Synthese (2007) 156:537–562
DOI 10.1007/s11229-006-9137-6

ORIGINAL PAPER

Are Newcomb problems really decisions?

James M. Joyce

Published online: 20 April 2007

Abstract Richard Jeffrey long held that decision theory should be formulated without recourse to explicitly causal notions. Newcomb problems stand out as putative counterexamples to this 'evidential' decision theory. Jeffrey initially sought to defuse Newcomb problems via recourse to the doctrine of ratificationism, but later came to see this as problematic. We will see that Jeffrey's worries about ratificationism were not compelling, but that valid ratificationist arguments implicitly presuppose causal decision theory. In later work, Jeffrey argued that Newcomb problems are not decisions at all because agents who face them possess so much evidence about correlations between their actions and states of the world that they are unable to regard their deliberate choices as causes of outcomes, and so cannot see themselves as making free choices. Jeffrey's reasoning goes wrong because it fails to recognize that an agent's beliefs about her immediately available acts are so closely tied to the immediate causes of these actions that she can create evidence that outweighs any antecedent correlations between acts and states. Once we recognize that deliberating agents are free to believe what they want about their own actions, it will be clear that Newcomb problems are indeed counterexamples to evidential decision theory.

Keywords Newcomb Problem · Richard Jeffrey · Causal Decision Theory · Evidential Decision Theory · Ratifiability · Freedom

Richard Jeffrey long held that decision theory should be formulated without invoking explicitly causal notions. The version of expected utility theory he championed in *The Logic of Decision*, Jeffrey (1983), did not use counterfactual conditionals, imaging functions, causally homogenous partitions, directed causal graphs, causal dependency hypotheses, or any of the other vehicles of causal information that "causal" decision

J. M. Joyce (✉)
Department of Philosophy, University of Michigan,
2215 Angel Hall, 435 South State Street,
Ann Arbor, MI 48109-1003, USA
e-mail: jjoyce@umich.edu

theories employ. Jeffrey did not, however, think that causal information is irrelevant to decision making. Rather, he maintained that, insofar as rational choice theory is concerned, an agent's beliefs about causal relationships can be adequately characterized in terms of her subjective conditional probabilities for non-causal propositions. Since these conditional probabilities characterize the agent's views about *evidential* relationships, Jeffrey's formalism is often referred to as *evidential* decision theory because it asks agents to choose *auspicious* acts that provide them with evidence for thinking that desirable outcomes will ensue.

Newcomb problems stand out as putative counterexamples to Jeffrey's theory. In these problems evidential decision theory asks agents to choose auspicious but *inefficacious* acts that provide evidence for desirable outcomes without doing anything to causally promote these outcomes. This has struck many decision theorists, including Jeffrey himself, as wrongheaded. The whole point of rational agency is to make choices that *change* things for the better, and this is something one can only do by performing acts that causally promote desirable consequences. When an agent chooses an auspicious but inefficacious act she gets "good news," but does nothing to improve her prospects.

Jeffrey initially sought to retain the "evidential" character of his theory in the face of Newcomb problems via recourse to the doctrine of *ratificationism*. This doctrine, which owes much to Eells (1982), advises you to "choose for the person you expect to be once you have chosen." (1983, p. 16) If an agent makes only ratifiable choices, Jeffrey argued, then the purely evidential import of her acts will be nullified. After she has made her choice she will know which action she plans to perform, and this knowledge will "screen off" any purely evidentiary relationships that might hold between her acts and states of the world. For reasons to be discussed below, Jeffrey came to see this response as problematic, and so gave up on ratificationism as a solution to Newcomb problems. Indeed, we shall see that the situation for ratificationism is worse than even Jeffrey imagined: it will be shown that the ratificationist argument presupposes the truth of causal decision theory.

In his final treatment of the topic, Jeffrey (1993, 2004), Jeffrey argued that Newcomb problems present no obstacle to evidential decision theory *because they are not really decision problems at all*! Agents who face them, he claimed, possess so much evidence about correlations between their actions and states of the world that they are unable to regard their deliberate choices as causes of outcomes. Such agents do not see themselves as making free choices. Jeffrey's reasoning goes wrong, I shall argue, because it fails to recognize that an agent's beliefs about her own (immediately available) actions are so closely tied to the *immediate causes* of these actions that she is in a position to *create evidence* that will outweigh any antecedent correlations might have existed between acts and the states. Once we recognize that that freedom consists, in part, in being free to believe what one wants about one's own actions, we will see that Newcomb problems are genuine decision problems, and that they do indeed serve as counterexamples to evidential decision theory.

1 Evidential versus causal decision theory

The central difference between the evidential and causal theories concerns, not what an agent should do, but whether we can adequately characterize her *rationale* for doing what she should do without making explicit reference to her beliefs about what

her choices are likely to cause. Causal decision theorists maintain that there is no avoiding causality.[1] To know what an agent should do in a given situation we must know what she believes about the *effects* of her actions, and this is only possible if we make reference either to her beliefs about propositions with explicitly causal content or to forms of belief revision that are subject to explicitly causal constraints.[2] Evidential theorists dispute this. They claim that, for purposes of decision theory, we can capture the relevant beliefs about causes and effects by appealing to nothing more than the agent's ordinary subjective probabilities for non-causal propositions.

In the simplest sorts of cases, the issue boils down to a dispute about the probabilities that figure into calculations of expected utility. Imagine a 2×2 decision matrix

	E	$\sim E$
A	$des(A \,\&\, E)$	$des(A \,\&\, \sim E)$
$\sim A$	$des(\sim A \,\&\, E)$	$des(\sim A \,\&\, \sim E)$

in which A and $\sim A$ are propositions that describe possible acts and E and $\sim E$ are propositions that describe the "states of the world" on which the outcomes of these acts depend. In Jeffrey's framework outcomes are act/state conjunctions, and each has a utility (or "desirability") as listed in the matrix. Early decision theorists, following Savage (1972), computed the expected utility of an act by weighting the utilities of its outcomes by the probabilities of the events that bring them about, so that

$$\textbf{EU} \quad V(A) = prob(E)des(A \,\&\, E) + prob(\sim E)des(A \,\&\, \sim E)$$
$$V(\sim A) = prob(E)des(\sim A \,\&\, E) + prob(\sim E)des(\sim A \,\&\, \sim E)$$

Among other things, this entails the unrestricted "sure-thing" principle of Savage (1972, pp. 21–22).

> *Sure-thing Principle* (**STP**). For any event E, if an agent prefers A to $\sim A$ both on the supposition that E obtains and on the supposition that E does not obtain, then she should prefer A to $\sim A$ unconditionally.

This rule, which relates an agent's conditional preferences to her unconditional preferences, endorses a strong kind of *dominance reasoning*: if A is the better act whatever E's truth-value, then A is the better act simpliciter.

Jeffrey recognized that **EU** and **STP** are only valid if the agent believes that her choice will not causally affect E's occurrence. Suppose the agent reasoned thus: "I can pay \$10 for an influenza vaccination or I can keep my \$10. Either I will get the flu this winter or I won't. If I am to get the flu, I'd rather have the extra \$10 to buy medicine. If I do not get the flu, I'd rather have the \$10 to buy hot cocoa. So, whether I get the flu this winter or not, I prefer having the extra \$10 to not having it. Thus, by **STP**, I should forgo the vaccination." This is clearly absurd: if the agent believes that the vaccination will inhibit influenza, then choosing to forgo it is choosing to make the undesirable outcome more probable and the desirable outcome less probable.

[1] For defenses of causal decision theory see Gibbard and Harper (1978), Cartwright (1979), Skyrms (1980), Lewis (1981), Sobel (1985), Armendt (1986), and Joyce (1999).

[2] These two options mark an important distinction among causal decision theorists. Some, like Lewis, Skyrms, and Gibbard and Harper, represent the relevant causal beliefs using unconditional subjective probabilities of propositions with causal content. Others, like Sobel and Joyce, prefer to incorporate the causal elements into the process of belief revision.

The moral Jeffrey drew from cases like this was that if the agent suspects that A or $\sim A$ might influence E, then the correct expected utilities are given not by **EU** but by[3]

$$\textbf{EDT}\ V(A) = prob(E/A)des(A\,\&\,E) + prob(\sim E/A)des(A\,\&\sim E)$$

$$V(\sim A) = prob(E/\sim A)des(\sim A\,\&\,E) + prob(\sim E/\sim A)des(\sim A\,\&\sim E)$$

Here the utility of each outcome is multiplied by the agent's subjective probability for its occurrence *conditioned on the evidence that the act in question is in fact performed*. This requires the deliberating agent to treat information about which action she will ultimately perform as "news items," on a par with information about any other aspect of the world. On this reading, the value of choosing an action is the same as the value of being told by a reliable soothsayer that one will perform it. Actions are thus evaluated on the basis of the evidence they provide for thinking that desirable results will ensue.

EDT sanctions only a restricted form of the sure-thing principle.

> **STP**$_{\mathrm{E}}$. For any event E that is evidentially independent of both A and $\sim A$, so that $prob(E/A) = prob(E/\sim A)$, if the agent prefers A to $\sim A$ both on the supposition that E obtains and on the supposition that E does not obtain, then she should prefer A to $\sim A$ unconditionally.

The requirement that E be evidentially independent of A and $\sim A$ is meant to prevent the application of **STP** in cases like the influenza example, where the agent thinks that she can causally influence the state of the world.

