What Counterfactuals Can Be Tested". Technical Report R-334, Cognitive Systems Laboratory, Department of Computer Science, UCLA. (See website http://bayes.cs.ucla.edu/jp_home.html).

Tversky and Quattrone.