Chapter 23
Causal Decision Theory

James M. Joyce and Allan Gibbard

- EXAMPLE: *Prisoner's Dilemma with Twin (PDT).* You are caught in a standard, one-shot prisoner's dilemma (diagram next page), and the other player is your twin. You don't know for sure what Twin will do, but you know that Twin is amazingly like you psychologically. What you do, he or she too will likely do: news that you were going to rat would be good indication that Twin will rat, and news that you were going to keep mum would be a good sign that Twin will keep mum. Your sole goal is to minimize your own time in jail: Family feelings affect you not, and you care not a whit about loyalty, returning good for good, or how long Twin spends in jail. What course of action is rational for you in pursuit of your goals?

Many will find the answer easy—though they may disagree with each other on which the answer is. A standard line on the prisoner's dilemma rests on *dominance:* What you do won't affect what Twin does. Twin may rat or keep mum, but in either

In the nearly 20 years since this article was written there has been a revolution in the understanding of causal and counterfactual reasoning. This revolution had its roots in early work by Rubin (1974), Holland (1986) and Robbins (1986), which gave rise to the so-called "potential outcomes" framework. At roughly the same time the closely related "structural equations/causal graphs" approach was being developed and used to great effect by Spirtes et al. (1993), and Pearl (2000). In both treatments counterfactual reasoning plays a leading role in causal inference, just as in causal decision theory. While the core claims of this article remain true, and the basic structure of causal decision theory remains intact, these new models of provide us with far more sophisticated ways of representing and identifying causal relationships than were available and widely known when we wrote. As a result, some of our remarks about "the need for new advances in understanding of localization in relation to rational belief" have been rendered moot. Readers are encouraged to investigate these new developments, which we see as great advances.

J.M. Joyce (✉) • A. Gibbard
University of Michigan, Ann Arbor, MI, USA
e-mail: jjoyce@umich.edu; gibbard@umich.edu

H. Arló-Costa et al. (eds.), *Readings in Formal Epistemology*, Springer Graduate
Texts in Philosophy 1, DOI 10.1007/978-3-319-20451-2_23

case, you yourself will do better to rat. Whichever Twin is doing, you would spend less time in jail if you were to rat than if you were to keep mum. Therefore the rational way to minimize your own time in jail is to rat.

	Mum	Rat
Mum	−1, −1	−10, 0
Rat	0, −10	−9, −9

Prisoner's Dilemma

Another line of argument leads to the opposite conclusion. Assess each act by its *auspiciousness*, by how welcome the news would be that you were about to perform it. News that you're about to rat would indicate that Twin is likewise about to rat. That's bad news; it means a long time in jail, for you as well as for Twin. News that you're about to keep mum, on the other hand, would be good news: It indicates that Twin is likewise about to keep mum, and your both keeping mum will mean a short time in jail. Keeping mum, then, is the *auspicious* act, and so—in terms of your selfish goals—you achieve best prospects by keeping mum.[1]

The two lines of reasoning, then, lead to opposite conclusions. One or the other, to be sure, may strike a reader as obviously wrong. Still, if one of them is cogent and the other not, decision theory should tell us why. Standard theories haven't spoken, though, with one voice on this matter. Savage himself (1972) was mostly silent on issues that would decide between the two lines: his system could be read in more than one way, and the few pertinent remarks he left us point in opposing directions. Various other decision-theoretic systems do have implications for this matter. Some imply that the argument from auspiciousness is correct: the principle of dominance, these systems entail, doesn't properly apply to a case like PDT. Taking the other side, a group called—perhaps somewhat misleadingly—*causal* decision theorists have formulated systems according to which the principle of dominance does apply to this case, and the rational thing to do is to rat.

"Causal" theorists maintain that decision theory requires a notion of causal dependency, explicit or implicit. Otherwise, they say, the theory will yield the wrong prescription for cases like PDT. We touch below on how causal notions might be made explicit for decision theorists' purposes, and how causality might be vindicated as empirically respectable. Auspiciousness theorists—or *evidential*

[1] An interesting model of Prisoner's Dilemma with a twin can be found in Howard (1988). Howard, who endorses a version of the auspiciousness argument, shows how to write a Basic program for playing the game which is capable of recognizing and cooperating with programs that are copies of itself.

decision theorists, as they are called in the literature—have no need for causal terms in their theory: they manage everything with standard subjective probabilities and conditional probabilities. Some evidential theorists deny that their theory, properly construed or developed, really does say not to rat in PDT. They deny, then, that causal notions must be introduced into decision theory, even if causal theorists are right about what to do in this case. We touch below on debates between "causal" theorists and this camp of "evidential" theorists, but mostly stick with the "causal" theory, explaining it and examining its potentialities.[2]

Cases with the structure of PDT can't be rare. The prisoner's dilemma itself is a parable, but economics, politics, war, and the like will be full of cases where one's own acts suggest how others are acting. Consider, for instance, a sophisticated speculator playing a market. Mustn't he reasonably take himself to model other sophisticated players? Why should he be unique? A rational agent interacting with others must escape the hubris of thinking that only he is smart and insightful—but then he'll have to take himself as a likely model for the schemings and reasonings of others. In such cases, different versions of decision theory may prescribe incompatible actions.

Dominance and Expected Utility: Two Versions

Savage (1972) encoded decision problems as matrices, with columns indicating "states of the world". For a Savage matrix with states $S_1, \ldots, S_n$, the expected utility of an act A is calculated by the Savage formula

$$\mathcal{V}(A) = \sum_{i=1}^{n} \rho(S_i)u(A, S_i), \tag{23.1}$$

where $u(A, S_i)$ is the utility of act A for state S_i and $\rho(S_i)$ is the subjective probability of S_i. From (23.1) follows a principle which we'll call the *Unqualified Principle of Dominance*, or *UPD:*

- UPD: If for each S_i, $u(A, S_i) > u(B, S_i)$, then $\mathcal{V}(A) > \mathcal{V}(B)$.

Which Savage matrix correctly represents a problem, though, must be decided with care, as is shown by a spoof due to Jeffrey (1967), p. 8:

[2] Nozick (1969) introduced PDT and other cases of this kind, focusing his discussion on Newcomb's problem, which he credits to physicist William Newcomb. He makes many of the points that causal theorists have come to accept, but recognizes only one kind of expected utility, the one we are calling auspiciousness. Stalnaker originated causal expected utility in a 1972 letter published only much later (Stalnaker 1981). Gibbard and Harper (1978) proselytize Stalnaker's proposal, and Lewis (1981) gives an alternative formulation which we discuss below. Gibbard and Harper (1978) and Lewis (1979b) also discuss PDT along with Newcomb's problem.

- EXAMPLE: *Better Red Than Dead (BRD)*. I'm an old-time American cold warrior with scruples, deciding whether or not my country is to disarm unilaterally. I construct a matrix as follows: My two possible states are that the Soviets invade and that they don't. In case they invade, better red than dead; in case they don't, better rich than poor. In either case, unilateral disarmament beats armament, and so by dominance, I conclude, it is rational to disarm.

Now whether or not unilateral disarmament would be rational all told, this argument can't be right. As the proponent of deterrence will point out, unilateral disarmament may decrease the likelihood of the Soviets' invading, and a scenario in which they don't invade is better than one in which they do. The argument from "dominance" treats these considerations as irrelevant—even if the Soviets are sure to invade if we disarm and to hold back if we arm.

Savage's states, then, must be act-independent: They must obtain or not independently of what the agent does. How, then, shall we construe this requirement? The first answer was developed, independently, by Jeffrey (1967) and by Luce and Krantz (1971): For dominance correctly to apply, they say, the states must be stochastically (or probabilistically) independent of acts. Where acts $A_1 \ \ldots \ A_m$ are open to the agent and $\rho(S/A_j)$ is the standard conditional probability of S given A_j,[3] S is stochastically act-independent iff

$$\rho\,(S/A_1) = \rho\,(S/A_2) = \cdots = \rho\,(S/A_m) = \rho(S) \tag{23.2}$$

The probabilities in question are subjective, or at any rate epistemic. Formula (23.2), then, means roughly that learning what one's going to do won't affect one's credence in S. One's act won't be evidence for whether S obtains. Requirement (23.2), then, is that any state S be *evidentially* act-independent. Theories that entail that (23.2) is what's needed for the principle of dominance to hold good we can thus call *evidential* theories of decision.

As applied to the prisoner's dilemma with your twin, evidential decision theory must treat the dominance argument as bad. Twin's act fails to be independent of yours evidentially. Let proposition C_t be that Twin cooperates, let C_y be that you cooperate, and let D_y be that you defect. C_t is not evidentially independent of what you do: $\rho(C_t/C_y)$ is high whereas $\rho(D_t/C_y)$ is low, since your cooperating is good evidence that she is cooperating, whereas your defecting is good evidence that she is defecting. Condition (23.2), then, won't hold for this case.

Evidential decision theory holds that the Savage formula (23.1) applies when condition (23.2) holds—that is, when the states of the matrix are evidentially act-independent. Requirements of act-independence, though, could be dropped if we changed formula (23.1) for computing expected utility. Use as weights not the probability of each state, as in (23.1), but its conditional probability—its probability conditional on the act's being performed. In this more general formulation, we have

[3] More precisely, A_i is the *proposition* that one performs a particular one of the alternative acts open to one in one's circumstances. We reserve the notation $\rho(S|A_j)$ for a more general use later in this chapter.

$$\mathcal{V}(A) = \sum_{i=1}^{n} \rho(S_i/A)u(A, S_i). \tag{23.3}$$

The Savage formula (23.1) is then a special case of (23.3), for conditions of evidential act-independence.[4] Since UPD follows from (23.1), it follows from (23.3) plus condition (23.2) of evidential act-independence. Evidential decision theory, then, has (23.3) as its general formula for expected utility. Its version of the principle of dominance is UPD qualified by condition (23.2) of evidential act-independence. In general, it recommends using "auspiciousness" to guide choices: a rational agent should select an act whose performance would be best news for him—roughly, an act that provides best evidence for thinking desirable outcomes will obtain.

Evidential theory has the advantage of avoiding philosophically suspect talk of causality: Its general formula (23.3) sticks to mathematical operations on conditional probabilities, and likewise, its requirement (23.2) of evidential act-independence—its condition, that is, for the Savage formula (23.1) to apply to a matrix—is couched in terms of conditional probabilities.

The causal theorist, in contrast, maintains that to apply a principle of dominance correctly, one can't avoid judgments of causality. One must form degrees of belief as to the causal structure of the world; one must have views on what is causally independent of what. Belief in Twin's causal isolation is a case in point. Dominance applies to PDT, their contention is, because you and your twin are causally isolated—and you know it. What you do, you know, will in no way affect what twin does. The argument, then, invokes a causal notion: the notion of what will and what won't causally *affect* what else. Causal decision theory then recommends using causal efficacy to guide choices: It holds, roughly, that a rational agent should select an act whose performance would be likely to bring about desirable results.

The causal decision theorist's requirement on a state S of a Savage matrix, then, is the following: The agent must accept that nothing he can do would causally affect whether or not S obtains. For each act A open to him, he must be certain that S's obtaining would not be causally affected by his doing A.