Jeffrey initially thought that the evidential relationships encoded in ordinary conditional probabilities would provide as much information about causal relations as decision theory requires. In any case where it mattered, he hoped, an agent's views about the differing causal powers of A and $\sim A$ vis-à-vis E would be revealed in the disparity between $prob(E/A)$ and $prob(E/\sim A)$. Since the difference $prob(E/A) - prob(E/\sim A)$ is one way to measure of the *amount of evidence* that A provides for E, the substance of Jeffrey's initial proposal can be rendered as follows:

> **J** For purposes of decision theory, an agent regards A as more efficacious than $\sim A$ as a cause of E when $prob(E/A) > prob(E/\sim A)$. Moreover, the quantity $prob(E/A) - prob(E/\sim A)$ is her quantitative estimate of the degree of A's causal efficacy as a promoter of E.[4]

Causal decision theorists deny this, arguing that *no* function of $prob(E/A)$ and $prob(E/\sim A)$ can ever adequately encode the agent's beliefs about what her acts are likely to cause. While the values of these conditional probabilities depend, partly, on the agent's judgments about causal linkages between E and A, they can also reflect

[3] Throughout I will use $prob(X/Y)$ to denote the conditional probability of X given Y. It is defined as $prob(X/Y) = prob(X\,\&\,Y)/prob(Y)$ when $prob(Y) > 0$.

[4] While Jeffrey does not make the second point explicitly, it does follow from his theory. To see why, consider the special case in which A and $\sim A$ produce equally desirable outcomes under both E and $\sim E$, with the better outcome being produced by E. We can then set $des(A\,\&\,E) = des(A\,\&\sim E) = 1$ and $des(\sim A\,\&\,E) = des(\sim A\,\&\sim E) = 0$, and write the difference in the expected utilities of A and $\sim A$ as $V(A) - V(\sim A)\ prob(E/A) - prob(E/\sim A)$. Since there is no reason for the agent to prefer A to $\sim A$ in this context unless she thinks that A's performance promotes E's occurrence, it follows that, as far as decision making is concerned, the difference $prob(E/A) - prob(E/\sim A)$ may be interpreted as a reflection of the degree to which the agent regards A as a cause of E.

the agent's knowledge of non-causal correlations between E and A. For example, $prob(E/A) - prob(E/\sim A)$ can be large when A and E are correlative effects of a common cause, but are otherwise causally disconnected.

The contrast between the causal and evidential approach to decision theory is brought out most clearly in Newcomb problems. The initial Newcomb problem was a bizarre science fiction story involving a master psychologist who could almost infallibly predict an agent's actions. Unfortunately, focus on this example has lead many to think that Newcomb problems are so recherché as to bear little relevance to actual decision making. In fact, they are fairly common. Consider, for example the *Twin's Dilemma*, in which two agents, Row and Column, must decide whether or not to take some cooperative action as described in the following decision table:

	C	$\sim C$
R	9, 9	0, 10
$\sim R$	10, 0	1, 1

Row chooses a row, Column chooses a column, and the listed pairs give their respective utilities for each outcome. Three conditions must hold for Row to be facing a true Newcomb problem:

- Row must know that Column's action is *causally independent* of her own, i.e., she must know that she has no way of causally influencing what Column will choose or do.
- Row must believe she and Column are somewhat like-minded. Specifically, her subjective probabilities should meet the (rather mild) condition that $prob(C/R) - prob(C/\sim R) > 1/9$.
- Row must take herself to be a free agent; she must believe that she has the ability to freely choose either R or $\sim R$.

The first of these conditions is easily met. Indeed, we can ensure it by having the players make simultaneous choices at distant locations so that no causal signal can pass between them. To satisfy the second condition we need only imagine that the players acquired their decision making skills in a common setting: perhaps they grew up in the same neighborhood, or had the same decision theory teacher, or have played the game many times before, or whatever. Given the symmetry of the situation, and their common background, Row might well believe that, whatever she chooses, Column is a bit more likely (probability $> 1/9$) to choose the same way. On the face of it, the third condition seems trivial as well. We can imagine that Row is sitting before two buttons marked R and $\sim R$, and that all she has to do is to reach out and push one. What could be freer than that? We will return to this question shortly.

When the above three conditions are met, the evidential theory encourages Row to reason thus: "Since Column is significantly more likely to perform C if I do R than if I do $\sim R$ (in the sense that $prob(C/R) - prob(C/\sim R) > 1/9$), and since I am better off by nine utiles if he chooses C, it follows that I should choose R." This reasoning recommends choosing the auspicious but inefficacious R, which provides Row with evidence for thinking that Column will perform C, over the efficacious but inauspicious $\sim R$, which will earn her an extra utile *whatever Column does* (despite being correlated with the highly undesirable $\sim C$).

Causal theorists see this as a fallacy. Row is passing up a sure utile to secure the "good news" that outcome $R \,\&\, C$ is probable. This would make sense if Row were, in

Jeffrey's phrase, "making the news" that R & C, but in this situation the only "news" Row makes concerns the row of the outcome to be received, COLUMN independently fixes its column. In effect, Row's control over the news extends only to the question of getting the extra utile. On the causal view, Row is obliged to take the course of action that is most likely to cause desirable results, and this is remains true even when this means getting some "bad news" on the side.

EDT goes wrong by weighting the utilities of outcomes by the conditional probabilities of states given acts. Instead, the relevant utilities should be computed according to the formulas

$$\textbf{CDT } \boldsymbol{U}(A) = prob(E\backslash A)des(A \,\&\, E) + prob(\sim E\backslash A)des(A \,\&\, \sim E)$$
$$\boldsymbol{U}(\sim A) = prob(E\backslash \sim A)des(\sim A \,\&\, E) + prob(\sim E\backslash \sim A)des(\sim A \,\&\, \sim E)$$

where $prob(E\backslash A)$ and $prob(E\backslash \sim A)$ are *causal probabilities* that captures the agent's views about the relative causal powers of A and $\sim A$ vis-à-vis E. Causal theorists interpret these quantities in a variety of ways,[5] but all agree that (a) the agent regards A as a promoting cause of E exactly if $prob(E\backslash A) - prob(E\backslash \sim A) > 0$, and (b) these are not ordinary conditional probabilities, i.e., $prob(E\backslash A) \neq prob(E/A)$ and $prob(E\backslash \sim A) \neq prob(E/\sim A)$.

Since **CDT** reduces to **EU** when acts have no causal influence over states, the causal theory entails **EU** when E is causally independent of A.[6] Thus, the causal decision theorist endorses the following version of the sure-thing principle

> **STP**$_C$. For any event E that is causally independent of A and $\sim A$, so that $prob(E\backslash A) = prob(E\backslash \sim A)$, if the agent prefers A to $\sim A$ both on the supposition that E obtains and on the supposition that E does not obtain, then she should prefer A to $\sim A$ unconditionally.

Accordingly, in Newcomb problems, causal decision theory always recommends that Row choose the efficacious act $\sim R$, even though this provides her with evidence for thinking that the undesirable state $\sim C$ obtains.

Most philosophers, Jeffrey included, now agree that choosing the merely auspicious R is a mistake. The vexed question is whether one needs the entire apparatus of causal decision theory, with its invocation of mysterious causal probabilities or causal independence conditions, to get this answer.

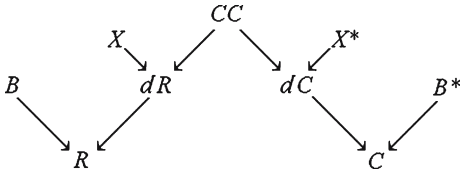
2 Ratificationism

Some evidentialists have sought to defuse Newcomb Problems by arguing that EDT *does not* recommend the merely auspicious choice because the agent will make her decision from an epistemic perspective in which the R and C are evidentially independent. In his "tickle defense," Eells (1982) argued that a fully rational agent will come to know what she intends to do before she acts, and that this information will

[5] See (Joyce, 1999, pp. 161–180) for a discussion of the options.

[6] Unlike evidential independence, the notion of causal independence, as it is being used here, is non-symmetric. When I say that E is causally independent of A I mean that changes in A's truth-value do not cause changes in E's truth-value. It does not follow that changes in E's truth-value do not cause changes in A's truth-value. Thus, E might be causally independent of A even though A is causally dependent on E. Indeed, this will happen whenever E is a cause of A.

Fig. 1



"screen off' any evidential connections there might have been between her acts and states of the world. In a similar vein, Jeffrey's doctrine of *ratificationism* suggests that a rational agent should choose for the person she expects to be once she has chosen by evaluating each action on the hypothesis that she will ultimately decide to perform it. On either view, a rational agent's *ability to anticipate her own decisions* nullifies any purely evidential correlations that might exist between states and acts.

To see how this works, let dA denote the *decision to perform A*. This is a different proposition from A itself; it is at least logically possible for the agent to decide on an act and yet not perform it, so that $\sim A \ \& \ dA$ and $A \ \& \ d \sim A$ are non-contradictory. We assume that all conditional probabilities of the form $prob(\pm C/\pm R \& d \pm R)$ make sense.[7] Jeffrey's ratificationist solution to the Newcomb problem is based on the following principles:

Screening. The decision to perform an act screens off any evidential correlations that might exist between that act and states of the world , so that in Twin's Dilemma one has

$$prob(C/R \,\&\, dR) = prob(C/\sim R \,\&\, dR) = prob(C/dR)$$
$$prob(C/R \,\&\, d\sim R) = prob(C/\sim R \,\&\, d\sim R) = prob(C/d\sim R)$$

Maxim of Ratifiability . An agent can rationally perform act A only if A is *ratifiable* in the sense that there is no alternative B whose expected utility exceeds that of A on the supposition that A is decided upon.

Screening is an instance of the *Causal Markov Condition*[8] that figures so prominently in contemporary approaches to Bayesian causal modeling. (See Pearl, 2000; Spirtes, Glymour, & Scheines, 2000) Roughly, the Causal Markov Condition says that, given a sufficiently complete description of the causal structure of a situation, knowledge of all an event's direct causes (its "causal parents") makes any other information statistically irrelevant to the event's occurrence (except for information about the event's own "causal descendents"). For the Twin's Dilemma, the picture Jeffrey has in mind is as described in (Fig. 1).

