Synthese (2021) 199:5125–5143
https://doi.org/10.1007/s11229-020-03018-y

Newcomb's problem isn't a choice dilemma

Zhanglyu Li[1] · Frank Zenker[2]

Received: 28 July 2020 / Accepted: 28 December 2020 / Published online: 18 March 2021

Abstract

Newcomb's problem involves a decision-maker faced with a choice and a predictor forecasting this choice. The agents' interaction seems to generate a choice dilemma once the decision-maker seeks to apply two basic principles of rational choice theory (RCT): maximize expected utility (MEU); adopt the dominant strategy (ADS). We review unsuccessful attempts at pacifying the dilemma by excluding Newcomb's problem as an RCT-application, by restricting MEU and ADS, and by allowing for backward causation. A probability approach shows that Newcomb's original problem-formulation lacks causal information. This makes it impossible to specify the probability structure of Newcomb's univocally. Once missing information is added, Newcomb's problem and RCT re-align, thus explaining Newcomb's problem as a seeming dilemma. Building on Wolpert and Benford (Synthese 190(9):1637–1646, 2013), we supply additional details and offer a crucial correction to their formal proof.

Keywords Newcomb's problem · Decision-making · Bayesian network · Causality · Joint probability · Probability structure

> *"To judge what one must do to obtain a good or avoid an evil, it is necessary to consider not only the good and the evil in itself, but also the probability that it happens or does not happen; and to view geometrically the proportion that all these things have together." (Arnauld and Nicole, 1662 [1996], The Port-Royal Logic).*

✉ Frank Zenker
fzenker@gmail.com

Zhanglyu Li
zhanglvli@swu.edu.cn

[1] Institute of Logic and Intelligence, Southwest University, Chongqing 400715, People's Republic of China

[2] Warsaw University of Technology, Warsaw, Poland

1 Introduction

Running parallel to the developing of mathematical logic in the mid-18th century, the development of a normative decision-making theory that assists humans in choosing rationally traces to Bernoulli (1738[1954]). By the mid-20th century, von Neumann and Morgenstern (1944) and Savage (1954) had developed a probabilistic version of *rational choice theory* (RCT) that is known as *Bayesian decision theory* (BDT). It generally concerns "how an agent ought to choose when faced with some decision problem" (Elliot 2019, p. 755). Given possible alternative actions, BDT provides models that express (combinations of) beliefs and desires in ways that guarantee a *definite* choice. This is known as BDT's normalization property, or simply its normative property.[1]

Among the challenges posed to RCT and BDT, a problem raised first, at some point between 1960 and 1963, by the physicist William Newcomb has counted as a hard case (e.g., Nozick 1969; Gibbard and Harper 1978; Skyrms 1980; Lewis 1981; Jeffrey 1983). Formalized by Nozick (1969), Newcomb's problem (aka Newcomb's paradox or dilemma) invokes an implausible interaction scenario where a rational choice seems impossible. Several scholars have suggested that the problem casts strong doubt on two basic rational choice principles: *maximize expected utility* (MEU-P), and *adopt the dominant strategy* (ADS-P). Brams (1975) observes that the conflict between MEU-P and ADS-P "is [at] the heart of the paradox" (ibid., 599). Indeed, Newcomb's problem is said to provide "[a] useful entry into the *inadequacies* of the current standard theory [of rational choice]" (Nozick 1993, p. 41).

We review several unsuccessful attempts at pacifying Newcomb's dilemma by excluding it as an RCT application, by restricting MEU-P and ADS-P, and by allowing for backward causation (Sects. 2 and 3). A probability approach shows conclusively that Newcomb's original problem-formulation lacks causal information, making it impossible to specify the problem's probability structure univocally (Sects. 4 and 5). If causal information is added, Newcomb's problem and RCT re-align. This shows that Newcomb's problem is a *seeming* choice dilemma. To our best knowledge, only Wolpert and Benford (2013) have pursued a probability approach to Newcomb's problem. We supply additional details and offer a crucial correction to their formal proof.

2 Newcomb's problem as a rational choice dilemma

Suppose a being, let's call it 'the predictor', forecasts your choices with great accuracy. You know that the predictor has *never* predicted your own past choices incorrectly. You also know that the predictor has *often* predicted other peoples'

[1] BDT, as most philosophers today pursue it, is a *normative* theory. How agents in fact behave, whether under natural or in laboratory conditions, is of no immediate relevance. The issue is how agents *ought* to behave in view of risky decision. To this end, BDT's models guide decisions by evaluating possible action-alternatives.