Can the causal theorist find a formula for expected utility that dispenses with this requirement of believed causal act-independence? A way to do so was proposed by Stalnaker (1968); see also Gibbard and Harper (1978). It requires a special conditional connective, which we'll render '$\Box\!\!\rightarrow$'. Read '$A \Box\!\!\rightarrow B$' as saying, "If A obtained then B would." In other words, either A's obtaining would cause B to obtain, or B obtains independently (causally) of whether or not A obtains. Then to say that S is causally independent of which act $A_1 \ldots A_n$ one performs is to say this: Either S would hold whatever one did, or whatever one did S would fail to hold. In other words, for every act A_i, we have $A_i \Box\!\!\rightarrow S$ iff S. We can now generalize the Savage formula (23.1) for the causal theorist's kind of expected utility. Use as weights, now, the probabilities of conditionals $\rho(A \Box\!\!\rightarrow Si)$, as follows:

[4] Formula (11.1) is introduced by Jeffrey (1967) and by Luce and Krantz (1971).

$$\mathcal{U}(A) = \sum_{i=1}^{n} \rho(A \,\Box\!\!\rightarrow S_i) u(A, S_i) \tag{23.4}$$

Call this $\mathcal{U}(A)$ the *instrumental expected utility* of act A. The Savage formula

$$\mathcal{U}(A) = \sum_{i=1}^{n} \rho(S_i) u(A, S_i), \tag{23.5}$$

is then (23.4) for the special case where the following condition holds:

$$\text{For each } S_i, \rho\,(A \,\Box\!\!\rightarrow S_i) = \rho\,(S_i)\,. \tag{23.6}$$

A sufficient condition for this to hold is that, with probability one, S_i is causally independent of A—in other words, that

$$\text{For each } S_i, \rho\,([A \,\Box\!\!\rightarrow S_i] \leftrightarrow S_i) = 1. \tag{23.7}$$

Note that for the prisoner's dilemma with twin, condition (23.7) does hold. Twin, you know, is causally isolated from you. You know, then, that whether Twin would defect if you were to defect is just a matter of whether twin is going to defect anyway. In other words, you know that $D_y \,\Box\!\!\rightarrow D_t$ holds iff D_t holds, and so for you, $\rho([D_y \,\Box\!\!\rightarrow D_t] \leftrightarrow D_t) = 1$. This is an instance of (23.7), and similar informal arguments establish the other needed instances of (23.7) for the case.

In short, then, causal decision theory can be formulated taking formula (23.4) for instrumental expected utility as basic. It is instrumental expected utility as given by (23.4), the causal theorist claims, that is to guide choice. The Savage formula is then a special case of (23.4), for conditions of known causal act-independence—where (23.7) holds, so that for each state S_i, $\rho(A \,\Box\!\!\rightarrow S_i) = \rho(S_i)$. The Unqualified Principle of Dominance for $\mathcal{U}$ is

- UPD: If for each S_i, $u(A, S_i) > u(B, S_i)$, then $\mathcal{U}(A) > \mathcal{U}(B)$.

Causal decision theory, in this formulation, has (23.4) as its general formula for the instrumental expected utility that is to guide choice, and its own version of the principle of dominance: UPD qualified by condition (23.7) of known causal act-independence.

Evidential and causal decision theorists, in short, accept different general formulas for the expected utility that is to guide choice, and consequently, they accept different conditions for the Savage formula to apply, and different principles of dominance. Causal theory—in the formulation we've been expounding—adopts (23.4) as its formula for expected utility, whereas evidential theory adopts (23.3). Causal theory, in other words, weighs the values of outcomes by the probabilities of the relevant conditionals, $\rho(A \,\Box\!\!\rightarrow S_i)$, whereas evidential theory weighs them by the relevant conditional probabilities $\rho(S_i/A)$. Different conditions, then, suffice, according to the two theories, for the Savage formula correctly to apply to a matrix, and consequently for UPD to apply. That makes for distinct principles of dominance: For the causal theorist, UPD qualified by condition (23.7) of known causal act-independence, and for the evidential theorist, UPD qualified by condition (23.2) of evidential act-independence.

Conditionals and Their Probabilities

What, then, is the contrast on which all this hinges: the contrast between the probability $\rho(A \;\Box\!\!\rightarrow S)$ of a conditional $A \;\Box\!\!\rightarrow S$ and the corresponding conditional probability $\rho(S/A)$? Where probability measure ρ gives your *credences*—your degrees of belief—the conditional probability $\rho(S/A)$ is the degree to which you'd believe S if you learned A and nothing else. In the prisoner's dilemma with your twin, then, $\rho(D_t/D_y)$ measures how much you'd expect twin to rat on learning that you yourself were about to rat. If $\rho(D_t/D_y) \neq \rho(D_t/C_y)$, that doesn't mean that D_t is in any way causally dependent on whether D_y or C_y obtains. It just means that your act is somehow *diagnostic* of Twin's. Correlation is not causation. Probability $\rho(D_y \;\Box\!\!\rightarrow D_t)$, on the other hand, is the degree to which you believe that if you were to defect, then Twin would. There are two circumstances in which this would obtain: Either Twin is about to defect whatever you do, or your defecting would cause Twin to defect. To the degree to which $\rho(D_y \;\Box\!\!\rightarrow D_t) > \rho(D_y \;\Box\!\!\rightarrow C_t)$, you give some credence to the proposition $[D_y \;\Box\!\!\rightarrow D_t] \& \neg [D_y \;\Box\!\!\rightarrow C_t]$, that twin would defect if you did, but not if you cooperated. This is credence in the proposition that your act will make a causal difference.

In daily life, we guide ourselves by judgments that seem to be conditional: What would happen if we did one thing, or did another? What would be the effects of the various alternatives we contemplate? We make judgments on these matters and cope with our uncertainties. Classic formulations of decision theory did not explicitly formalize such notions: notions of causal effects or dependency, or of "what would happen if". In Ramsey's and Savage's versions, causal dependency may be implicit in the representational apparatus, but this formal apparatus is open to interpretation. Other theorists had hoped that whatever causal or "would" beliefs are involved in rational decisions could be captured in the structure of an agent's conditional probabilities for non-causal propositions or events.

This last maneuver might have great advantages if it worked, but causal decision theorists argue that it doesn't. Causal or "would" notions must somehow be introduced into decision theory, they claim, if the structure of decision is to be elucidated by the theory. The introduction can be explicit, as in the general forumula (23.4) for $\mathcal{U}$ above, or it can be in the glosses we give—say, in interpretations of the Savage formula (23.1) or (23.5). If causal theorists are right, then, the theory of subjective conditional probability won't give us all we need for describing the beliefs relevant to decision. We'll need some way of displaying such beliefs and theorizing about them.

Causal theorists have differed, though, on how causal beliefs are best represented. So far, we've spoken in Stalnaker's terms, but we need to say more on what his treatment consists in, and what some of the alternatives might be for representing causal decision theory.

First, some terminology. Savage spoke of "states" and "events", and distinguished these from "acts" or "strategies". The philosophers who developed causal decision theory often spoke of "propositions", and include as propositions not only Savage's "events" and "states", but also acts and strategies. That is to say, propositions can characterize not only what happens independently of the agent, but also what the agent does—or even what he *would* do in various eventualities. A proposition can say that I perform act a or adopt strategy s. Such propositions can be objects of belief and of desire, and so can be assigned credences (subjective probabilities) and utilities.

Let A, then, be the proposition that I perform act a. Stalnaker constructs a conditional proposition $A \mathbin{\square\!\!\rightarrow} B$, which we read as "If I did A, then B would obtain." How does such a conditional proposition work? Much as Savage treats an event as a set of states, so Stalnaker treats a proposition as a set of *possible worlds* or maximally specific ways things might have been. Abstractly, the connective '$\square\!\!\rightarrow$' is a two-place propositional function: To each pair of propositions it assigns a proposition.

Stalnaker hoped originally that this conditional function $\square\!\!\rightarrow$ could be defined so that the probability of a conditional is always the corresponding conditional probability: So that whenever $\rho(C/A)$ is defined, $\rho(A \mathbin{\square\!\!\rightarrow} C) = \rho(C/A)$. Lewis (1976) proved that—with trivial exceptions—no such equality will survive conditionalization. Read ρ_A, in what follows, as probability measure ρ conditioned on A, so that by definition, $\rho_A(C) = \rho(C/A)$. What Lewis showed impossible is this: that for all propositions A, C and B for which $\rho(A\&B) > 0$, one has $\rho_B(A \mathbin{\square\!\!\rightarrow} \mathrm{C}) = \rho_B(C/A)$. For if this did obtain, then one would have both $\rho_C(A \mathbin{\square\!\!\rightarrow} C) = \rho_C(\mathrm{C/A})$ and $\rho_{\neg C}$ $(A \text{ W } C) = \rho_{\neg C}(C/A)$. But then

$$\begin{aligned}
\rho\,(A \mathbin{\square\!\!\rightarrow} C) &= \rho_C\,(A \mathbin{\square\!\!\rightarrow} C)\,\rho(C) + \rho_{\neg C}\,(A \mathbin{\square\!\!\rightarrow} C)\,\rho\,(\neg C)\\
&= \rho_C\,(C/A)\,\rho(C) + \rho_{\neg C}\,(C/A)\,\rho\,(\neg C)\\
&= 1 \cdot \rho(C) + 0 \cdot \rho\,(\neg C)\\
&= \rho(C)
\end{aligned}$$

We'd have $\rho(A \mathbin{\square\!\!\rightarrow} C) = \rho(C/A)$, then, at most when $\rho(C) = \rho(C/A)$. No such equality can survive conditionalization on an arbitrary proposition.

How, then, should we interpret the probability $\rho(A \mathbin{\square\!\!\rightarrow} C)$ of a conditional proposition $A \mathbin{\square\!\!\rightarrow} C$, if it is not in general the conditional probability $\rho(C/A)$. Many languages contrast two forms of conditionals, with pairs like this one[5]:

[5] Adams (1975) examines pairs like this.

$$\text{If Shakespeare didn't write } \textit{Hamlet}\text{, someone else did.} \tag{23.8}$$

$$\text{If Shakespeare hadn't written } \textit{Hamlet}\text{, someone else would have.} \tag{23.9}$$

Conditionals like (23.8) are often called *indicative*, and conditionals like (23.9) *subjunctive* or *counterfactual*. Now indicative conditional (23.8) seems epistemic: To evaluate it, you might take on, hypothetically, news that Shakespeare didn't write Hamlet. Don't change anything you now firmly accept, except as you would if this news were now to arrive. See, then, if given this news, you think that someone else did write Hamlet. You will, because you are so firmly convinced that Hamlet was written by someone, whether or not the writer was Shakespeare. The rule for this can be put in terms of a thinker's subjecive probabilities—or her *credences*, as we shall say: indicative conditional (23.8) is acceptable to anyone with a sufficiently high conditional credence $\rho(E/D)$ that someone else wrote Hamlet given that Shakespeare didn't. Subjunctive conditional (23.9) works differently: If you believe that Shakespeare did write Hamlet, you will find (23.9) incredible. You'll accept (23.8), but have near zero credence in (23.9). Your conditional credence $\rho(E/D)$ in someone else's having written Hamlet given that Shakespeare didn't will be high, but your credence $\rho(D \mathbin{\Box\!\!\to} E)$ in the subjunctive conditional proposition (23.9) will be near zero. Here, then, is a case where one's credence in a conditional proposition (23.9) diverges from one's corresponding conditional credence. Speaking in terms of the subjective "probabilities" that we have been calling credences, we can put the matter like this: the probability $\rho(D \mathbin{\Box\!\!\to} E)$ of a subjunctive conditional may differ from the corresponding conditional probability $\rho(E/D)$.

The reason for this difference lies in the meaning the $\Box\!\!\to$ operator. Stalnaker (1968) puts his account of conditionals in terms of alternative "possible worlds". A world is much like a "state" in Savage's framework (except that it will include a strategy one might adopt and its consequences). Think of a possible world as a maximally specific way things might have been, or a maximally specific consistent proposition that fully describes a way things might have been. Now to say what the proposition $A \mathbin{\Box\!\!\to} C$ is, we have to say what conditions must obtain for it to be true. There is no difficulty when the antecedent A obtains, for then, clearly, $A \mathbin{\Box\!\!\to} C$ holds true if and only if C obtains. The puzzle is for cases where A is false. In those situations, Stalnaker proposes that we imagine the possible world w^A in which A is true, and that otherwise is most *similar* to our actual world in relevant respects. $A \mathbin{\Box\!\!\to} C$, then, holds true iff C holds in this world w^A. Stalnaker and Thomason offered a rigorous semantics and representation theorem for this explication. Stalnaker's distinctive axioms are these[6]:

- *Intermediate strength:* If A necessitates B, then $A \mathbin{\Box\!\!\to} B$, and if $A \mathbin{\Box\!\!\to} B$, then $\neg(A \& \neg B)$
- *Conditional non-contradiction:* For possible A, $\neg[(A \mathbin{\Box\!\!\to} B) \& (A \mathbin{\Box\!\!\to} \neg B)]$.