Letters denote causal variables, and each arrow indicates that the variable at its tail is a causal parent of the one at its head. An assignment of values to all parents of a given variable causally determines the variable's value. The graph is assumed to be *complete* in the sense that variables, like B and B^*, with no common causal ancestors are both causally and statistically independent. dR is the variable of Row's decision,

[7] To ensure this Jeffrey introduced a "trembling hand" condition that requires the agent to believe that, for *any* two actions, there is some positive probability that she will decide to do one of them but end up doing the other. If one thinks conditional probabilities can be well-defined even when their conditions have probability zero (Joyce, 1999, pp. 201–214), then such a trembling hand condition is unnecessary.

[8] This misnamed principle is a generalization of Reichenbach's "common cause" principle (1956, Ch. 3).

while dC is the column player's choice. These decisions, it is assumed, are correlated in virtue of having a common cause (or set of causes) CC. Both decisions, as well as the player's subsequent actions, might depend on other causal variables, X and B or X^* and B^* as the case may be. Crucially though, these other variables are *not* correlated, so that, e.g., $prob(B/B^*) = prob(B)$. This ensures that the only causal or evidential relationships that obtain between R and C hold in virtue of the path through CC that connects them.

In Twin's Dilemma, the existence of this path provides the basis for a "backtracking" inference from R to C. Since effects provide information about their causes, Row can reason "up the graph" from a truth-value for R to some conclusion about CC. The sort of inference she is able to make will depend on the exact relationship between B, X, X^*, dR and CC, but in a genuine Newcomb problem she will learn enough about CC's value to reason "down" to a conclusion about C's value (using the principle that causes provide evidence for their effects). Again, the sort of inference she can make will depend on the precise relationship between B^*, X, X^*, dC and CC, but in a Newcomb problem she will be able to assign subjective probabilities such that $prob(C/R) - prob(C/{\sim}R) > 1/9$.

The Markov condition blocks such "backtracking" inferences by ensuring that every variable in the graph, other than B, is evidentially independent of R conditional on a value for dR. Consequently, Row will not see any evidential connection between her acts and Column's acts *after* she comes to know her decision. Moreover, she will prefer news of ${\sim}R$ to news of R conditional on *whatever* decision she makes. If she decides on R, and so learns dR, then her evidential utilities are $V(R/dR) = 9prob(C/dR) < V({\sim}R/dR) = 9\,prob(C/dR) + 1$. If she decides on ${\sim}R$, and so learns $d{\sim}R$, then $V(R/d{\sim}R) = 9prob(C/d{\sim}R) < V({\sim}R/d{\sim}R) = 9prob(C/d{\sim}R) + 1$. Since ${\sim}R$ has an additional unit of "news value" either way, Row can be certain that she will prefer ${\sim}R$ to R once she knows her decision, *no matter what that decision turns out to be*. This highlights a clear asymmetry between the two acts. Row knows in advance that if she decides on R she will regret her choice because, from the perspective of her *new* epistemic state, ${\sim}R$ will be better news. She also knows that if she decides on ${\sim}R$, she will be pleased with what she has done and so will "ratify" her decision.

But why should the fact that ${\sim}R$ is ratifiable in this way make a difference to what Row chooses? Even though ${\sim}R$'s news value exceeds that of R when Row is *certain* of her decision, the reverse is true while she is still trying to make up her mind. The Maxim of Ratifiability supplies the last, crucial step in Jeffrey's argument. In the context of evidential decision theory, it requires the agent to subordinate her *current* estimates of the auspiciousness of acts to her *future* estimates of their auspiciousness, at least to the extent that she can determine what her future estimates are. Thus, the combination of the Maxim of Ratifiability and evidential decision theory imposes the following necessary condition on rational choices: An agent can rationally perform the act A only if A is *evidentially ratifiable* in the sense that $V(A/dA) > V(B/dA)$ for every alternative act B.

Jeffrey initially maintained that, in the presence of Screening, this solves the Newcomb without leaving the confines of the evidential theory. Even though performing R on the basis of a decision to perform R provides Row with better news than does performing ${\sim}R$ on the basis of a decision to perform ${\sim}R$, only the efficacious act ${\sim}R$ is ratifiable. Thus, as Jeffrey put it,

> the [Twin's dilemma] exemplifies a class of problems in which the agent foresees during deliberation that the choice will change [her] subjective probabilities so as to make the dominant act be the one whose performance will have the highest estimated desirability once a decision (either decision) is made, even though it did not have the highest estimated desirability when the agent's uncertainty about his decision was taken into account, before the choice. It seems clear that one should choose an act whose performance one expects to have highest expected desirability if and when it is chosen. (1983, p. 17)

Jeffrey soon came to reject this response to Newcomb problems because, as Bas van Fraassen pointed out, Screening is implausible in some versions of Twin's Dilemma. Consider an example given in (Joyce, 1999, p. 159). Suppose Row is a bumbler who recognizes that she sometimes fails to carry out her decisions. One can think of the variable B in Fig. 1 as controlling Row's bumbling. Column might also be a bumbler, with his bumbling controlled by B^*. If we suppose that both players tend to bumble in a similar fashion, and that both know this, then it can happen that R rather than $\sim R$ is the uniquely ratifiable act. If the probabilities are

$$\begin{array}{ll} prob(R\,\&\,C/dR) = 0.6 & prob(R\,\&\,C/d{\sim}R) = 0.1 \\ prob(R\,\&\,{\sim}C/dR) = 0.1 & prob(R\,\&\,{\sim}C/d{\sim}R) = 0.2 \\ prob({\sim}R\,\&\,C/dR) = 0.1 & prob({\sim}R\,\&\,C/d{\sim}R) = 0.1 \\ prob({\sim}R\,\&\,{\sim}C/dR) = 0.2 & prob({\sim}R\,\&\,{\sim}C/d{\sim}R) = 0.5 \end{array}$$

then R and $\sim R$ provide better evidence about C than do dR and $d{\sim}R$ alone. The news values work out so that $V(R/dR) = 7.71 > V({\sim}\mathrm{R/dR}) = 4$ and $V(R/d{\sim}R) = 3 > V({\sim}R/d{\sim}R) = 2.5$, which makes R uniquely ratifiable.

While this sort of argument led Jeffrey to give up on ratificationism as a solution to Newcomb's problem, considerations introduced in Eells (2000) suggest that Jeffrey might have been too hasty. Eells argues that van Fraassen's example should be "disqualified" because "rational deliberation loses its point" when there is a correlation among acts that cannot be screened off by decisions. (p. 896) The crux of the matter is that Row, who perceives a correlation between her bumbling tendencies and those of Column, will regard R's truth-value as lying outside of the sphere of things that she can directly control. Moreover, she will seek to explain the correlations between the R and C that remain after conditioning on dR or $d{\sim}R$ by postulating the existence of a common cause, which lies outside her control (Fig. 2). For her, the correct picture is not Fig. 1, in which there is an *unexplained* correlation between B and B^* (thus violating the Causal Markov Condition). Here the players' bumbling has a common cause CC^* that jointly influences the truth-values of R and C *via pathways that do not pass through their decisions*. Thus, each player believes that her action depends on factors that causally influence the other player's action without affecting his decision. In such a situation, Row cannot regard herself as truly free in the matter of R since its truth-value is not determined solely by her deliberate choices. As Eells observes, "to the extent to which you believe that your actual act is influenced by factors other than your decision (the result of rational deliberation), it would seem that, to you rational deliberation will lose its point." (p. 896) In effect, Eells is arguing that van Fraassen's version of the Newcomb problem is not a genuine decision.

I have come to think Eells is right. Row will see herself as fully free in the matter of R only if she believes that external factors influence R's truth-value solely

Fig. 2

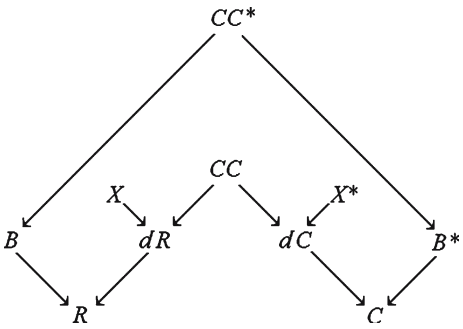
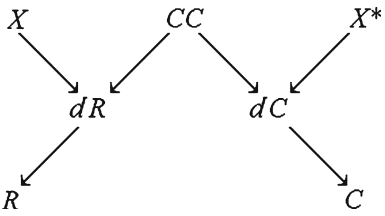


Fig. 3



by influencing her decision. She must, that is, see herself as facing a problem like this (Fig. 3).

If this is Row's situation, then knowledge of her decision will screen off any purely evidential correlations that may hold between her acts and those of the other player. So, to concede Eells's point, Screening is plausible in any Twin's Dilemma played by agents who see themselves as making a genuinely free choice. Accordingly, I grant that "the ratificationist defense is stronger if one makes the (perhaps idealized) assumption that the [agent's] act is not correlated with decision relevant causes that bypass the process of rational deliberation." (p. 896)

Still, the evidential ratificationist is not home free. Since the objections to Screening have seemed so decisive (even to Jeffery), critics have largely neglected the role of the Ratifiability Maxim in this argument. This is regrettable because, as we will see, its use is not nearly so straightforward as ratificationists make it seem. Despite appearances, the combination of evidential decision theory and the Maxim of Ratifiability does *not* provide a sound rationale for choosing efficaciously in Twin's Dilemma and other Newcomb problems (even granting Screening). To justify the efficacious choice one must augment the Maxim with explicitly causal principles. In reality, then, the evidential ratificationist's rationale for choosing the dominant act in Twin's Dilemma and other Newcomb problems relies on a hidden appeal to causal decision theory!

To see the lacuna in the ratificationist argument note first that, as Jeffrey stressed, the Maxim of Ratifiability is a only necessary condition; it merely gives agents a way of ruling out acts that cannot be rationally performed. As a necessary condition it has a great deal of plausibility. Unratifiable acts do seem genuinely defective: if one cannot choose to perform A without thereby giving oneself a compelling reason not to perform A, then one should not choose A. This necessary condition, however, is too weak to justify the efficacious choice in Twin's Dilemma. Ruling out R as a rational choice is not the same as showing that $\sim R$ is the rational choice. First, it could be that

there is *no* rational choice in Twin's Dilemma.[9] Second, if $\sim R$ really is the unique rational choice, then it should be possible to explain why by showing that $\sim R$ is better than R is at achieving the agent's ends.

To put the point in a more general way, evidential ratificationists seem to be imagining rational choice as a two-stage process. First, the Maxim of Ratifiability separates those acts that can be ratified from those that cannot. Second, the agent chooses *from among her ratifiable options* on the basis of expected utility considerations. The fact that some act outside the ratifiable set may have a higher expected utility than any act within the set is deemed irrelevant. Unratifiable acts are simply left out of the picture; they are treated like actions that the agent has lost the power to perform (though, it is crucial to keep in mind, she has not lost this power). Thus, the evidential ratificationist's argument requires the following strengthening of the Maxim of Ratifiability:

> *Strong Ratificationism.* If an agent faces a decision that contains at least one ratifiable act, then she should choose a ratifiable act whose expected utility is not exceeded by that of any other ratifiable act. The chosen act's expected utility is *not* required to meet or exceed that of all the unratifiable acts.

This ensures that there will be a rational choice in any decision with at least one ratifiable act. Moreover, when there is exactly one ratifiable act, as in Twin's Dilemma, then that act is the unique rational choice no matter what its expected utility might be.

The challenge for evidentialists is to show why Strong Ratificationism is reasonable by explaining why the fact that a given act has maximal expected utility among the ratifiable options makes that act choiceworthy even though it does not maximize utility *tout court*. The core of the problem is that an act like $\sim R$ in Twin's Dilemma, though ratifiable, is still lousy news, which means that the agent faces a choice between an unratifiable act of high news value and a ratifiable act of low news value. It looks as though an evidentialist should read this as an indictment of both acts: R is a bad choice in virtue of being unratifiable, while $\sim R$ is a bad choice in virtue of failing to maximize news value. To rebut this charge, evidentialists must explain how, in light of R's unratifiability, the fact that $\sim R$ is ratifiable provides the agent with a reason for doing $\sim R$ that is sufficient to override or outweigh its low news value. $\sim R$'s ratifiability must somehow be portrayed as a reason *in favor* of performing it.

Twin's Dilemma would seem to afford evidentialists with their best chance of doing this because its one ratifiable act $\sim R$ is also *uniformly ratifiable*: it uniquely maximizes news value conditional on *every* decision that she might make, so that $V(A/dB) > V(A^*/dB)$ for all acts $A^* \neq A$ and B. In this special case, Strong Ratifiability reduces to

> **SR***. If the set of an agent's possible options contains an option A whose performance is preferable to all others conditional on any decision that she might make, then A is her uniquely rational choice.

Since $\sim R$ is uniformly ratifiable, evidentialists can close the hole in their argument by supplying a compelling justification for **SR***.

At first glance, it seems easy to justify **SR***. Indeed, there is a plausible rationale for the Maxim of Ratifiability that extends to **SR*** almost without modification. Why

[9] For an example, see the "Death in Damascus" example in Gibbard and Harper (1978).

should "one choose for the person one expects to be once one has chosen?" The answer seems obvious: it is because one will be better informed at that time. The main difference between a decision maker who has yet to settle on a definite course of action and her "post-choice self" is that the latter knows her choice, and so possesses more relevant information about the consequences of her actions. Generally, a person with more relevant information about a decision is better positioned to make a wise choice than a person with less information. Thus, it seems reasonable to introduce the following general principle:

> *Preference Reflection*.[10] A rational agent should conform her preferences among acts to those of her better-informed post-decision self, at least insofar as she can determine what those preferences are.

Now, if an act is unratifiable, the agent knows she will rue her decision to perform it. It would therefore be unwise for her to perform an unratifiable act since this would be to do something that her better-informed post-choice self is certain to regard as ill advised. So, it is irrational to choose unratifiable acts. In this way, the Maxim of Ratifiability is justified on the basis of the principle that rational agents should be guided by the preferences of their better-informed post-decision selves.

To extend this justification to cover **SR***, note first if an action is uniformly ratifiable then the agent is sure to prefer it after making her decision, no matter what decision she makes. Since the agent should be guided by her better-informed post-choice self, and since her post-choice self is sure to prefer the uniformly ratifiable act, it follows that she should perform this act. It seems, then, that we have found a justification for **SR*** based, again, on Preference Reflection.

To evaluate this justification we must clarify the import and status of Preference Reflection. To this end, imagine two agents, "Doer" and "Knower," who are considering the same decision. Doer is charged with choosing an act, but both agents will receive whatever outcome results from her choice. We assume that

(a) Both agents have identical interests—their utilities *for outcomes* are the same—and Doer knows this.[11]
(b) Each agent evaluates prospects on the basis of the same sort of subjective expected utility (evidential or causal)—and Doer knows this.
(c) Knower is better situated epistemically—he knows everything Doer knows, but has reliable information Doer lacks—and Doer knows this.

When (c) holds Doer thinks that Knower's beliefs are likely to be more accurate than her own. In the terminology of Gaifman (1988), she regards Knower as an *epistemic expert* in the sense that, given their respective subjective probabilities, *prob* for Doer

[10] This name should remind readers of the Bas van Fraassen's *Reflection Principle*, which is defended in his (1984).

[11] Note that in a well-formed decision problem, the utilities of outcomes do not change when an agent acquires new information (though the expected utilities of less specific prospects might well change). In the current context, this means that neither Doer nor Knower will change their utilities *for outcomes* when they acquire information about what the other believes or prefers.

and *PROB* for KNOWER,[12] the following holds

- For any proposition X (relevant to the decision) and any number $0 \leq x \leq 1$, $prob(X/PROB^*(X) = x) = x$ whenever $prob(PROB^*(X) = x) > 0$.[13]

This says that DOER aligns her subjective probabilities with KNOWER's to the extent that she can determine what KNOWER's subjective probabilities are. In addition to seeing KNOWER as an epistemic expert, DOER might also regard him as a *decision-making expert* in the sense that:

- For any acts A and B, DOER prefers A to B conditional on the hypothesis that KNOWER prefers A to B.

This entails that DOER *unconditionally* prefers A to B whenever she is convinced that KNOWER prefers A to B.[14]

It is easy to find situations in which it would be reasonable for DOER to view KNOWER as an epistemic expert. Suppose, for example, that DOER knows that KNOWER's epistemic state differs from her own only insofar as he has undergone a *learning experience*[15] that has revealed the truth-value of some proposition Y about which she remains uncertain. Since DOER is then certain both that $PROB^*(\bullet) = prob(\bullet/Y)$ if Y and that $PROB^*(\bullet) = prob(\bullet/\sim Y)$ if $\sim Y$, it follows that she regards KNOWER as an epistemic expert.[16]

When (a) and (b) hold, DOER should view KNOWER as a decision-making expert, and so emulate his preferences, whenever she regards KNOWER as an epistemic expert. For DOER can reason as follows: "Since KNOWER and I have the same interests and are equally rational, and since he knows more than I do about the decision I face, he is clearly better positioned than I am to see which act is best for me. Hence, I should want to do whatever he would want me to do (to the extent that I can determine what this is)." The upshot of this reasoning is encapsulated in the following principle, which is a more precise version of Preference Reflection:

K When (a) and (b) hold, if DOER regards KNOWER as an epistemic expert, then she should also regard him as a decision-making expert.

[12] Hereafter the "*" indicates *non-rigid* designation. Whereas "$PROB(X) = x$" expresses a (necessary) identity between two numbers, "$PROB^*(X) = x$" expresses the contingent identity that the probability (whatever it is) that KNOWER assigns to X is the number x. One needs to be careful about rigid and non-rigid designation here to avoid *Miller's fallacy*. See van Fraassen (1984).

[13] For the sake of simplicity, I am leaving a few things out of this definition. For instance, DOER must have perfect knowledge of her own subjective probabilities, and must be convinced that KNOWER has perfect knowledge of his subjective probabilities. This requirement is important in contexts where experts might be *self-undermining* in the sense that learning the probabilities that they assign to propositions will alter those very probabilities. These cases need not concern us here.

[14] DOER can regard KNOWER as an epistemic or a decision-making expert even though she does not know what he believes or prefers. In these cases, DOER's subjective probability for a proposition X will be her expectation of KNOWER's subjective probability for X, and her expected utility for any act A will be her expectation of KNOWER's expected utility for A.

[15] DOER must see this as an experience that increases the accuracy of KNOWER's beliefs. Belief changes that are not the result of such learning experiences will not lead DOER to regard KNOWER as an epistemic expert. For a discussion of learning experiences see Skyrms (1987).

[16] Proof: Since DOER knows her own subjective probabilities, $prob(PROB^*(X) = x) > 0$ only when x is either $prob(X/Y)$ or $prob(X/\neg Y)$. Since $prob(PROB^*(X/Y) = prob(X/Y)$ iff $Y) = 1$ it follows that $prob(X/\ PROB^*(X/Y) = prob(X/Y)$ iff $Y) = prob(X/Y)$. The same holds true with Y replaced by $\sim Y$.