[6] Stalnaker (1968) p. 106, Stalnaker and Thomason (1970), slightly modified.

- *Distribution:* If $A \mathrel{\Box\!\!\rightarrow} (B \vee C)$, *then* $(A \mathrel{\Box\!\!\rightarrow} B) \vee (A \mathrel{\Box\!\!\rightarrow} C)$.
- *Suppositional equivalence:* If $(A \mathrel{\Box\!\!\rightarrow} B)$ and $(B \mathrel{\Box\!\!\rightarrow} A)$, then $(A \mathrel{\Box\!\!\rightarrow} C)$ iff $(B \mathrel{\Box\!\!\rightarrow} C)$.

All this leaves mysterious the notion of *relevant similarity* invoked by Stalnaker's account. Formal axioms are easy: Stalnaker speaks of a *selection function* f which assigns a world $f(A, w) = w^A$ to each proposition A that has the possibility of being true. A compelling logic for $\Box\!\!\rightarrow$ can be developed on this basis.

How, though, do we interpret the notion of "relevant similarity" when applying this formal apparatus to real-life decision problems? Intuitive overall likeness of worlds won't do. Nixon, imagine, had in the Oval Office a red button to launch the missiles.[7] In his despair he considered pushing it, but drew back. We can say, "If Nixon had pushed the button, nuclear holocaust would have ensued." This is true, we would judge, if the apparatus was in working order and without safeguards. Of the worlds in which Nixon pushes the button, though, the one most similar overall to the actual world would be not one in which all was destroyed, but one in which the apparatus malfunctioned—a wire became temporarily nonconducting, say. After all, a world in which the missiles were launched would surely have a future radically different from the actual world. Little in any city would look the same.

Clearly, then, we cannot look to overall similarity of worlds to cash out the type of "relevant" similarity needed in the evaluation of subjunctive conditionals. Rather, we want to know what would have ensued from initial conditions that were much like those that actually obtained, but differed in some slight respect in Nixon's decision-making apparatus—differed in such a way that by natural laws, the outgrowth of those modified initial conditions would have been nuclear holocaust. Thus, the rough idea is that one evaluates it $A \mathrel{\Box\!\!\rightarrow} C$ by considering a world in which A obtains that is as much like the actual world as possible both with regard to particular facts about the past as well as general facts about what might follow causally from what in the future. Lewis (1979a) attempts a general account of the kind of "relevant similarity" that fits our causal judgments, and derives from it an account of the regularities that give time its direction. As decision theorists, though, we need not be concerned whether such lofty philosophical ambitions can be fulfilled. We need only understand that where w_0 is the actual world, the value $f(A, w_0)$ of the Stalnaker selection function is the world as it would be if A obtained. It is that world, with all its ensuing history. In many situations, it will be clear that agents do have subjective probabilities for what that world would be like, and so the application of Stalnaker's apparatus to an agent's decision situation will be clear enough.[8]

Stalnaker's framework allows us to be more precise about how probabilities of subjunctive conditionals differ from ordinary conditionals probabilities. It is useful

[7] Fine (1975) gives this example to make roughly this point.

[8] Shin (1991a), for instance, devises a metric that seems suitable for simple games such as "Chicken".

to think of both $\rho(A \;\square\!\!\rightarrow C)$ and $\rho(C/A)$ as capturing a sense of *minimal* belief revision. $\rho(C/A)$, as we have seen, is the credence that an agent with prior ρ should assign C if she gets pure and certain news of A's truth. Thus, the function $\rho_A = \rho(\bullet /A)$ describes the outcome of a belief revision process in which an agent learns that A. This revision is minimal in the sense that it changes ρ so as to make A certain without thereby altering any ratios of the form $\rho(X\&A) : \rho(Y\&A)$. In terms of possible worlds, ρ_A is obtained from ρ by setting the credence of $\neg A$ equal to zero, and spreading the residual probability uniformly over the worlds in which A obtains—thus leaving undisturbed any evidential relationships that might obtain among propositions that entail A. The function $\rho^A = \rho(A \;\square\!\!\rightarrow \bullet)$ defines a rather different sort of minimal belief revision, *imaging* (Lewis 1976, 1981) or better, *imagining*. Instead of distributing the probability of $\neg A$ uniformly over the A-worlds, spread it with an eye to relevant similarities among worlds. Transfer the probability of each world w in which A is false to w^A, the A-world most similar to w, adding this probability to the probability w^A has already.

This whole treatment, though, rests on an assumption that is open to doubt: that there always exists a unique "most similar" world w^A. Perhaps there's no one definite way the world would be were I, say, now to flip this coin. The world might be indeterministic, or the supposition that I now flip the coin might be indeterminate—in that the exact force and manner in which I'd be flipping the coin isn't specified and isn't even under my control. It may then be the case neither that definitely were I to flip the coin it would land heads, nor that definitely were I to flip the coin it would not land heads. The law of *conditional excluded middle* will be violated:

$$(F \;\square\!\!\rightarrow H) \vee (F \;\square\!\!\rightarrow \neg H) \tag{23.10}$$

Conditional excluded middle obtains in Stalnaker's model. There are models and logics for conditionals in which it does not obtain.[9] In a case like this, however, the right weight to use for decision is clearly not the probability $\rho(F \;\square\!\!\rightarrow H)$ of a conditional proposition $F \;\square\!\!\rightarrow H$. If I'm convinced that neither $F \;\square\!\!\rightarrow H$ nor $F \;\square\!\!\rightarrow \neg H$ obtains, then my subjective probability for each is zero. The weight to use if I'm betting on the coin, though, should normally be one-half.

A number of ways have been proposed to handle cases like these. One is to think that with coins and the like, there's a kind of conditional chance that isn't merely subjective: the chance with which the coin would land heads *were* I to flip it. Write this $\pi_F(H)$. Then use as one's decision weight the following: one's *subjectively expected value for* this *objective conditional chance*. Suppose you are convinced that the coin is loaded, but don't know which way: You think that the coin is loaded either .6 toward heads or .6 toward tails, and have subjective probability of .5 for each of these possibilities:

[9] Lewis (1973) constructs a system in which worlds may tie for most similar, or it may be that for every A-world, there is an A-world that is more similar. He thus denies Conditional Excluded Middle: It fails, for instance, when two A-worlds tie for most similar to the actual world, one a C-world and the other a $\neg C$-world.

$$\rho\,(\pi_F(H) = .6) = .5 \qquad \text{and} \qquad \rho\,(\pi_F(H) = .4) = .5. \tag{23.11}$$

Your subjectively expected value for $\pi_F(H)$, then, will be the average of .6 and .4. Call this appropriate decision weight $\varepsilon_F(H)$. We can express this weighted averaging in measure theoretic terms, so that in general,

$$\varepsilon_A(C) = \int_0^1 x \cdot \rho\,(\pi_A(C) \in dx)\,. \tag{23.12}$$

$\varepsilon_A(C)$ thus measures one's subjective expectation of C's obtaining were A to occur: the sum of (i) the degree to which A's obtaining would tend to bring it about that C obtained, plus (ii) the degree to which C would tend to hold whether or not A obtained.

We can now write formulas for $\mathcal{U}$ using $\varepsilon_A(C)$ where we had previously used $\rho(A \,\Box\!\!\rightarrow C)$. Formula (23.4) above for instrumental expected utility now becomes

$$\mathcal{U}(A) = \sum_{i=1}^{n} \varepsilon_A(S_i)u(A, S_i) \tag{23.13}$$

Lewis (1981) gives an alternative formulation of causal decision theory, which is equivalent to the Stalnaker formulation we've been presenting whenever the Stalnaker framework applies as intended. He speaks of *dependency hypotheses:* complete hypotheses concerning what depends on what and how. Which dependency hypothesis obtains is causally independent of what the agent does, and so a Lewis dependency hypothesis can serve as a "state" in the Savage framework. He allows dependency hypotheses to contain an element of objective chance.[10]

Can an empiricist believe in such things as objective conditional chance or objective dependencies? This is a hotly debated topic, mostly beyond the scope of this article. Some philosophers think that objective dependency can be defined or characterized, somehow, in purely non-causal terms. Others doubt that any such direct characterization is possible, but think that a more indirect strategy may be available: Characterize a thinker's *beliefs* in causal dependencies—or his *degrees* of belief, his subjective probabilities, or as we have been calling them, his *credences*. His credences in causal propositions, this contention is, can be cashed out fully in terms of complex features of his credences in non-causal propositions—propositions that don't involve causal notions.

We ourselves would argue that causal propositions are genuine propositions of their own kind, basic to thinking about the world. They can't be fully explained in other terms, but they can be vindicated. We can be just as much empiricists about

[10] Skyrms (1980) offers another formulation, invoking a distinction between factors that are within the agent's control and factors that aren't. Lewis (1981) discusses both Skyrms and unpublished work of Jordan Howard Sobel, and Skyrms (Skyrms 1984), 105–6, compares his formulation with those of Lewis (1981) and Stalnaker (1981).

causes as we can about other features of the layout of the world. A rational thinker forms his credences in causal propositions in much the same Bayesian way he does for any other matter: He updates his subjective probabilities by conditionalizing on new experience. He starts with reasonable prior credences, and updates them. Subjective probability theorists like de Finetti long ago explained how, for non-causal propositions, updating produces convergence. The story depends on surprisingly weak conditions placed on the thinker's prior credence measure. The same kind of story, we suspect, could be told for credences in objective chance and objective dependence.

Lewis (1980) has told a story of this kind for credence in objective chance. His story rests on what he labels the "Principal Principle", a condition which characterizes reasonable credences in objective chances. Take a reasonable credence measure ρ, and a proposition about something that hasn't yet eventuated—say, that the coin I'm about to flip will land heads. Let me conditionalize his credences on the proposition that as of now, the objective probability of this coin's landing heads is .6. Then his resulting conditional credence in the coin's landing heads, the principle says, will likewise be .6. Many features of reasonable credence in objective chance follow from from this principle. From a condition on reasonable prior credences in objective chance follows an account of how one can learn about them from experience.

A like project for objective dependency hasn't been carried through, so far as we know, but the same broad approach would seem promising.[11] In the meantime, there is much lore as to how experience can lead us to causal conclusions—and even render any denial of a causal depencency wildly implausible. The dependence of cancer on smoking is a case in point. Correlation is not causation, we all know, and a conditional probability is not a degree of causal dependence. (In the notation introduced above, the point is that $\rho(C/A)$ need not be $\varepsilon_A(C)$, the subjectively expected value of the objective chance of C were one to do a.) Still, correlations, examined with sophistication, can *evidence* causation. A chief way of checking is to "screen off" likely common causes: A correlation between smoking and cancer might arise, say, because the social pressures that lead to smoking tend also to lead to drinking, and drinking tends to cause cancer. A statistician will check this possibility by separating out the correlation between smoking and cancer among drinkers on the one hand, and among non-drinkers on the other. More generally, the technique is this: A correlation between A and C, imagine, is suspected of being spurious—suspected *not* to arise from a causal influence of A on C or *vice versa*. Let F be a suspected common cause of A and C that might account for their correlation. Then see if the correlation disappears with F held constant. Econometricians elaborate such devices to uncover causal influences in an economy. The methodological literature on gleaning causal conclusions from experience includes classic articles by Herbert Simon (see Simon (1957), chs. 1–3).

[11] The work of Spirtes et al. (1993) and Pearl (2000) goes a long way toward realizing this goal.

Screening off is not a sure test of causality. A correlation might disappear with another factor held constant, not because neither factor depends causally on the other, but because the causal dependency is exactly counterbalanced by a contrary influence by a third factor.[12] Such a non-correlation might be robust, holding reliably with large sample sizes. But it will also be a coincidence: opposing tendencies may happen to cancel out, but we can expect such cases to be rare. Lack of correlation after screening off is *evidence* of lack of causal influence, but doesn't *constitute* lack of causal influence.