This captures the core of the idea that, all else equal, agents who know more are in a better position to make wise choices. In the special case where Doer regards Knower as an epistemic expert because he undergoes a learning experience that reveals the truth-value of some proposition Y, **K** becomes

> **K+** When (a) and (b) hold, if Doer knows that Knower's epistemic state is identical to her own except insofar as he learned Y's truth-value by a reliable process, then Doer should prefer A to act B if she can deduce that Knower will prefer A to B no matter what he learns about Y's truth-value.

To see **K+** in action, consider the following (in my view unassailable) rationale for choosing efficaciously in Twin's Dilemma.[17] Imagine that Row has an identical twin Rosie who is every bit as rational and who stands to gain exactly what Row gains from her decision. Unlike Row, who has no direct knowledge of Column's choices, Rosie is observing Column through a one-way mirror. She is also receiving continuous updates about the state of Row's deliberations, so Rosie knows every relevant fact that Row knows. Moreover, if Column acts before Row completes her deliberations, then Rosie will come to know something Row does not know: C's truth-value. Finally, suppose that Rosie has the power to send Row a signal that tells her that Column has acted, but without indicating which act he has performed. What should Row do if she receives this signal? The obvious answer is that Row should perform the efficacious act $\sim R$ because she can deduce that this is the act that her better-informed, equally rational, identically interested twin favors. Row can deduce this by reasoning as follows: "Either Rosie will have received the good news that Column chose C or Rosie will have received the bad news that he chose $\sim C$. In the first case, Rosie will hope that I put the icing on the cake by choosing $\sim R$. In the second, she will hope that I minimize the damage by choosing $\sim R$. So, whether she learns C or $\sim C$, Rosie will want me to choose $\sim R$. By **K+**, it follows that I should choose $\sim R$." Of course, Rosie is a merely prop in this little drama. Row can easily see that she would prefer $\sim R$ to R if she were to have a learning experience that revealed C or $\sim C$. Since her "better informed self" would want her to act efficaciously, she should act efficaciously too, just as **K+** recommends.

Evidential ratificationists offer a superficially similar rationale for choosing $\sim R$, but in their argument Rosie learns Row's choice rather than Column's action. Imagine that Rosie has somehow come to know either dR or $d\sim R$. If Row knows this then, even though she might not know what her decision will be, she can reason as follows: "Either Rosie knows that I will decide on R or Rosie knows that I will decide on $\sim R$. In the first case, she will have strong evidence for believing C, and will therefore hope I put the icing on the cake by performing $\sim R$ (contrary to my decision). In the second case, she will have strong evidence for believing $\sim C$, and will hope I mitigate the damage by carrying through on my decision to do $\sim R$. So, whether she learns dR or $d\sim R$, Rosie will want me to choose $\sim R$. By **K+** , I should choose $\sim R$."

If we replace Rosie in this scenario by Row's post-decision self, then this becomes an argument for **SP*** on the basis of **K+**. The post-decision Row is a rational agent (we assume) who shares Row's basic desires, and who has come to learn either dR or $d\sim R$

[17] I first heard about this justification from Don Hubin some years ago. He attributed it to J. H. Sobel.

via deliberative processes.[18] Since $\sim R$ is uniformly ratifiable the pre-decision Row can deduce that her post-decision self will prefer it to R whether she learns dR or $d\sim R$. Thus, **K+** entails that the pre-decision Row should also prefer $\sim R$ to R, just as **SR*** says. Thus, **SR*** is justified on the grounds that an agent's choices should be guided by the judgments of her post-decision self – an equally rational, identically interested self who is better informed about her decision. In this way, the ratificationist justification for choosing the efficacious act in Twin's Dilemma comes to rest squarely on **K+**.

Unfortunately for friends of Ratificationism, **K+** can fail when the agent is in a position to influence Y's truth-value by her actions. Indeed, **K+** is nothing more than a disguised version of Savage's sure-thing principle. If (a) and (b) hold, and if Knower's epistemic state is identical to Doer's except insofar as Knower has learned Y's truth-value, then Knower's beliefs and preferences are identical to those of Doer conditional on either Y or $\sim Y$. Accordingly, **K+** entails that Doer should unconditionally prefer A to B if she can deduce that she will prefer A to B both if she learns Y and if she learns $\sim Y$. This is just the sure-thing principle! As we have seen, however, the sure-thing principle can fail when the process of learning Y's truth-value involves making choices and taking acts that have undesirable causal consequences. No one should be tempted by the following reasoning: "I can pay $10 for an influenza vaccination or not. At the end of the flu season I will have learned whether or not I contracted the flu. If learn that I did contract it, then I will prefer having $10 in my pocket to having spent it on the vaccine. If I learn that I did not contract it, then I will also prefer having $10 in my pocket to having spent it on the vaccine. So, by **K+**, I should forgo the vaccination since my identically interested, equally rational self, who will know more than I do about my prospects for getting the flu, will prefer that I forgo the vaccination." All decision theorists, be they of causal or evidential persuasion, will see this as a fallacy. The agent's future self will, in all probability, only come to know that she contracted the flu as a direct causal consequence of the fact that her past self did not take a precaution that would have causally inhibited the illness. In cases like this, where the pre-decision self can manipulate the information that the post-decision self learns, **K+** and the sure-thing principle are unsound. So, if the ratificationist rationale for choosing efficaciously in Newcomb problems is to succeed, then **K+**, and Preference Reflection, must somehow be restricted.

What restriction is appropriate? Here, evidentialists and causalists must part ways since consistency requires that each treat **K+** the same way she treats the sure-thing principle. Evidentialists must restrict **K+** to cases where the agent's acts provide no evidence about Y's truth-value, while causalists must restrict it to cases where these acts do not causally influence Y's truth-value. Both correctly prohibit **K+**'s use when the agent has the power to manipulate her post-decision knowledge about Y. In these cases, the post-decision self's extra knowledge is not a useful indicator of the desirability of acts because the content of this knowledge is causally influenced by which action is chosen. Evidentialists are forced to go even farther: they must bar **K+**'s use in any context where the agent can act to provide her future self with evidence about Y, even contexts where she cannot causally influence Y's truth-value. In particular, evidentialists must enjoin the use of **K+** in any decision problem whose acts are (i) evidentially correlated with Y's truth-value, but (ii) do not causally influence Y's truth-value.

[18] We are assuming that the deliberative processes that generate the decision constitute a genuine learning experience by which Row ultimately comes to know dR or $d\sim R$. This is a substantive assumption, but we will not question it here.

In such cases, they must also reject the strong form of ratifiability reasoning based on **SR***.

This makes it impossible for evidentialists to consistently deploy **SR*** in Twin's Dilemma or other Newcomb problems. In any genuine decision, facts about which act is performed are highly correlated with facts about which decision is made because, with high probability, the decision causes the act. Once we know what the agent did, we have strong evidence about what she decided. In Twin's Dilemma, for example, R and $\sim R$ provide strong evidence for dR and $d\sim R$ in just this way: $prob(\mathrm{d}R/R) >> prob(\mathrm{d}R/\sim R)$. Of course, these disparities in probability are due to the fact that decisions *cause* the acts, but appeals to such explicitly causal facts are anathema to evidentialists, who are restricted to the use of the evidential version of **K+**. As we have just seen, however, this version of **K+** explicitly *prohibits* ratifiability reasoning when the propositions that the post-decision self learns are evidentially tied to her acts. Thus, there is no ratificationist argument for the efficacious choice in Twin's Dilemma that appeals exclusively to evidentialist principles.

The causal version of **K+** does sanction ratifiability reasoning in Twin's Dilemma. Since Row's acts do not cause her decisions—it's the other way around—the "backtracking" evidential correlations between the acts and decisions do not prevent her from using her post-decision desires as guides to current decisions. So it is the causal decision theorist, not the evidentialist, who is in a position to employ **SR*** to rationalize the efficacious choice.[19] Evidentialists who seek to do so are subtly begging the question by invoking a strengthened version of the Maxim of Ratifiability that can only be justified within the confines of a causal decision theory.

3 Are Newcomb problems genuine decisions?

As already noted, Jeffrey ultimately rejected ratificationism as a solution to Newcomb problems (albeit for the wrong reasons). In his final word on the topic, (1993) and (2004), he alleged that Newcomb problems are not really *decisions* at all because those who face them know too much about their own actions to see themselves as making free choices. In any genuine decision an agent must be able to see her acts as *causes* of outcomes. This cannot happen in Newcomb problems, Jeffrey maintained, because in any *plausible* Newcomb problem the agent will be able to explain the correlation between acts and states in terms of the existence of a "deep state" that serves as a common cause of both. In Twin's Dilemma, the picture is precisely as given in Fig. 1 with the variable *CC* serving as the "deep state." This is, at best, a "quasi decision problem" because, in the presence of the deep state, *the agent cannot regard her acts as causes of outcomes*. Newcomb problems thus do not refute evidential decision theory, which is only meant to apply to "real" decisions in which an agent sees her acts as potential causes of outcomes.