When controlled experiments can be done, in contrast, reasonable credences in a degree of objective dependency can be brought to converge without limit as sample size increases. Subjects are assigned to conditions in a way that we all agree has no influence on the outcome: by means of a chance device, say, or a table of pseudo-random numbers. Observed correlations then evidence causal dependence to whatever degree we can be confident that the correlation is no statistical fluke. With the *right* kind of partition, then, screening off does yield a reliable test of causality. But what makes a partition suitable for this purpose, we would claim, must be specified in terms that somehow invoke causality—in terms, for instance, of known causal independence.

How we can use evidence to support causal conclusions needs study. Standard statistical literature is strangely silent on questions of causation, however much the goals of statistical techniques may be to test and support causal findings. If we are right, then one class of treatments of causality will fail: namely, attempts to characterize causal beliefs in terms of the subjective probabilities and the like of non-causal propositions. Correct treatments must take causality as somehow basic. A constellation of relations—cause, chance, dependence, influence, laws of nature, what would happen if, what might likely happen if—are interrelated and may be intercharacterizable, but they resist being characterized purely from outside the constellation. Our hope should be that we can show how the right kind of evidence lets us proceed systematically from causal truisms to non-obvious causal conclusions. Fortunately, much of decision and game theory is already formulated in terms that are causal, implicitly at least, or that can be read or interpreted as causal. (When games are presented in normal form, for instance, it may be understood that no player's choice of strategies depends causally on the choice of any other.) A chief aim of causal decision theory is to make the role of causal beliefs in decision and game theory explicit.

Ratificationism

While some proponents of auspiciousness maximization have taken the heroic course and tried to argue that cooperating in Prisoner's Dilemma with Twin is rational,[13] most now concede that only defection makes sense. Nevertheless,

[12] Gibbard and Harper (1978), 140–2, construct an example of such a case.

[13] See, for example, Horgan (1981).

the promise of an analysis of rational choice free from causal or counterfactual entanglements remains appealing. A number of writers have therefore sought to modify the evidential theory so that it endorses the non-cooperative solution in PDT and yet does not appeal any unreconstructed causal judgements. The hope is that one will be able to find statistical techniques of the sort discussed toward the end of the last section to distinguish causal relationships from suprious correlations, and then employ these techniques to formulate a decision theory that can be sensitive to causal considerations without making explicit use of causal or subjunctive notions. The most influential of these attempts are found in Eells (1982) and Jeffrey (1983). Since the two approaches are similar, we focus on Jeffrey.

As we have seen, statisticians sometimes use "screening off" techniques to detect the effects of a common cause. Jeffrey's strategy is based on the insight that an agent's ability to anticipate her own decisions typically screens off any purely evidential import that her actions might possess. Prior to performing an act she will generally come to realize that she has decided on it. The act itself then ceases to be a piece of evidence for her since she has already discounted it. Letting $\triangle^A$ denote the decision to do *a*, we can put Jeffrey's claim like this:

- *Screening*. The decision to perform *A* screens off any purely evidential correlations between acts and states of the world, in the sense that $\rho(S/B\&\triangle^A) = \rho(S/\triangle^A)$ for all acts *B* and states *S*.

To ensure that these conditional credences are well-defined, Jeffrey must assume that the agent always assigns some positive probability to the prospect that she will fail to carry out a decision—due to a "trembling hand," a lack of nerve, or other factors beyond her control—so that $\rho(B\&\triangle^A)$ is non-zero for all acts *B*.

To see what happens when screening is introduced into PDT, imagine that during the course of your deliberations but prior to performing any act, you become certain that you will decide to cooperate, so that your credence in Δ_y^C moves to one. Since you are likely to carry out whatever decision you make, your probability for C_y also moves close to one, which gives you strong grounds for thinking your twin will cooperate. Indeed, if Screening obtains, you will have strong evidence for thinking that Twin is about to cooperate, no matter what news you get as to what you yourself are about do, because $\rho\left(C_t/C_y\&\Delta_y^C\right) = \rho\left(C_t/D_y\&\Delta_y^C\right) = \rho\left(C_t/\Delta_y^C\right) \approx 1$. Condition (23.2) is thus satisfied. You can then correctly apply the evidential dominance principle to conclude that defection is your most auspicious option. In this way, Screening ensures that if you are certain that you will eventually decide to cooperate, then you will assign defecting a higher auspiciousness than cooperating. On the other hand, if you are certain that you will decide to defect, you then have strong reason to suspect that twin is about to defect, whatever you learn that you yourself are about to do—and again, defecting turns out to be the more auspicious option. Therefore, no matter how auspicious cooperating might seem *before* you make up your mind about what to do, defecting is sure to look more auspicious *afterwards*—and this will be true no matter what decision you have made. Jeffrey proposes to uses this basic asymmetry—between what one decides and what one does—to argue that defection in PDT is the only rational course of action for an auspiciousness maximizer.

The case has already been made for an agent who is already certain about what she will decide. But what about agents who have yet to made up their minds? Here things get dicey. If you have not yet decided what to do, then the probabilities you assign to Δ_y^C and Δ_y^D will be far from one. This puts the auspiciousness values of C_y and D_y near those of $(C_y\&\Delta_y^C)$ and $(D_y\&\Delta_y^D)$ respectively, and since $\mathcal{V}(C\&\Delta_y^C) > \mathcal{V}(D\&\Delta_y^D)$, evidential decision theory tells you to choose cooperation. However, as soon as you make this choice, you will assign Δ_y^D a credence close to one, and as we have seen, you will then favor defection. Thus, the pursuit of good news forces you to make choices that you are certain to rue from the moment you make them—clearly something to avoid.

Jeffrey hopes to circumvent this difficulty by denying that evidential decision theory requires one to maximize auspiciousness as one *currently* estimates it. If you are savvy, he argues, you will realize that any choice you make will change some of your beliefs, thereby altering your estimates of auspiciousness. Thus, given that you want to make decisions that leave you better off for having made them, you should aim to maximize auspiciousness not as you currently estimate it, but as you *will* estimate it once your decision is made. You ought to, "choose for the person you expect to be when you have chosen,"[14] by maximizing expected utility computed relative to the personal probabilities you will have *after* having come to a firm decision about what to do. This is only possible if your choices conform to the maxim

- *Evidential Ratifiability*. An agent cannot rationally choose to perform A unless A is *ratifiable*, in the sense that $\mathcal{V}(A\&\Delta^A) \geq \mathcal{V}(B\&\Delta^A)$ for every act B under consideration.

This principle advises you to ignore your current views about the evidentiary merits of cooperating versus defecting, and to focus on maximizing future auspiciousness by making choices that you will regard as propitious from the epistemic perspective you will have once you have made them. Since in the presence of Screening, defection is the only such choice, the maxim of Evidential Ratifiability seems to provide an appropriately "evidentialist" rationale for defecting in PDT.

Unfortunately, though, the Screening condition need not always obtain. There are versions of PDT in which the actual performance of an act provides better evidence for some desired state than does the mere decision to perform it. This would happen, for example, if you and twin are bumblers who tend to have a similar problems carrying out your decisions. The fact that you were able to carry out a decision would then be evidentially correlated with you twin's act, and this correlation would not be screened off by the decision itself. In such cases the evidential ratifiability principle sanctions cooperation.[15] Therefore, the Jeffrey/Eells

[14] Jeffrey (1983) p. 16.

[15] Jeffrey, in his original published treatment of ratificationism (Jeffrey 1983), 20, gives this counterexample and credits it to Bas van Fraassen. Shin (1991b) treats cases in which the respective players' "trembles" are independent of each other.

strategy does not always provide a satisfactory evidentialist rationale for defecting in PDT. We regard this failure as reinforcing our contention that any adequate account of rational choice must recognize that decision makers have beliefs about causal or counterfactual relationships, beliefs that cannot be cashed out in terms of ordinary subjective conditional probabilities—in terms of conditional credences in non-causal propositions.

Ratificationism and Causality in the Theory of Games

Despite its failure, the Jeffrey/Eells strategy leaves the theory of rational choice an important legacy in the form of the Maxim of Ratifiability. We will see in this section that it is possible to assimilate Jeffrey's basic insight into casual decision theory and that so understood, it codifies a type of reasoning commonly employed in game theory. Indeed, the idea that rational players always play their part in a Nash equilibrium is a special case of ratificationism.

The notion of a ratifiable act makes sense within any decision theory with the resources for defining the expected utility of one act given the news that another will be chosen or performed. In causal decision theory the definition would be this:

$$\mathcal{U}(B/A) = \sum_{i=1}^{n} \rho((B \,\Box\!\!\rightarrow S_i)/A)u(B, S_i)$$

Jeffrey's insight then naturally turns into the maxim,

- *Causal Ratifiability*. An agent cannot rationally choose to perform act A unless $\mathcal{U}(A/A) \geq \mathcal{U}(B/A)$ for every act B under consideration.

This says that a person should never choose A unless once she does, her expection of A's efficacity in bringing about desirable results is at least as great as that for any alternative to A. Notice that one no longer needs a "trembling hand" requirement to define the utility of one act conditional on another, since the conditional credence $\rho(B \,\Box\!\!\rightarrow S/A)$ is well-defined even when A and B are incompatible. This means that an agent who is certain that she will perform A can still coherently wonder about how things *would* have gone *had* she done B—even though she can no longer wonder about how things are set to go if she's *going* to do B. This ability to assign utilities to actions that definitely will not be performed can be used to used to great advantage in game theory, for instance, when considering subgames and subgame perfection.

To see how an act might fail to be causally ratifiable, imagine yourself playing Matching Pennies with an opponent who you think can predict your move and will make a best response to it (see table). Neither pure act then turns out to be ratifiable. Suppose that you are strongly inclined to play [HEADS], so that $\rho(H_y)$ is close to one. Since your conditional probability for Twin playing heads given that you do is high, it also follows that your subjective probability for H_t will be high. Thus by recognizing that you plan to play [HEADS], you give yourself evidence for thinking that Twin will also play [HEADS]. Note, however, this does nothing to alter the fact that you still judge Twin's actions to be causally independent of your own; your subjective probabilities still obey

	HEADS	TAILS
HEADS	−1, 1	1, −1
TAILS	1, −1	−1, 1

Matching Pennies

$$\rho\left(\left(H_y \Box\!\!\rightarrow H_t\right)/H_y\right) \approx \rho\left(\left(T_u \Box\!\!\rightarrow H_t\right)/H_y\right).$$

Since $\rho(Ht) \approx 1$, your overall position is this: because you are fairly sure that you will play heads, you are fairly sure that Twin *will* play heads too; but you remain convinced that she *would* play heads even if (contrary to what you expect) you were to play tails. Under these conditions, the conditional expected utility associated with [TAILS] larger than that associated with [HEADS] on the supposition that [HEADS] is played. That is to say, [HEADS] is unratifiable. Similar reasoning shows that tails is also unratifiable. In fact, the only ratifiable act this game is the mixture ½ [HEADS] + ½ [TAILS].[16]

It is no coincidence that this mixture also turns out to be the game's unique Nash equilibrium; there is a deep connection between ratifiable acts and game-theoretic equilibria. Take any two-person game, such as Matching Pennies, for which the unique Nash equilibrium consists of mixed strategies (which the players in fact adopt). If a player has predicted the other's strategy correctly and with certainty, then by playing the strategy she does, she maximizes her expected utility. But this isn't the only strategy that, given her credences, would maximize her expected utility; any other probability mixture of the same pure strategies would do so too. The

[16] Piccione and Rubinstein (1997) present another kind of case in which considerations of ratifiability may be invoked: the case of the "absent-minded driver" who can never remember which of two intersections he is at. One solution concept they consider (but reject) is that of being "modified multi-selves consistent". In our terms, this amounts to treating oneself on other occasions as a twin, selecting a strategy that is ratifiable on the following assumption: that one's present strategy is fully predictive of one's strategy in any other situation that is subjectively just like it. This turns out to coincide with the "optimal" strategy, the strategy one would adopt if one could choose in advance how to handle all such situations.

strategy she adopts is unique, though, in this way: it is the only strategy that could be ratifiable, given the assumption that her opponent has predicted her strategy and is playing a best response to it. It should be clear that this argument extends straightforwardly to the *n*-person case. In any Nash equilibrium, all players perform causally ratifiable acts.