In making his argument Jeffrey appeals to a specific proposal, due to Arntzenius (1992), about what it means for an agent to see one event as a cause of another. To say what it is for an agent to see her actions as causes of outcomes we are *not* required, Jeffrey thinks, to represent her as having explicitly causal beliefs. We need only suppose that her (non-causal) beliefs about her own acts evolve in a specific way during

[19] Note, however, that the causal decision theorist does not need to introduce ratifiability considerations to justify $\sim R$ since it is the act that maximizes causal expected utility.

deliberation. "Imputations of causal influence," he writes, "are not shown simply by momentary features of probabilistic states of mind, but by intended or expected features of their evolution." (1993, p. 139) The picture is one in which the decision maker's subjective probabilities and expected utilities change as a result of deliberation. We can think of our agent's mental state at time t as being represented by a probability $prob_t$ and an expected utility des_t that jointly obey **EDT**. During deliberation, the agent's mental state changes in accordance with a belief/desire revision process that maps an initial probability/utility pair ($prob_0$, des_0) through a sequence of temporal stages ($prob_t$, des_t), $0 \leqslant t \leqslant 1$, to a final state ($prob_1$, des_1). At $t = 0$ the agent is undecided about what she ought to do; at $t = 1$ she has made up her mind. To focus on essentials, let us assume that the process of revision is purely epistemic, so that the agent's *basic* desires remain fixed and all changes in ($prob_t$, des_t) are ultimately traceable to changes in $prob_t$, so that ($prob_t$, des_t) = ($prob_t$, des_0).

Within this framework Jeffrey's proposes to analyze "imputations of causation" using

ARNTZENIUS'S TEST: During the time interval $0 \leqslant t \leqslant 1$, an agent regards one event C as a *promoting (or inhibiting) cause* of another event E only if

- *Correlation*. $prob_0(E/C) - prob_0(E/\sim C) > 0$ (or < 0).
- *Rigidity*. $prob_t(E/C)$ and $prob_t(E/\sim C)$ remain fixed as $prob_t(C)$ varies.
- *Variability*. $prob_0(C)$ starts out at some intermediate value, and the belief revision process takes $prob_1(C)$ either to zero or to one.

Correlation ensures that a promoting cause always provides evidence in favor of its effect. Rigidity requires the belief revision process that alters C's probability to behave like a *Jeffrey shift* on $\{C, \sim C\}$ with respect to the propositions in the set $\{C\ \&\ E, C\ \&\ \sim E, \sim C\ \&\ E, \sim C\ \&\ \sim E\}$. Variability requires that C's probability change over time. Rigidity then forces E's probability to change over time as well. The manner in which these changes occur is supposed to determine "imputations of causal influence."

We can start to see why by noting that Arntzenius's Test introduces a partial asymmetry between causes and their effects. As long as the agent's probabilities for C and E change over time, as required by Variability, then Rigidity holds only if $\text{prob}_t(C/E)$ and $prob_t(C/\sim E)$ vary with $prob_t(E)$. This is the result of the following

FACT: If $prob_t(E/C)$, $prob_t(E/\sim C)$, and *either* $prob_t(C/E)$ or $prob_t(C/\sim E)$ remain fixed over time, then so must $prob_t(C)$ and $prob_t(E)$ because

$$prob_t(C) = \frac{prob_t(C/E)prob_t(E/\sim C)}{prob_t(\sim C/E)Prob_t(E/C) + Prob_t(C/E)Prob_t(E/\sim C)}$$

$$Prob_t(E) = Prob_t(C)Prob_t(E/C) + prob_t(C)Prob_t(E/\sim C)$$

In other words, if you fix any three of the four conditional probabilities, then you also fix the unconditional probabilities of E and C as well. Thus, if the conditions of Arntzenius's Test are satisfied, then it is impossible to both regard C as a cause of E and to regard E as a cause of C.

While this does introduce a kind of asymmetry between causes and effects, it does not fully distinguish them since Arntzenius's Test can hold relative to one sort of

belief revision process with C and E as they are, and yet hold relative to another belief revision process with C and E reversed. For example, any time the agent has an *E-learning experience* modeled by a Jeffrey shift

$$prob_t(\bullet) = prob_t(E)prob_t(\bullet/E) + prob_t(\sim E)prob_t(\bullet/\sim E)$$

the conditional probabilities $prob_t(C/E)$ and $prob_t(C/\sim E)$ remain fixed as $prob_t(E)$, $prob_t(C)$, $prob_t(E/C)$ and $prob_t(E/\sim C)$ all vary. Suppose, for example, that I know that C, an excess of bilirubin in the blood, causes E, a characteristic yellow hue of the skin (when viewed in sunlight). When looking at you in poor light I might have a learning experience in which my level of confidence in E increases from 0.2 to 0.4 while my probabilities for you having excess bilirubin conditional on having or not having yellow skin remain the same. If the light slowly improves and I go through a series of such Jeffrey shifts until I become certain that you have yellow skin, then my subjective probabilities satisfy the conditions of Arntzenius's Test with E in for the putative cause and C in for the putative effect.

This is no refutation of the Arntzenius's Test, which is only intended to be a necessary condition anyhow, but it does show that one needs to be careful when applying it. The test can only reliably distinguish causes from effects when it is restricted to specific sorts of belief revisions. In the bilirubin case the changes in probabilities for C and E were driven by learning experiences involving their *effects*, specifically the effects of E. Arntzenius's Test works best when the changes in probabilities for C and E are driven by learning experiences involving their causes.

I do not think that Jeffrey should find anything objectionable in this reading. Indeed, he is quite clear about the fact that in decision-making contexts there is a "driving mechanism."

> In decision-making it is deliberation, not observation, that changes your probabilities. To think that you face a decision problem rather than a question of fact about the rest of nature is to expect whatever changes arise in your probabilities... during your deliberations to stem from changes in your probabilities for choosing options.... As a decision maker you regard probabilities of options as inputs driving the mechanism, not driven by it. (Jeffrey, 1993, p. 8)

Jeffrey's claim, then, is that Arntzenius's Test tells us when the agent genuinely regards her acts as causes of outcomes in cases where changes in her subjective probabilities for actions are directly induced by processes of *rational deliberation*. Thus, we have

> Jeffrey's Test: During the time interval $0 \leqslant t \leqslant 1$, an agent regards an act A as a *promoting (or inhibiting) cause* of an outcome $A \ \& \ E$ only if
>
> - *Correlation*. $prob_0(E/A) - prob_0(E/\sim A) > 0$ (or < 0).
> - *Rigidity*. $prob_t(E/A)$ and $prob_t(E/\sim A)$ remain fixed as $prob_t(A)$ varies.
> - *Variability*. $prob_0(A)$ starts out at some intermediate value, and is driven by the process of rational deliberation to a state in which $prob_1(A)$ is either to zero or to one.

Since an agent facing a genuine decision must see her acts as causes of outcomes, Jeffrey requires that the agent's subjective probabilities satisfy this test for each act A and state E in the decision under consideration.

It is crucial to Jeffrey's view that the deliberative process can alter an agent's act probabilities. Her initial state of indecision is modeled by a time-0 probability that does not assign any act a probability close to one. As a result of deliberation, this indecisive state is eventually replaced by a time-1 state in which the probability of the chosen act is one or nearly one. In effect, deliberation is a process by which the agent learns what she will ultimately do. Skyrms (1990) provides a formal model of deliberation in which the agent iteratively revises her beliefs in light of information about the expected utilities of acts, and then recalibrates these utilities in light of her revised beliefs. She uses her $t = 0$ probabilities to calculate expected utilities for A, $\sim A$, and the "status quo" $A \vee \sim A$, and raises her degree of confidence in any act whose expected utility exceeds that of the status quo. The procedure is repeated, using the revised probabilities for A and $\sim A$ as inputs, until some stable equilibrium is reached. Changes in act probabilities *seek the good* in the sense that $prob_{t+1}(A) > prob_t(A)$ whenever the time-t expected utility of A exceeds that of $A \vee \sim A$.[20] The character of these direct changes in probability will, of course, depend on how expected utilities are computed. For Jeffrey, A's probability increases between t and $t+1$ just in case A is good news at t. For a causal decision theorist, such an increase will occur only if A is judged to be more effective than $\sim A$ at causing desirable outcomes.

No matter what sort of expected utility calculation is driving these changes in act probabilities, they ramify through the agent's beliefs via Jeffrey conditioning, so that $prob_{t+1}(\bullet) = prob_{t+1}(A)prob_t(\bullet/A) + prob_{t+1}(\sim A)prob_t(\bullet/\sim A)$.[21] Rigidity is then automatically satisfied. This has a number of implications. First, any correlations that exist between states and acts are preserved, so that $prob_t(E/A) = prob_0(E/A)$ for all states E, acts A, and times t. Second, since we are assuming that the utilities of outcomes do not change over time (though the utilities of other propositions might), the news values A and $\sim A$ remain *fixed* at their time zero values as t varies. Thus, deliberation, in the sense captured by the model, is *not* a process by which an agent revises her views about news values of actions: she merely revises her views about acts' *probabilities* while keeping their news values fixed. These changes induce modifications in her estimates of the news value of the "status quo", $V_t(A \vee \sim A)$, which varies until $t = 1$, at which point it coincides with $V_0(A)$ or $V_0(\sim A)$ depending upon whether the former or latter is greater. The agent will then have made up her mind: she will have decided to perform the act that maximizes her *time-0* news value, and will have come to assign this act a probability near one. The upshot is that in any *genuine* decision in which changes in act probabilities are driven by considerations about news values a rational agent will settle on the act sanctioned by evidential decision theory.

[20] Skyrms and others have discussed a number of belief revision rules with this property. One is *Nash* rule, which, in the case of evidential decision theory, sets $prob_{t+1}(A) = (prob_t(A) + k)/(1 + k)$ where $k = (V_t(A) - V_t(A \vee \sim A)) > 0$. Another is the *Darwin* rule, which sets $prob_{t+1}(A) = k\, prob_t(A)/V_t(A \vee \sim A)$.