	C_1	C_2
R_1	1, 1	0, 4
R_2	4, 0	−15, −15

Chicken

In fact, the least restrictive of all equilibrium solution concepts—Aumann's correlated equilibrium—can be understood as a straightforward application of the maxim of ratifiability.[17] Aumann sought to generalize the Nash equilibrium concept by relaxing the assumption—implicit in most previous game-theoretic thinking—that players in normal-form games believe that everyone acts independently of everyone else. Take a game of Chicken. The table shown might give the utilities of drivers traveling along different roads into a blind intersection, who must decide whether to stop and lose time (R_1 and C_1) or to drive straight through and risk an accident (R_2 and C_2). Let the players assume that they will choose, respectively, strategies $p \cdot R_1 + (1-p) \cdot R_2$ and $q \cdot C_1 + (1-q) \cdot C_2$. What credences will they give to the various joint pure actions they may end up performing?

Game theorists have traditionally assumed that the players will treat the chance devices that implement their mixed strategies as evidentially independent, ascribing the credences in Table 23.1. Aumann imagines that they might ascribe the credences of Table 23.2: To see how the correlations in Table 23.2 might arise, imagine a "managed" game of Chicken in which, to minimize the chances of a disastrous outcome, things have been arranged so that an impartial arbitrator will throw a fair die and illuminate a traffic light at the intersection according to the "Arbitrator's Scheme" shown in the table. If each player intends to heed the signal, stopping on red and driving through on green, and believes his opponent intends to do so as well, then both will have the correlated priors of Table 23.2. Aumann showed how to define an equilibrium solution concept even when players' acts are correlated in this way. An *adapted* strategy is a rule *s* which dictates a unique pure act *s*(*A*) for

[17] More precisely, correlated equilibrium is the weakest equilibrium solution concept which assumes that all players have common beliefs. When this assumption is relaxed one obtains a subjectively correlated equilibrium. For details see Aumann (1974, 1987).

Table 23.1 Chicken with Independence

	C_1	C_2
R_1	pq	$p(1-q)$
R_2	$(1-p)q$	$(1-p)\ (1-q)$

Table 23.2 Chicken with Correlation

	C_1	C_2
R_1	$p+q-1$	$1-q$
R_2	$1-p$	0

each act A signalled by the arbitrator: for instance, "Stop on red and go on green" or "Run all lights." A correlated equilibrium is simply a pair of adapted strategies r and c such that, for all alternatives r^* and c^*, the following condition holds (call it *CE*):

	one	two	three	four	five	six
ROW	Red	Red	Red	Red	Green	Green
COL	Red	Red	Green	Green	Red	Red

Arbitrator's Scheme

$$\sum_{ij} \rho(R_i \& C_j) \cdot \mathcal{U}_{\text{row}}(r(R_i), C_j) \geq \sum_{ij} \rho(R_i \& C_j) \cdot \mathcal{U}_{\text{row}}(r^*(R_i), C_j)$$

$$\sum_{ij} \rho(R_i \& C_j) \cdot \mathcal{U}_{\text{col}}((R_i), c(C_j)) \geq \sum_{ij} \rho(R_i \& C_j) \cdot \mathcal{U}_{\text{col}}(R_i c^*(C_j))$$

Table 23.3 Equilibrium for Chicken with Independence

	C_1	C_2
R_1	⅓	⅓
R_2	⅓	0

Thus, *r* and *c* constitute a correlated equilibrium iff no player has reason to deviate from her adapted strategy given her beliefs and the signal she receives. The reader may verify that, when things are as described in Table 23.3, both players' resolving to obey the light is a correlated equilibrium, whereas there is no correlated equilibrium in which either player decides to run red lights.

Definition CE makes it appear as if the existence of correlated equilibria depends on the availability of external signaling mechanisms. This was the view presented by Aumann in (1974), but in (1987), he shows that acts themselves can serve as appropriate signals. Specifically, Aumann established that a common probability distribution[18] ρ defined over strategy combinations comprises a correlated equilibrium iff given any pair of acts *R* and *C* assigned positive probabilities, one has *CE*:*

$$\sum_i \rho(C_j/R)\cdot \mathcal{U}_{\text{row}}(R, C_j) \geq \sum_i \rho(C_j/R)\cdot \mathcal{U}_{\text{row}}(R^*, C_j)$$

$$\sum_i \rho(R_j/C)\cdot \mathcal{U}_{\text{col}}(R_i, C) \geq \sum_i \rho(R_j/C)\cdot \mathcal{U}_{\text{col}}(R_i, C^*)$$

for all alternatives *R** and *C**. CE* requires, then, that agents assign zero prior probability to any act, either of their own or of their adversaries', that does not maximize expected utility on the condition that it will be performed. Aumann regards this condition as "an expression of Bayesian rationality," since players satisfy it by maximizing expected utility at the time they act.

As a number of authors have noted, CE* is an application of the maxim of ratifiabilty.[19] It requires players to give zero credence to non-ratifiable acts. Hence for a group of players to end up in a correlated equilibrium, it is necessary and sufficient that all choose ratifiabile acts and expect others to do so as well. Aumann's "Bayesian rationality" thus coincides with the notion of rationality found in Jeffrey's ratificationism. More specifically, we contend that Aumann is requiring rational players to choose *causally* ratifiable actions.

[18] Note that such a distribution determines a unique mixed act for each player. Thus, it makes no difference whether one talks about the players' acts or the players' beliefs being in equilibrium.

[19] Shin (1991b), Skyrms (1990).

While Aumann has rather little to say about the matter, he clearly does not mean the statistical correlations among acts in correlated equilibrium to reflect causal connections. This is particularly obvious when external signaling devices are involved, for in such cases each player believes that his opponents would heed the arbitor's signal whether or not he himself were to heed it. Each player uses his knowledge of the arbitor's signal to him to make inferences about the signals given to others, to form beliefs about what his opponents expect him to do, and ultimately to justify his own policy of following the arbitor's signal. What he does not do is suppose that the correlations so discovered *would* continue to hold no matter what he decided to do. For example, if [ROW]'s credences are given in Table 23.3, it would be a mistake for him to run a red light in hopes of making it certain that [COLUMN] will stop; [COLUMN's action, after all, is determined by the signal she receives, not by what [ROW] does. Notice, however, that a straightforward application of evidential decision theory recommends running red lights in this circumstance—further proof that the view is untenable. In cases, then, where correlations are generated via external signaling, a causal interpretation of CE* clearly is called for. To make this explicit, we can rewrite CE* as the following condition *CE***:

$$\sum_j \rho((R \,\Box\!\!\rightarrow C_j)/R) \cdot \mathcal{U}_{\mathrm{row}}(R, C_j) \geq \sum_j \rho((R^* \,\Box\!\!\rightarrow C_j)/R) \cdot \mathcal{U}_{\mathrm{row}}(R^*, C_j)$$

$$\sum_j \rho((C \,\Box\!\!\rightarrow R_i)/C) \cdot \mathcal{U}_{\mathrm{col}}(R_i, C) \geq \sum_i \rho((C^* \,\Box\!\!\rightarrow R_i)/C) \cdot \mathcal{U}_{\mathrm{col}}(R_i, C^*)$$

This reduces to CE* on the assumption that [ROW] and [COLUMN] cannot influence each other's acts, so that both $\rho\left(\left(R* \;\; \Box\!\!\rightarrow C_j\right)/R\right) = \rho\left(C_j/R\right)$ and $\rho\left(\left(C* \;\; \Box\!\!\rightarrow R_i\right)/C\right) = \rho\left(R_i/C\right)$ hold for all R^* and C^*.

The situation is not appreciably different in cases where the correlation arises without any signaling device. Imagine playing [ROW] in the coordination game in Table 23.4, and consider the correlated equilibrium described by Table 23.5: These correlations need not have arisen through signaling. You might find the (R_1, C_1) equilibrium salient because it offers the vastly higher payoff, and, knowing that

Table 23.4 Correlation without Signaling

	C_1	C_2
R_1	25, 1	0, 0
R_2	0, 0	1, 2

Table 23.5 Equilibrium for Correlation without Signaling

	C_1	C_2
R_1	0.7	0.01
R_2	0.01	0.28

your opponent can appreciate its salience for you, you might conclude that she expects you to play it. That makes C_1 the better play for her, which reinforces your intention to play R_1 and your belief that she will play C_1, and so on. It would not be unreasonable under these conditions for the two of you to end up in the correlated equilibrium of Table 23.5. Still, there is no suggestion here that your initial inclination to play R_1 is somehow responsible causally for your opponent's credences. Her credences are what they are because she suspects that you are inclined to play R_1, but neither your decision to play R_1 nor your actually playing of R_1 is the cause of these suspicions – she develops them solely on the basis of her knowledge of the game's structure. As in the signaling case, the acts in a correlated equilibrium are evidentially correlated but causally independent.

This explicitly causal reading of Aumann's discussion helps to clear up a perplexing feature of correlated equilibria. CE only makes sense as a rationality constraint if agents are able to treat their own actions as bits of information about the world, for it is only then that the expressions appearing to the right of the "$\geq$" signs can be understood as giving utilities for the starred acts. As Aumann notes, his model (like that of Jeffrey before him)

> does away with the dichotomy usually perceived between uncertainty about acts of nature and of personal players... . In traditional Bayesian decision theory, each decision maker is permitted to make whatever decision he wishes, after getting whatever information he gets. In our model this appears not to be the case, since the decision taken by each decision maker is part of the description of the state of the world. This sounds like a restriction on the decision maker's freedom of action. (Aumann 1987), 8.

The problem here is that the utility comparisons in CE seem to portray acts as things that happen to agents rather than things they do. Moreover, it is not clear why an agent who learns that he is surely going to perform R should need to compare it with other acts that he is sure he will not perform. Aumann tries to smooth things over by suggesting that CE describes the perspective of agents, not as choosers, but as "outside observers." He writes,

> The "outside observer" perspective is common to all differential information models in economics... . In such models, each player gets some information or "signal"; he hears only the signal sent to him, not that of others. In analyzing his situation, [the] player must

> first look at the whole picture as if he were an outside observer; he cannot ignore the possibility of his having gotten a signal other than he actually got, even though he knows that he actually did not get such a signal. This is because the other players do not know what signal he got. [He] must take the ignorance of the other players into account when deciding on his own course of action, and he cannot do this if he does not explicitly include in the model signals other than the one he actually got. (Aumann 1987), 8.

The problem with this response is that it does not tell us how a player is supposed to use the knowledge gained from looking at things "externally" (i.e., as if he were not choosing) to help him with his "internal" decision (where he must choose how to act). The point is significant, since what's missing from the outside observer standpoint is the very thing that makes something a decision problem—the fact that a choice has to be made.

In our view, talk about outside observers here is misleading. Aumann's third-person, "outside observer" perspective is the first-person *subjunctive* perspective: a view on what *would* happen if *I were to do* something different. It is an advantage of causal decision theory that it allows one to assess the rationality of acts that certainly will not be performed in the same way as one assesses the rationality of any other option. One simply appeals to whatever facts about objective chances, laws of nature, or causal relations are required to imagine the "nearest" possible world in which the act is performed, the one most similar, in relevant respects, to the actual world in which one won't perform the act. One then sees what would ensue under those circumstances. Even, for instance, if one assigns zero credence to the proposition that one will intentionally leap off the bridge one is crossing, it still makes sense on the causal theory to speak of the utility of jumping. Indeed the abysmally low utility of this action is the principal reason why one is certain not perform it. When we interpret correlated equilibria causally, then, along the lines of CE**, there is no need for an "external" perspective in decision making. All decisions are made from the first-person subjunctive perspective of "What would happen if I performed that act?"

A number of other aspects of game-theoretic reasoning can likewise be analyzed in terms of causal ratifiability.[20] Some of the most interesting work in this area is due to Harper, who has investigated the ways in which causal decision theory and ratifiability can be used to understand extensive-form games (Harper 1986, 1991). This is a natural place for issues about causation to arise, since players' choices at early stages in extensive-form games can affect other players' beliefs—and thus their acts—at later stages.

Harper proposes using the maxim of ratifiability as the first step in an analysis of game-theoretic reasoning that is "eductive" or "procedural", an analysis that seeks to supply a list of rules and deliberative procedures that agents in states of indecision can use in order to arrive at an intuitively correct equilibrium choice.[21] Rational

[20] See Shin (1991b), for instance, for an interesting ratificationist gloss on Selten's notion of a "perfect" equilibrium.

[21] On the need for such "eductive" procedures see Binmore (1987, 1988).

players, he suggests, should choose actions that maximize their unconditional causal expected utility *from among the ratifiable alternatives*. (Harper regards cases where no ratifiable strategies exist as genuinely pathological.) The idea is not simply to choose acts that maximize unconditional expected utility, since these need not be ratifiable. Nor is it to choose acts with maximal expected utility on the condition that they are performed, since these may have low unconditional utility. Rather, one first eliminates unratifiable options, and then maximizes unconditional expected utility with what is left. In carrying out the second step of this process, each player imagines her adversaries choosing among their ratifiable options by assigning zero probability to any option that wouldn't be ratifiable.

Harper shows that both in normal and in extensive-form games, players who follow these prescriptions end up choosing the intuitively "right" act in a wide range of cases. In extensive-form games, his method produces choices that are in sequential equilibrium—and perhaps most interesting, the method seems to promote a strong form of "forward induction." To illustrate the latter point, and to get a sense of how Harper's procedure works in practice, consider the game of Harsanyi and Selten (1988) shown here. [ROW]'s act *A* yields fixed payoffs, whereas [ROW]'s act *B* leads to a strategic subgame with [COLUMN]. Harsanyi and Selten argue that (*C, e*) is the unique rational solution to the subgame, and they use backwards induction to argue that (*AC, e*) is the only rational solution in the full game. Their thought is that at the initial choice point [ROW] will know that (*C, e*) would be played if the subgame were reached, so he actually faces the "truncated game" as shown below, which makes *A* the only rational choice.

Kohlberg and Mertens (1986) have objected to this reasoning on the grounds that [ROW] can perform *B* as a way of signaling [COLUMN] that he has chosen *BD* (since it would be crazy to pass up *A* for *C*), and can thus force [COLUMN] to play *f* rather than *e*. Harsanyi and Selten respond by claiming that [COLUMN] would have to regard [ROW]'s playing *B* as a mistake since, "before deciding whether [ROW] can effectively signal his strategic intentions, we must first decide what strategies are rational for the two players in the subgame, and accordingly what strategy is the rational strategy for [ROW] in the truncation game" (Harsanyi and Selten 1988), 353 (see table). Thus, we have a dispute over what sorts of effects the playing of *B* would have on [COLUMN]'s beliefs at the second choice point, and thereby on her decision. This is just the sort of case where causal decision theory can be helpful.

Harper's procedure endorses Kohlberg's and Mertens' contention. Harsanyi and Selten's preferred act for [ROW] will be unratifiable, so long as each player knows that the other chooses only ratifiable options. For *A* to be ratifiable, it would at least have to be the case that

$$U(A/A) = 4 \geq \mathcal{U}(BD/A) = \rho((BD \,\square\!\!\rightarrow e)/A) \cdot 0 + \rho((BD \,\square\!\!\rightarrow f)/A) \cdot 10. \tag{23.14}$$

However, since [COLUMN] must choose at the third choice point knowing that [ROW] has played *B*, but without knowing whether he has played *C* or *D*, it follows that [ROW]'s credence for $BC \,\square\!\!\rightarrow f$ must be $\rho(f/B)$. Now at the final choice point, [COLUMN] would know that [ROW] had chosen either *BC* or *BD* or some mixture of the two, and she would have to assume that the option chosen was ratifiable (if such an option is available). *BC* clearly cannot be ratifiable, since it is dominated by *A*. Harper also shows, using a somewhat involved argument, that no mixture of *BC* and *BD* can be ratifiable.[22] *BD*, however, is ratifiable, provided that $\rho\,(BD\ \Box\!\!\rightarrow f/BD) = \rho\,(f/B) \geq {}^{4}\!/_{10}$. Thus, since only one of [ROW]'s B-acts can be ratifiable, [COLUMN] would have to assign it a probability of one if she were to find herself at the second choice point. [ROW], knowing all this and knowing that f is [COLUMN]'s only ratifiable response to *BD*, will indeed assign a high value to $\rho(f/B)$, viz. $\rho(f/B) = 1$. This in turn ensures that $\mathcal{U}(BD/A) > \mathcal{U}(A/A)$, making *A* unratifiable. *BD* is thus the only ratifiable solution to the Harsanyi/Selten game. Hence, if Harper is correct, it seems that Kohlberg and Mertens were right to reject the backwards induction argument and to think that [ROW] can use *B* to warn [COLUMN] of his intention to play *D*.

We hope this example gives the reader something of the flavor of Harper's approach. Clearly his proposal needs to be elaborated more fully before we will be able to make an informed judgment on its merits. We are confident, however, that any adequate understanding of game-theoretic reasoning will rely heavily on causal decision theory and the maxim of ratifiability.

Foundational Questions

Before any theory of expected utility can be taken seriously, it must be supplemented with a representation theorem that shows precisely how its requirements are reflected in rational preference. To prove such a theorem one isolates a small set of axiomatic constraints on preference, argues that they are requirements of rationality, and shows that anyone who satisfies them will automatically act in accordance with the theory's principle of expected utility maximization. The best known result of this type is found in Savage (1972), where it was shown that any agent whose preferences conform to the well-known Savage axioms will maximize expected utility, as defined by Eq. (23.1), relative to a (unique) probability ρ and a (unique up to positive linear transformation) utility u. Unfortunately, this result does not provide a fully satisfactory foundation for either CDT or ETD, because, as we have seen, Savage's notion of expected utility is ambiguous between a causal and an evidential interpretation. This leaves some important unfinished business for evidential and causal decision theorists, since each camp has an obligation to present a representation theorem that unambiguously captures its version of utility theory.

Evidential decision theorists were first to respond to the challenge. The key mathematical result was proved in Bolker (1966), and applied to decision theory in Jeffrey (1983). The Jeffrey/Bolker approach differs from Savage's in two significant

[22] Harper (1991), 293.

ways: First, preferences are defined over a σ-algebra of propositions that describe not only states of the world, but consequences and actions as well. Second, Savage's "Sure-thing" Principle (his postulate P3) is replaced by the weaker

- *Impartiality Axiom*: Let X, Y and Z be non-null propositions such that (a) Z is incompatible with both X and Y, (b) X and Y are indifferent in the agent's preference ordering, and (c) $(X \vee Z)$ is indifferent with $(Y \vee Z)$ but not with Z. Then, $(X \vee Z^*)$ must be indifferent with $(Y \vee Z^*)$ for any proposition Z^* incompatible with both X and Y.

In the presence of the other Bolker/Jeffrey axioms, which do not differ substantially from those used by Savage, Impartiality guarantees that the agent's preferences can be represented by a function that satisfies Eq. (23.3).[23] It also ensures that the representation will be *partition independent* in the sense that, for any partitions $\{X_i\}$ and $\{Y_j\}$ and any act A, one will always have $\mathcal{V}(A) = \sum \rho(X_i/A)\mathcal{V}(A\&X_i) = \sum \rho(Y_j/A)V(A\&Y_j)$. Thus, in EDT it does not matter how one chooses the state partition relative to which expected utilities are computed.

This contrasts sharply with Savage's theory. Formula (23.1) shows how to compute expected utilities with respect to a single partition of states, but it gives no guarantee that different choices of partitions yield the same value for $\mathcal{V}(A)$. As a consequence, Savage had to place restrictions on the state partitions that could be legitimately employed in well-posed decision problems. Strictly speaking, he said, his axioms only apply to *grand-world* decisions whose act and state partitions slice things finely enough to ensure that each act/state pair produces a consequence that is sufficiently detailed to decide every question the agent cares about. Thus, on the official view, we can only be confident that (23.1) will yield the correct value for $\mathcal{V}(A)$ when applied to state partitions in "grand-world" decisions. Savage recognized this as a significant restriction on his theory, since owing to the extreme complexity of grand-world decisions, no actual human being can ever contemplate making one. He tried to make this restriction palatable by suggesting that his axioms might be usefully applied to certain "small-world" decisions, and expressing the hope that there would be only a "remote possibility" of obtaining values for $\mathcal{V}(A)$ inconsistent with those obtained in the grand-world case. This hope, however, was never backed-up by any rigorous proof. We take it as a mark in favor of EDT that it is can solve this "problem of the small-world" by giving such a proof.

The easiest way to prove a representation result for CDT is to co-opt Savage's theorem by stipulating that well-posed decision problems must be based on partitions of states that are certain to be causally independent of acts, and then imposing Savage's axioms on such problems. A number of causal decision

[23] This representation will not be unique (except in the rare case where $\mathcal{V}$ is unbounded), for, as a simple calculation shows, if the function $\mathcal{V}(A) = \sum \rho(S_i/A)u(A, S_i)$ represents a preference ordering, and if k is such that $1 + k\mathcal{V}(X) > 0$ for all propositions X in the algebra over which $\mathcal{V}_k(A) = \sum \rho k(S_i/A)u_k(A, S_i)$ is defined, then $V_k(A) = \sum \rho k\,(S_i/A)\,u_k\,(A, S_i)$ will also represent the ordering, when $pk(X) = \rho(X)(1 + k\mathcal{V}(X))$ and $\mathcal{V}_k(X) = [\mathcal{V}(X)(1 + k)]/(1 + k\mathcal{V}(X))$.

theorists have endorsed the view that there is a “right” partition of states to be used for computing instrumental expected utility.[24] The Lewis formulation of CDT in terms of dependency hypotheses mentioned in the section “Conditionals and their probabilities” above is an example of this strategy (Lewis 1981). Minor intramural squabbles aside, the idea is that each element in the privileged partition, often denoted $\mathbf{K} = \{K_j\}$ following Skyrms (1980), should provide a maximally specific description of one of the ways in which things that the agent cares about might depend on what she does.[25] It is characteristic of such partitions that they will be related to the agent’s subjective probability by[26]

- *Definiteness of Outcome*: For any proposition O that the agent cares about (in the sense of not being indifferent between O and $\neg O$), any action A, and any $K_j \in \mathbf{K}$, either $\rho((A \mathbin{\Box\!\!\rightarrow} O)/K_j) = 1$ or $\rho((A \mathbin{\Box\!\!\rightarrow} \neg O)/K_j) = 1$.
- *Instrumental Act Independence*: $\rho([A \mathbin{\Box\!\!\rightarrow} K_j] \leftrightarrow K_j) = 1$ for all acts A and states K_j.

The first of these ensures that $u\ (A,\ S_i)$ has the same value for each Savage-state S_i in K_j, and thus that

$$\mathcal{U}(A) = \sum_i \rho(A \mathbin{\Box\!\!\rightarrow} S_i) u(A, S_i) = \sum_j \rho(A \mathbin{\Box\!\!\rightarrow} K_j) \mathcal{U}(A, K_j)$$

The second condition then guarantees that $\mathcal{U}(A) = \sum_j \rho(K_j) \mathcal{U}(A, K_j)$. Since this equation has the form (23.5), it follows that if there exists a partition $\mathbf{K}$ that meets these two requirements, then one can appropriately apply the Savage axioms to actions whose outcomes are specified in terms of it. Fortunately, the required partition is certain to exist, because it is always possible to find a $\mathbf{K} = \{K_i\}$ such that (i) K_j entails either $(A \mathbin{\Box\!\!\rightarrow} O)$ or $(A \mathbin{\Box\!\!\rightarrow} \neg O)$ for every O the agent cares about, and (ii) $([A \mathbin{\Box\!\!\rightarrow} K_j] \leftrightarrow K_j)$ is a truth of logic for all A and K_j.[27] The general existence of such partitions lets us use Savage’s representation theorem as a foundation for CDT, subject to the proviso that the Savage axioms should only be applied to decisions framed in terms of a **K**-partition.