[21] Admittedly, this is only (an idealization of) a small part of the deliberative process. When speaking of deliberation we often have in mind the whole raft of psychological processes by which agents come to arrive at decisions. Parts of this process involve the decision maker coming to have settled views about desirabilities of outcomes and probabilities of states of the world. Deliberation, in this broader sense, thus includes what fixes, or makes us aware of, the utilities of outcomes and the probabilities of states of the world that are relevant to our choices. This is not what the Jeffrey/Skyrms model is meant to capture. It presents us with an account of only the *end* of the deliberative process by explaining how an agent who already possesses fixed views about the desirabilities of outcomes and the probabilities of states is able to move from being undecided about her best option to having chosen an action.

It follows that Newcomb Problems cannot be genuine decisions. These problems require the agent to know a great deal about the conditional probabilities of her acts given the world's state, e.g., Row must knows that she is likely to do R if Column does C. If Row knows enough to fix sharp values for $prob_0(R/C)$ and $prob_0(R/\sim C)$, and if these values are, as Jeffrey claims, "fixed once given" (1993, p. 141), then the **Fact** mentioned above entails that she cannot satisfy both Rigidity and Variability. If Rigidity holds, so that $prob_t(C/R)$ and $prob_t(C/\sim R)$ remains fixed for all t, then $prob_t(A) = prob_0(A)$. In a symmetric Twin's Dilemma, for example, Row's initial subjective probabilities might be $prob_0(C/R) = prob_0(R/C) = 0.9$ and $prob_0(C/\sim R) = prob_0(R/\sim C) = 0.3$, which mandates $prob_0(A) = prob_0(C) = 0.875$. Thus, enforcing Rigidity in Twin's Dilemma entails that Row's act probabilities cannot vary during her deliberations; her beliefs about her own acts are effectively hemmed in by the evidence that she has about the correlations between her acts and states of the world. According to Jeffrey, this shows that Row is not facing a real decision. Her choice is not free because her act probabilities are not free to vary; she knows too much about the correlations between her actions and states of the world to see her acts as causes of outcomes.

If this is right, then evidential decision theory has nothing to fear. According to Jeffrey's test, in any bona fide decision problem both probabilities of outcomes conditional on acts and news values of acts must remain constant throughout deliberation. Accordingly, an agent can confidently use an act's news value as a "figure of its merit" (1993, p. 140) because she can be sure that its present news value will coincide with the news value she will accord to it after she deliberates herself into a state of certainty about what she will do. Newcomb problems fail as counterexamples to evidentialism because they are not really decisions at all!

While there is some plausibility to this argument, it contains a subtle flaw. To ferret it out, we need to understand how an agent's *beliefs about her own acts* are related to her decisions. Jeffrey's argument rests on the claim that the conditional probabilities $prob_t(R/C)$ and $prob_t(R/\sim C)$ are "fixed once given" because they are constrained by Row's antecedent evidence concerning the correlations between her actions and those of Column. But many decision theorists (both evidential and causal) have suggested that free agents can legitimately *ignore* evidence about their own acts. Judea Pearl (a causalist) has written that while "evidential decision theory preaches that one should never ignore genuine statistical evidence... [but] actions—by their very definition—render such evidence irrelevant to the decision at hand, for actions change the probabilities that acts normally obey." (2000, p. 109) Pearl took this point to be so important that he rendered it in verse:

> Whatever evidence an act might provide
> On facts that precede the act,
> Should never be used to help one decide
> On whether to choose that same act. (2000, p. 109)

Huw Price (an evidentialist) has expressed similar sentiments: "From the agent's point of view contemplated actions are always considered to be *sui generis*, uncaused by external factors... This amounts to the view that free actions are treated as probabilistically independent of everything except their effects." (1993, p. 261) A view somewhat similar to Price's can be found in Hitchcock (1996).

These claims are basically right: a rational agent, *while in the midst of her deliberations*, is in a position to legitimately ignore any evidence she might possess about what

she is likely to do. She can readjust her probabilities for her currently available acts[22] at will, including her probabilities for acts conditional on states of the world. Row can, for instance, say to herself, "though I have overwhelming evidence for thinking that Column and I are going to make the same choice, I am nevertheless free to believe anything I want about whether or not I am going to do *R* in the event that he does *C*." If this is right, then there is then no conflict between Rigidity, Variability and the idea that Row has enough evidence to fix values for $prob_0(R/C)$ and $prob_0(R/{\sim}C)$. Jeffrey's mistake was to suppose that these latter probabilities are "fixed once given." He thought this, I suspect, because he felt that the strength of the agent's opinions about her own acts should, like her opinions about other matters, be proportioned to her antecedent evidence for them. Pearl, Price, Hitchcock and I all deny this. A deliberating agent who regards herself as free need not proportion her beliefs about her own acts to the antecedent evidence that she has for thinking that she will perform them. Let's call this the *evidential autonomy thesis*.

It is clear that Jeffrey's attempt to exclude Newcomb problems from the realm of decision theory fails if we accept the autonomy thesis. One might, however, wonder why is it legitimate for deliberating agents to ignore evidence about their own actions when it is not legitimate for them to ignore evidence about other matters? While Row may ignore evidence about the causes of her acts, she may not ignore the evidence that her acts provide for their effects. What makes the difference? On the face of things, it seems unreasonable for an agent to ignore evidence for any belief, even a belief about what she will do. So, it seems as if the evidential autonomy thesis runs afoul of the requirement of

> *Proportionism*. An person's level of confidence in a proposition *X* should always be proportioned to her balance of evidence for *X*'s truth, in particular it should be responsive to evidence about the causes of *X*.

To refute Jeffrey one must either explain why a deliberating agent's beliefs about her own currently available acts are not bound by this requirement, or to show that, contrary to appearances, they do not violate it. Neither Pearl, Price nor Hitchcock says as much as one would like on this topic. I shall try to do better by showing how a deliberating agent's beliefs about her own actions can be justified on the basis of the totality of her evidence at the time of her decision.

One initially tempting strategy for doing this has affinities with Jeffrey's ratificationism and with Eells's "tickle defense." The idea would be to argue that, at the moment of choice, Row can ignore all evidence about correlations between her acts and Column's acts because at that point she will have learned her decision, so that either $prob_1(dR) = 1$ or $prob_1(d{\sim}R) = 1$, and this new knowledge will screen off all evidence about prior causes or correlations. Unfortunately, this strategy merely replaces the question of how changes in Row's beliefs about *A* and $\sim A$ can be justified with the question of how changes in her beliefs about dR and $d{\sim}R$ can be justified. Jeffrey will insist that, insofar as Row sees her decisions as causes of her acts, the conditional probabilities $x = prob_t(R/dR) > y = prob_t(R/d{\sim}R)$ will remain constant during deliberation while $prob_t(dR)$ and $prob_t(d{\sim}R)$ vary. But, since $prob_t(dR) = (prob_t(R) - y)/(x - y)$ this prevents $prob_t(dR)$ from varying during deliberation unless $prob_t(R)$ varies

[22] It is important to understand that this freedom only extends to propositions that describe actions about which the agent is currently deliberating, and whose performance she sees as being exclusively a matter of the outcome of her decision. It does not, for example, apply to acts that will be the result of future deliberations.

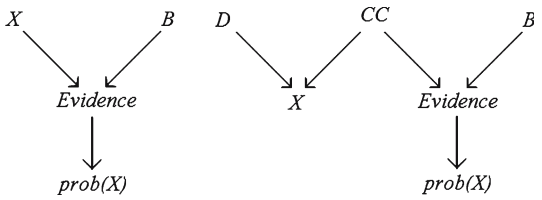


Fig. 4

as well. So, just as in the case of A, we are left with the task of explaining why it is legitimate for Row to change her opinions about dA during deliberation even though she has enough evidence to determinately fix a value for $prob_t(dR)$. The general point is that, insofar as she regards her decision a cause of her act, Row's beliefs about the former will be evidentially tied to her beliefs about the latter. Thus, with respect to autonomy from evidence, Row's beliefs about A and her beliefs about dA are in the same boat.

We can reconcile the autonomy thesis with the requirement that beliefs be proportioned to the evidence by recognizing that an agent's beliefs about her own free decisions and actions provide evidence for their own truth. Normally, the fact that a person believes a proposition X says nothing about X's truth-value that is not indicated by evidence she already possesses. This is because, in normal cases, (a) the person's level of confidence in X is, at least partly, a result of her having the evidence she does, and (b) X's truth or falsity either causes her to have this evidence or is tied to it as joint effects of a common cause. So, normal cases look like this (Fig. 4).

Situations depicted on the left are quite familiar, as when the fact that the sun is shining, together with my standing outside, causes me to see my shadow, and thereby to believe that the sun is shining. Those on the right are common as well. For example, the fact that a coin has a certain physical constitution, combined with my observing it being tossed 100 times, might cause me to have the evidence that eighty of the last 100 tosses have been heads, and thus to be highly confident that the coin will come up heads between 65 and 95 times in the next 100 tosses.

Neither (a) nor (b) hold in the situations that interest us. In these cases there is a causal path running between the belief and the proposition believed that is not mediated by the believer's evidence, and the existence of this path makes it possible for the person to use the belief as evidence for its own truth. There are two relevant cases of this type.[23] In the first, the person's belief about X may not be fully responsive to her evidence for X, e.g., it might also depend on her desire for X's truth. While this typically indicates an epistemic failing, it need not be problematic when the belief in question causally contributes to its own truth. For example, as William James argued, wishful thinking can be legitimate for beliefs that are *self-fulfilling prophecies*, i.e., when the fact that a person believes X to a sufficiently high degree causally promotes X's truth. When there is a "power to positive thinking" in this way, a believer can acquire evidence in favor of a belief solely in virtue of holding it. This added evidence must be balanced off against the evidence she already has, but its general effect will be to justify a higher level of confidence. For example, it might be that those who stutter

[23] A third sort of case, in which X directly causes the agent to believe that X, seems irrelevant to decision theory. If the agent learns that she will perform A as a causal consequence of A performance, then it is hard to interpret A as being the result of a free decision.

have fewer problems speaking when they believe they will not stutter than when they believe they will stutter. If so, then a stutterer who has strong inductive evidence for thinking that he will stutter the next time he speaks might nevertheless be in a position to reasonably believe that he will speak fluently. For, it might be that in the past he has generally believed that he will stutter. So, if he can somehow convince himself that he will not stutter this time, which he may or may not be able to do, then the fact that he has this belief will actually make stuttering less likely. If this "confidence effect" is great enough, then it can be permissible, from a purely epistemic perspective, for the person to believe that he will not stutter. Even thought all the evidence points the other way *before* he comes to hold the belief, the evidence he comes to have *after* adopting the belief might well tip the balance.