The trouble with this strategy is that the partition dependence of Savage’s theory is carried over into CDT. The need for a partition-independent formulation of CDT

[24] See, for example, Skyrms (1980), Lewis (1981), Armendt (1986).

[25] Notice that states are being viewed here as functions from acts to outcomes, whereas acts are taken as unanalyzed objects of choice (that is, as propositions the agent can make true or false as she pleases). This contrasts with Savage’s well-known formalization in which acts are protrayed as functions from states to outcomes, and states are left as unanalyzed objects of belief. Less hangs on this distinction than one might think. When one adopts the perspective of Jeffrey (1967, 1983) and interprets both states and actions as propositions, and views outcomes as conjunctions of these propositions, the two analyses become interchangeable.

[26] Here we are following Gibbard (1986).

[27] An explicit construction of **K** can be found in Gibbard and Harper (1978).

has been argued by Sobel (1989), and Eells (1982) has suggested that EDT should be preferred to CDT on this basis alone. The main difficulty is the problem of "small-worlds", which threatens to make the theory inapplicable, in the strictest sense, to all the decision problems people actually consider. Gibbard (1986) goes a certain distance toward alleviating this difficulty by, in effect, showing how to find the smallest **K**-partition for a given decision problem, but his partition is still rather "grand", and a fully partition-invariant version of CDT would still be desirable.

Armendt (1986) proves a representation result that does provide a formulation of CDT that is partition-independent, even though it does not do away with the notion of a **K**-partition. He takes the conditional decision theory of Fishburn (1974) as his starting point. The basic concept here is that of the utility of a prospect X on the hypothesis that some condition C obtains. If $\{C_1, C_2, \ldots, C_n\}$ is a partition of C, these conditional utilities are governed by the (partition independent) equation:

$$\mathcal{U}(X/C) = \sum \rho(C_i/C)\mathcal{U}(X/C_i) \tag{23.15}$$

which shows how X's utility given C depends on its utilities given the various ways in which C might be true. Notice that (23.15) allows for a distinction between an act A's unconditional utility and its utility conditional on its own performance. These are given respectively by

$$\mathcal{U}(A) = \mathcal{U}(A/A \vee \neg A) = \sum \rho(S_i)\mathcal{U}(A/S_i)$$

$$\mathcal{U}(A/A) = \sum \rho(S_i/A)\mathcal{U}(A/A \mathbin{\&} S_i)$$

where the state partition may be chosen arbitrarily. In a suggestion that bears some similarities to Jeffrey's ratificationist proposal, Armendt argues that decision problems in which an agent's unconditional preference for A differs from her preference for A conditional on itself are just the kinds of cases in which A's auspiciousness diverges from its instrumental expected utility. This suggests a way of characterizing **K**-partitions directly in terms of the agent's preferences. Armendt's thought is that the elements of an appropriate **K**-partition should "screen-off" differences in value between unconditional A and A-conditional-on-A, so that the agent is indifferent between A given K_i and A given $(A\&K_i)$ for every i. When this is so, we will have $\sum \rho(K_i)\mathcal{U}(A/K_i) = \sum \rho(K_i/A)\mathcal{U}(A/K_i)$, and Armendt shows that the conditional utilities $\mathcal{U}(A/K_i)$ can be eliminated in favor of the unconditional news values $\mathcal{V}(A\&K_i)$ as long as there exists at least one partition of "consequences" $\mathbf{O} = \{O_j\}$ such that the agent's unconditional utility for $(A\&O_j\&K_i)$ is equal to her utility for A conditional on $(A\&O_j\&K_i)$. When such a **K** and **O** exist, we have $\mathcal{U}(A) = \sum \rho(K_i)\mathcal{V}(A \mathbin{\&} K_i)$, which is precisely the condition in which CDT and EDT coincide. What Armendt shows, then, is that an appropriate representation theorem for CDT can be obtained by supplementing Fishburn's conditional decision theory with the assumption that every act A can be associated with a **K**-partition such that $A/K_i \approx A/(A\&K_i)$ for all i, and a partition of consequences **O** (dependent on A and **K**) such that $(A\&O_j\&K_i) \approx A/(A\&O_j\&K_i)$.

This is a nice result. Since Fishburn's theory is partition independent, it follows that CDT will be as well, provided that at least one pair of partitions **K, O** exist for each act A. The crucial questions are whether such partitions do exist in general, and whether we should think that the condition that defines **K** really does guarantee that $\mathcal{U}(A)$ and $\mathcal{V}(A)$ coincide. On this latter point we have our doubts, but even if it is granted, it seems unlikely to us, in the absence of further argument, that the appropriate **K**-partitions will exist in all cases where we would want to apply CDT. Indeed, it would be useful to have a representation theorem for CDT that does not need to assume the existence of any special partition of states or any canonical form for a decision problem to take.

A representation theorem with these desirable features is proven in Joyce (1999). Joyce sees both EDT and CDT as instances of an abstract conditional expected utility whose basic concept is that of the utility of an act A on the supposition that some condition C obtains. Joyce begins by characterizing supposition, or provisional belief revision, in terms sufficiently general to subsume Bayesian conditioning and Lewis's imaging as special cases. Given a subjective probability ρ defined over a σ-algebra Ω, and a distinguished subset $\mathcal{C}$ of *conditions* in Ω,[28] a *supposition for ρ relative to C* is a function $\rho(\bullet \mid \bullet)$ from $\Omega \times \mathcal{C}$ into the real numbers that satisfies

(a) $\rho(\bullet \mid C)$ is a countably additive probability on Ω for every $C \in \mathcal{C}$.
(b) $\rho(C \mid C) = 1$ for all $C \in \mathcal{C}$.
(c) $\rho(X \mid C \vee \neg C) = \rho(X)$ for all $X \in \Omega$.
(d) $\rho(X \mid B\&C) \geq \rho(X\&B \mid C)$ for all $X \in \Omega$ whenever $(B\&C) \in \mathcal{C}$.
(e) Let B and C be mutually incompatible conditions in $\mathcal{C}$. Then if one has $\rho(X \mid B) \geq \rho(X \mid C)$, then one has $\rho(X \mid B \vee C) \geq \rho(X \mid C)$, with equality if either $\rho(X \mid B) = \rho(X \mid C)$ or $\rho(B \mid B \vee C) = 0$.

The reader can verify that the ordinary conditional probability $\rho(X/C)$—that is, $\rho(X\&C)/\rho(C)$—is a supposition for p relative to $\mathcal{C} = \{C \in \Omega : \rho(C) > 0\}$. In fact, for any set of conditions C (even those containing conditions with zero prior probability), one can show that any map $\rho(\bullet \mid \bullet)$ defined on $\Omega \times \mathcal{C}$ that satisfies (a) – (c) plus

- *Bayes's Law*: $\rho(B \mid C)\rho(X \mid B\&C) = \rho(X \mid C)\rho(B \mid X\&C)$ whenever we have $(B\&C), (X\&C) \in C$.

is a supposition.[29] The imaging function $p^C = \rho(C \;\Box\!\!\rightarrow \bullet)$ associated with a similarity relation among possible worlds is also a supposition, but not typically

[28] The set C always takes the form $C = \Omega \sim I$, where the *ideal I* is a collection of Ω-propositions that contains the contradictory event $(X\&\neg X)$, is closed under countable disjunctions, and which contains $(X\&Y)$ whenever $X \in I$ and $Y \in \Omega$.

[29] These Bayesian suppositions were defined in Renyi (1955), and have come to be called "Popper measures" in the philosophical literature, after Popper (1934). Interested readers may consult van Fraassen (1995), Hammond (1994), and McGee (1994) for informative discussions of Popper measures.

one that satisfies Bayes's Law. There are also suppositions that are neither Bayesian nor instances of imaging.

Joyce, impressed by the need for partition-independent formulations of utility theories, uses the notion of a supposition to define an abstract conditional expected utility to be thought of as "utility under a supposition". Its (partition independent) basic equation is

$$\begin{aligned}\mathcal{V}(A|C) &= \sum_i \frac{\rho(S_i \,\&\, A|C)}{\rho(A|C)} u(A, S_i) \\ &= \sum_i \frac{\rho(X_i \,\&\, A|C)}{\rho(A|C)} \mathcal{V}(A \,\&\, X_i|C) \text{ for any partition } \{X_i\},\end{aligned} \tag{23.16}$$

where $\rho(\bullet \mid \bullet)$ might be any supposition function for ρ defined relative to a set of conditions C that contains propositions describing all an agent's actions (as well as other things). Since (23.16) is just EDT's (23.3) with $\rho(\bullet \mid C)$ substituted in for $\rho(\bullet)$, $\mathcal{V}(A|C)$ gives A's auspiciousness on the supposition that condition C obtains, where this supposition may be any provisional belief revision that satisfies (a) – (e).

As with Fishburn's theory, there is no guarantee that A's unconditional utility, which is now just $\mathcal{V}(A)$, will coincide with its utility conditional on itself,

$$\mathcal{V}(A|A) = \sum_i \rho(S_i|A)u(A, S_i)$$

The sole exception occurs when $\rho(\bullet \mid \bullet)$ is Bayesian, for in that case we have $\mathcal{V}(A) = \mathcal{V}(A \mid A)$, because $\rho(Si \mid A) = \rho(S_i\&A)/\rho(A)$ for all i. Given that $\mathcal{V}(A)$ and $\mathcal{V}(A \mid A)$ can differ in general, it becomes a live question whether a decision maker should choose acts that maximize her unconditional expected utility or choose acts that maximize expected utility conditional on the supposition that they are performed. Joyce argues, on grounds having nothing to do with the conflict between CDT and EDT, that a choiceworthy action is always one that maximizes expected utility on the condition that it is performed. The rational decision maker's objective, in other words, should always be to choose an A such that $\mathcal{V}(A \mid A) \geq \mathcal{V}(B \mid B)$ for all alternatives B. Neither evidential nor causal decision theorists will dispute this point, since the former endorse the prescription to maximize $\mathcal{V}(A \mid A)$ when the supposition function is $\rho(A \mid C) = \rho_c(A)$, which makes $\mathcal{V}\ (A|A)$ equal A's auspiciousness, and the latter endorse it when $\rho(A \mid C) = \rho(C \,\Box\!\!\rightarrow A)$, which makes $\mathcal{V}\ (A \mid A)$ equal A's instrumental expected utility. Thus, EDT and CDT are both instances of abstract conditional utility theory. The difference between them has to do not with the basic form of the utility function or with the connection between expected utility and choiceworthiness, but with the correct type of supposition to use in decision making contexts.

Once we recognize this, it becomes clear that the problem of proving a representation theorem for CDT can be subsumed under the more general problem of proving a representation theorem for an abstract conditional utility theory. And, since the function $\mathcal{V}$ (• | C) obeys Eq. (23.3) relative to any fixed condition C, this latter problem can be solved by extending the Bolker/Jeffrey axioms for unconditional preferences to conditional preferences, and showing that anyone who satisfies the axioms is sure to have conditional preferences that can be represented by some function $\mathcal{V}$ (A | C) of form (23.16) that is defined relative to a supposition function ρ(• | •) for her subjective probability ρ.

Joyce was able to accomplish this. We refer to reader to (1999) for the details, which turn out to be rather complicated, but the basic idea is straightforward. One starts by imagining an agent with a system of conditional preferences of the form: X on the supposition that B is weakly preferable to Y on the supposition that C, written $X \mid B \geq Y \mid C$. One assumes that this ranking obeys the usual axioms: transitivity, connectedness, a continuity principle, an Archimedean axiom, and so on. One also requires each *section* of the ranking $X \mid C \geq Y \mid C$, for C fixed, to satisfy the Bolker/Jeffrey axioms. Bolker's theorem then ensures that each section will be associated with a family r_C of ($\mathcal{V}$, ρ_C) pairs that satisfy Eq. (23.3) and that represent $X \mid C \geq Y \mid C$. Different r_C-pairs will be related by the equations $\rho_C{}^*(X) = \rho_C(X)[1 + k\mathcal{V}_C(X)]$ and $\mathcal{V}_C{}^*(X) = \mathcal{V}_C(X)[(k+1)/(1+k\mathcal{V}_C(X)]$, where k is any real number such that $[1 + k\mathcal{V}(X)] > 0$ for all propositions X such that $\mathcal{V}(X)$ is defined. The trick to proving a representation theorem for conditional decision theory is to find further constraints on conditional preferences under which it is possible to select a unique ($\mathcal{V}$, P_C) pair from each r_C in such a way that $\mathcal{V}(X) \geq \mathcal{V}$ (X) is guaranteed to hold whenever $X \mid B \geq Y \mid C$. The main axiom that is needed is the following generalization of Impartiality:

- Let X_1, X_2, and X_3, and Y_1, Y_2, and Y_3, be mutually incompatible, and suppose that

$$X_1 \,|B \approx Y_1|\, C > (X_1 \vee X_2)\,|B \approx (Y_1 \vee Y_2)|\, C > X_2\,|B \approx Y_2|\, C \tag{23.17}$$

holds for some conditions B and C. Then, if

$$X_1|B \approx Y_1|C \not\approx X_3|B \approx Y_3|C \not\approx (X_1 \vee X_3)|B \approx (Y_1 \vee Y_3)|C \tag{23.18}$$

then

$$(X_1 \vee X_2 \vee X_3)\,|B \approx (Y_1 \vee Y_2 \vee Y_3)|\, C. \tag{23.19}$$

This is not as complicated as it looks. If clause (23.17) holds, the only sort of conditional utility that will represent $X \mid C \geq Y \mid B$ will be one in which $\rho(X_1 \mid B)/\rho(X_1 \vee X_2 \mid B) = \rho(Y_1 \mid C)/\rho(Y_1 \vee Y_2 \mid C)$. Likewise, if (23.18) holds, then the representation must be one in which $\rho(X_1 \mid B)/\rho(X_1 \vee X_3 \mid B) = \rho(Y_1 \mid C)/\rho(Y_1 \vee Y_3 \mid C)$. Together these two equalities entail that $\rho(X_1 \mid B)/\rho(X_1 \vee X_2 \vee X_3 \mid B) = \rho(Y_1 \mid C)/\rho(Y_1 \vee Y_2 \vee Y_3 \mid C)$, and this is just what (23.19) guarantees.

Using this axiom as the main formal tool, Joyce is able to construct a full conditional expected utility representation for the ranking $X \mid C \geq Y \mid B$. By adding further conditions, one can ensure either that the representation's supposition function will be Bayesian or that it will arise via imaging from some similarity relation among possible worlds. In this way, both EDT and CDT are seen to have a common foundation in the abstract theory of conditional expected utility.

Conclusion

While the classical theory of Ramsey, de Finetti and Savage remains our best account of rational choice, its development has yet to be completed. An adequate theory should explain the role in decision making of causal thinking. True, a decision theory that did without causal propositions would have been nice: Cause and effect have long puzzled and divided philosophers and scientists, and theoretical discussion of causal methodology remains underdeveloped.[30] In decision making, though, we are stuck with causal thinking. Rational choice always involves judgements of how likely an option is to have various desirable consequences—and such judgements, we have argued, require the decision maker to have views, explicit or implicit, about causal or counterfactual relationships. Nothing else can substitute for these causal beliefs. The conditional credences employed by evidential decision theory cannot, because they are unable to distinguish causation from correlation. More refined "screening" techniques, while better at capturing causal connections, fail to apply in a important class of cases. To specify the kind of case to which they do apply, we must, one way or another, invoke causal relations.

We should not find this need for causal notions distressing. We draw causal conclusions all the time, after all, and scientists are able to glean causal tendencies from experiment and statistical data, using methods of high sophistication. Still, no one theory of causal notions has the precision and orthodox status of, say, the standard theory of subjective probability. Thus, an adequate decision theory, if we are right, must depend on new advances in our understanding of causation and its relation to rational belief. We might have wished that theoretical life had turned out easier, but as matters stand, important work on the foundations of utility theory remains to be done.

References

Adams, E. W. (1975). *The logic of conditionals*. Dordrecht: Reidel.

Armendt, B. (1986). A foundation for causal decision theory. *Topoi, 5*, 3–19.

Aumann, R. (1974). Subjectivity and correlation in randomized strategies. *Journal of Mathematical Economics, 1*, 67–96.

[30] Again, giant steps have been taken in this area since this article first appeared, especially by Spirtes et al. (1993) and Pearl (2000).

Aumann, R. (1987). Correlated equilibrium as an expression of Bayesian rationality. *Econometrica, 55*, 1–18.

Binmore, K. (1987). Modeling rational players I. *Economics and Philosophy, 3*, 179–212 (Reprinted, Binmore 1990).

Binmore, K. (1988). Modeling rational players II. *Economics and Philosophy, 4*, 9–55 (Reprinted Binmore 1990).

Binmore, K. (1990). *Essays on the foundations of game theory*. Oxford: Blackwell.

Bolker, E. (1966). Functions resembling quotients of measures. *Transactions of the American Mathematical Society, 124*, 292–312.

Eells, E. (1982). *Rational decision and causality*. Cambridge: Cambridge University Press.

Fine, K. (1975). Review of Lewis [24]. *Mind, 84*, 451–458.

Fishburn, P. (1974). A mixture-set axiomatization of conditional subjective expected utility. *Econometrica, 41*, 1–25.

Gibbard, A. (1986). A characterization of decision matrices that yield instrumental expected utility. In L. Daboni, A. Montesano, & M. Lines (Eds.), *Recent developments in the foundations of utility and risk theory* (pp. 139–148). Dordrecht: Reidel.

Gibbard, A., & Harper, W. L. (1978). Counterfactuals and two kinds of expected utility. In C. A. Hooker, J. J. Leach, & E. F. McClennen (Eds.), *Foundations and applications of decision theory* (Vol. I). Dordrecht: Reidel.

Hammond, P. (1994). Elementary non-Archimedean representations of probability for decision theory and games. In P. Humphries (Ed.), *Patrick Suppes: Scientific philosopher* (Vol. 1, pp. 25–61). Dordrecht: Kluwer Academic Publishers.

Harper, W. (1986). Mixed strategies and ratifiability in causal decision theory. *Erkenntnis, 24*, 25–26.

Harper, W. (1991). Ratifiability and refinements in two-person noncooperative games. In M. Bacharach & S. Hurley (Eds.), *Foundations of decision theory* (pp. 263–293). Oxford: Basil Blackwell.

Harsanyi, J., & Selten, R. (1988). *A general theory of equilibrium selection in games*. Cambridge: MIT Press.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81*, 945–960.

Horgan, T. (1981). Counterfactuals and Newcomb's problem. *Journal of Philosophy, 68*, 331–356.

Howard, J. V. (1988). Cooperation in the prisoner's dilemma. *Theory and Decision, 24*, 203–213.

Jeffrey, R. (1967). *The logic of decision*. New York: McGraw Hill.

Jeffrey, R. (1983). *The logic of decision* (2nd ed.). Chicago: University of Chicago Press.

Joyce. (1999). *The foundations of causal decision theory*. Cambridge: Cambridge University Press.

Kohlberg, E., & Mertens, J. (1986). On the strategic stability of equilibria. *Econometrica, 54*, 1003–1037.

Lewis, D. K. (1973). *Counterfactuals*. Cambridge: Harvard University Press.

Lewis, D. K. (1976). Probabilities of conditionals and conditional probabilities. *Philosophical Review, 85*, 297–315 (Reprinted, Lewis 1986, 133–152).

Lewis, D. K. (1979). Counterfactual dependence and time's arrow. *Noûs, 13*, 455–476 (Reprinted, Lewis 1986, 32–52).

Lewis, D. K. (1979). Prisoner's dilemma is a Newcomb problem. *Philosophy and Public Affairs, 8*, 235–240 (Reprinted, Lewis 1986, 299–304).

Lewis, D. K. (1980). A subjectivist's guide to objective chance. In Richard C. Jeffrey (Ed.), *Studies in inductive logic and probability* (Vol. 2). Berkeley: University of California Press. Reprinted (Lewis 1986), 83–113.

Lewis, D. K. (1981). Causal decision theory. *Australasian Journal of Philosophy, 59*, 5–30 (Reprinted, Lewis 1986, 305–337).

Lewis, D. K. (1986). *Philosophical papers* (Vol. 2). Oxford: Oxford University Press.

Luce, R. D., & Krantz, D. H. (1971). Conditional expected utility. *Econometrica, 39*, 253–271.

McGee, V. (1994). Learning the impossible. In E. Eells & B. Skyrms (Eds.), *Probability and conditionals* (pp. 179–197). Cambridge: Cambridge University Press.

Nozick, R. (1969). Newcomb's problem and two principles of choice. In N. Rescher (Ed.), *Essays in honor of Carl G. Hempel*. Dordrecht-Holland: Reidel.

Piccione, M., & Rubinstein, A. (1997). On the interpretation of decision problems with imperfect recall. *Games and Economic Behavior, 20*, 3–24.

Popper, K. (1934). *Logik der Forschung.* Vienna: Springer. Translated as *The logic of scientific discovery* (London: Hutchinson, 1959).

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. New York: Cambridge University Press.

Renyi, A. (1955). On a new axiomatic theory of probability. *Acta Mathematica Academiae Scientiarium Hungaricae, 6*, 285–335.

Robins, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period – applications to control of the healthy workers survivor effect. *Mathematical Modeling, 7*, 1393–1512.

Rubin, D. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology, 66*, 688–701.

Savage, L. J. (1972). *The foundations of statistics.* New York: Dover (First edition 1954).

Shin, H. S. (1991a). A reconstruction of Jeffrey's notion of ratifiability in terms of counterfactual beliefs. *Theory and Decision, 31*, 21–47.

Shin, H. S. (1991b). Two notions of ratifiability and equilibrium in games. In M. Bacharach & S. Hurley (Eds.), *Foundations of decision theory* (pp. 242–262). Oxford: Basil Blackwell.

Simon, H. A. (1957). *Models of man*. New York: Wiley.

Skyrms, B. (1980). *Causal necessity*. New Haven: Yale University Press.

Skyrms, B. (1984). *Pragmatics and empiricism*. New Haven: Yale University Press.

Skyrms, B. (1990). Ratifiability and the logic of decision. *Midwest Studies in Philosophy, 15*, 44–56.

Sobel, J. H. (1989). Partition theorems for causal decision theories. *Philosophy of Science, 56*, 70–95.

Spirtes, P., Glymour, C., & Scheines, R. (1993). *Causation, prediction, and search* (Lecture Notes in Statistics, Vol. 81). New York: Springer. ISBN: 978-1-4612-7650-0 (Print) 978-1-4612-2748-9 (Online).

Stalnaker, R. (1968). A theory of conditionals. In *Studies in logical theory* (American philosophical quarterly monograph series 2). Oxford: Blackwell.

Stalnaker, R. (1972). Letter to David Lewis. In W. Harper, R. Stalnaker, & G. Pearce (Eds.), *Ifs: Conditionals, belief, decision, chance, and time* (pp. 151–152). Dordrecht: Reidel.

Stalnaker, R. (1981). Letter to David Lewis of May 21, 1972. In W. L. Harper, R. Stalnaker, & G. Pearce (Eds.), *Ifs: Conditionals, belief, decision, chance, and time*. Dordrecht-Holland: Reidel.

Stalnaker, R., & Thomason, R. (1970). A semantic analysis of conditional logic. *Theoria, 36*, 23–42.

Van Fraassen, B. (1995). Fine-grained opinion, probability, and the logic of belief. *Journal of Philosophical Logic, 24*, 349–377.