Conditions (a) and (b) can also fail when X's truth-value and the belief that X are joint effects of a common cause whose occurrence is only know to the person in virtue of being aware of the fact that she believes X. To illustrate, suppose that, in contrast with the story of the last paragraph, each incident of stuttering is caused by a specific brain event that also causes the stutterer to lose confidence in his ability to speak fluently, a confidence that is otherwise present. The stutterer's belief that he will/will not stutter can then provide strong evidence for itself since its presence is a reliable indicator of the presence/absence of the underlying mental event.

The following diagram depicts the "confidence effect" on the left, and the "hidden common cause" on the right. (The broken arrow indicates a weak or non-existent causal connection (Fig. 5)).

In each scenario, it is reasonable for a believer who knows the causal structure of the situation to invoke her high degree of confidence in X as evidence for X's truth. Her ability to do this does *not* depend on the character of her background evidence, except insofar as this evidence relates to her knowledge of the causal structure. Call beliefs that arise in these situations *self-supporting*.

When the possibility of self-support is taken into account, apparent violations of Proportionism can be benign. If the "confidence effect" is strong enough, then the wishful thinker's belief might be properly proportioned to her total evidence for it, which includes the fact that she holds the belief, even though it is not properly proportioned to the evidence she had before coming to hold the belief. The same thing can happen when a believer recognizes that her high degree of confidence in a proposition indicates the occurrence of a common cause of both it and the truth of the proposition believed.

On the model of deliberation considered here, there is a deep connection between decision-making and self-supporting beliefs. The model has the agent becoming more confident in both A and dA upon learning that she prefers A to the status

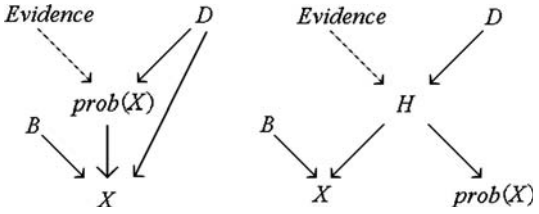


Fig. 5

quo. Why isn't this just wishful thinking? The answer is that it *is* wishful thinking, just not wishful thinking of the fallacious sort because the beliefs involved are self-supporting. There are a number of "metaphysical" accounts of decision-making that make A and/or dA self-supporting. They key to all of them is the idea that there is a causal pathway, not mediated by the agent's evidence, that runs between agent's beliefs about A and dA and her performance of the act A. Rather than trying to decide which of these views is correct, I will simply sketch out some of the possibilities so as to make clear what the options are.

One might hold a view on which an agent's beliefs about her own acts are self-supporting in virtue of being self-fulfilling prophecies. David Velleman defends such a position in his (1989). For our purposes, Velleman's crucial insights are these:

- An agent's beliefs about her own potential acts can, under conditions of deliberation, play a *direct causal role* in her behavior, and so become self-fulfilling prophecies.
- An agent who sees herself as acting freely will recognize that her beliefs about her own actions have this "self-fulfilling" character.
- This creates a kind of "epistemic freedom" that can *make it reasonable* for the agent to adopt beliefs about her own actions that run contrary to other evidence that she might have about what she is likely to do.
- The basis of this epistemic freedom is the fact that by believing that she will do a certain act A an agent *creates the evidence* that justifies this belief.

If all this is true, then an agent's beliefs about her own free actions can be both evidentially autonomous and correctly proportioned to the evidence. For example, even though Row's prior evidence requires that $prob_0(C/R) = prob_0(R/C) = 0.9$ and $prob_0(C/{\sim}R) = prob_0(R/{\sim}C) = 0.3$, and so mandates that $prob_0(R) = prob_0(C) = 0.875$, she can still become justifiably certain of ${\sim}R$ during deliberation because ${\sim}R$'s objective probability grows with increases in her confidence in ${\sim}R$.

Even if one does not agree with Velleman that an agent's beliefs about her own acts directly cause those acts, one might think this is true for beliefs about decisions. One might think, that is, that the direct cause of A is the agent's decision to A, and that the direct cause of her decision to A is her belief that she has decided to A. The agent's belief that she will A will not then be a cause of A, but, since the agent's beliefs about A are directly tied to her beliefs about dA, it will be connected to A by a causal chain that is not mediated by background evidence. The last bit of the causal chain will then look like this: $A \leftarrow dA \leftarrow prob_1(dA) \rightarrow prob_1(A)$. Here, the agent's belief that A is not itself self-fulfilling, but it is justified by a belief that is both self-fulfilling and a cause of A. Again, the requirement of evidential autonomy is met without violating Proportionism.[24]

An alternative way of reconciling evidential autonomy and Proportionism is to maintain that an agent's beliefs about her own acts are reliable indicators of some hidden, immediate common cause of both the act and the belief. On this view, Row's changing degrees of confidence in R and ${\sim}R$ (or dR and $d{\sim}R$) track vacillations in some aspect of her underlying mental state, we might call it her state of indecision, that has direct causal influence over what she will ultimately do. If she eventually becomes certain that she will do ${\sim}R$, this indicates that she has come to rest in a

[24] Even more generally, this will be true if there is a proposition X such that the agent rightly believes that causal relationships are such that $A \leftarrow X \leftarrow prob_1(X) \rightarrow prob_1(A)$.

decisive state that causes $\sim R$. She can then use the fact that she is certain about $\sim R$ as evidence for the conclusion that she is in such a state, thereby justifying that very belief. Again, the belief is autonomous of believer's background evidence, but it does not run afoul of Proportionism because it is self-supporting.

This brief discussion has indicated a few ways in which an agent's beliefs about her own current actions can be self-supporting. For our purposes, it is less important to know which one of these accounts is correct than it is to recognize that the self-supporting nature of act beliefs makes it legitimate for agent to ignore any antecedent evidence she might possess about what she is likely to choose or to do. Appreciating this makes it clear that, contrary to what Jeffrey suggests, the values of $prob_t(R/C)$ and $prob_t(R/\sim C)$ are not "fixed once given." No matter how much evidence ROW might have for the correlation between her acts and COLUMN's acts, if she regards herself as entirely free in the matter of R, then *her own beliefs about the causal structure of her decision situation* put in her in a position to legitimately disregard this evidence. Jeffrey's attempt to disqualify Newcomb problems thus fails. If there is to be a successful argument for the conclusion that these problems are not real decisions it will have but made on grounds other than epistemic ones. The beliefs of Newcomb deciders are *not* constrained by the evidence at their disposal; in the context of deliberation, free agents can believe what they want about their current acts because such beliefs provide their own justification.

References

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

Arntzenius, F. (1992). The common cause principle. *Proceedings of the 1992 PSA Conference*, Vol. 2, 227–237.

Cartwright, N. (1979). Causal laws and effective strategies. *Nous, 13*, 419–437.

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

Eells, E. (2000). *Review: The foundations of causal decision theory*, by James M. Joyce. *The British Journal for the Philosophy of Science, 51*, 893–900.

Gaifman, H. (1988). A Theory of higher order probabilities. In B. Skyrms, & W. Harper (Eds.), *Causation, chance, and credence.* Dordrecht: Kluwer.

Gibbard, A. & Harper, W. (1978). Counterfactuals and two kinds of expected utility. In C. Hooker, J. Leach, & E. McClennen (Eds.), *Foundations and applications of decision theory* (pp. 125–162). Dordrecht: Reidel.

Hitchcock, C. (1996). Causal decision theory and decision-theoretic causation. *Nous, 30*, 508–526.

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

Jeffrey, R. (1993). Causality and the logic of decision. *Philosophical Topics, 21*, 139–151.

Jeffrey, R. (2004). *Subjective probability the real thing*. Cambridge, UK: Cambridge University Press.

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

Lewis, D. (1981). Causal decision theory. *Australasian Journal of Philosophy, 59*, 5–30.

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

Price, H. (1993). The direction of causation: Ramsey's ultimate contingency. In D. Hull, M. Forbes, & K. Okruhlik (Eds.), *PSA 1992*, Vol. 2, pp. 253–267. East Lansing: Philosophy of Science Association.

Reichenbach, H. (1956). *The direction of time*. Berkeley: University of Los Angeles Press.

Savage, L. J. (1972). *The Foundations of Statistics*. New York: Dover.

Spirtes, P. Glymour, C. & Scheines, R. (2000). *Causation, prediction, and search*, 2nd edn. Cambridge, MA: MIT Press.

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

Skyrms, B. (1990). *The Dynamics of Rational Deliberation*. Cambridge: Harvard University Press.

Skyrms, B. (1987). The value of knowledge. In C. Wade Savage (Ed.), *Justification, discovery, and the evolution of scientific theories*. Minneapolis: University of Minnesota Press.
Sobel, J. H., (1985). Circumstances and dominance in a causal decision theory. *Synthese, 62*, 167–202.
van Fraassen, B. (1984). Belief and the will. *Journal of Philosophy, 81*, 235–256.
Velleman, J. D. (1989). Epistemic freedom. *Pacific Philosophical Quarterly, 70*, 73–97.