Table 1 *Game theoretic* payoff matrix for Newcomb's problem

	S_1 = *Box 2* holds \$1,000,000	S_2 = *Box 2* holds \$0
A_1 = choose only *Box 2*	1,000,000	0
A_2 = choose *both boxes*	1,001,000	1,000

choices correctly, many of whom are similar to you. Both assumptions together yield a *prima facie* sufficient reason that the predictor will almost certainly forecast your own choice accurately when he offers to play the following game:

The predictor gives you a choice of selecting either one or both of two boxes. The transparent *Box 1* contains 1,000 dollars; the opaque *Box 2* contains either one million dollars or nothing. Which state obtains depends *only* on the predictor's forecast of your choice: if "choose both boxes" is predicted, then the predictor leaves *Box 2* empty; if "choose only *Box 2*" is predicted, then *Box 2* holds one million dollars. What should you do? Should you "one-box" (*Box 2*) or "two-box" (*Box 1* and *Box 2*)?

If your expected utility (*eu*) *is* the expected monetary value (\$), i.e., if $eu(\$n) = n$, then Newcomb's problem has the payoff matrix in Table 1. Here, A_1 and A_2 are your possible choice-actions, and S_1 and S_2 describe the state of each box.

"One-boxers" choose only *Box 2* (e.g., Horgan 1981; Horwich 1985; Price 1986; Spohn 2012), because "two-boxing" provides strong evidence that *Box 2* is empty (stipulated payoff: \$1,000), whereas "one-boxing" provides strong evidence that *Box 2* holds one million dollars. The evidence is strong because, by assumption, the probability is very high that the predictor forecasts your choice correctly. This means your choice and your payoff are highly correlated. The MEU-P thus dictates to choose only *Box 2*.

By contrast, "two-boxers" stress that Newcomb's original problem-formulation states clearly that the predictor forecasts your choice action *before* you choose (Gibbard and Harper 1978; Lewis 1981; Jeffrey 1983; Fischer 1994). The moment you exercise your choice, therefore, neither choosing nor not choosing *Box 1* can change the content of *Box 2*. Hence, choosing *Box 1 in addition* to *Box 2* increases *eu* by \$1,000. This means "two-boxing" dominates "one-boxing." The ADS-P thus dictates to choose both boxes.

Both options arise *equally* in Newcomb's problem: "one-boxing" looks just as rational as "two-boxing." Short of being able to reject the problem in a motivated way, Newcomb's problem therefore appears as a rational choice *dilemma*. The challenge to BDT's normalization property that the problem poses thus targets BDT's normative center. Following formal demonstration, Priest (2002, p. 13) concluded (without irony) that "[w]e may now show that you ought to choose one box, and that you ought to choose both boxes," because BDT predicts that choosing both boxes *is* the most rational choice, and that it *isn't*. This conclusion is unacceptable, of course, yet both lines of reasoning it follows from are valid. Indeed, "[p]sychologically, Newcomb's problem is maddeningly paradoxical. Two deep-seated intuitions come head to head, and both refuse to budge" (Horgan 1981, p. 341).[2]

We first turn to eliminative approaches that argue Newcomb's problem away.

3 Eliminating Newcomb's problem

3.1 Problem rejection

Among those scholars who seek to reject Newcomb's problem as a "goofball case," Lewis (1979, p. 240; *italics added*) observes a tendency to "[…] not have, or […] not rely on, any intuitions about what is rational in goofball cases *so unlike* the decision problems of real life." Rejecting intuition in such contexts, however, incurs significant explanatory costs. For what counts why, and for whom, as a goofball case? A flat out rejection of Newcomb's problem thus appears unprincipled. Incidentally, Jeffrey (1983, 1988, 1993) had initially accepted, yet later dismissed Newcomb's as a well-formed problem (Jeffrey 2004, p. 113).

Principled rejections of Newcomb's problem as a *problem for RCT* cite that "the circumstances which allegedly define Newcomb's problem generate a previously unnoticed regress" (Maitzen and Wilson 2003, p. 152) such that "Newcomb's problem is insoluble because it is ill-formed." Similarly, "[w]hen we do understand [these circumstances] properly we recognize the logical incoherence of the problem and the pointlessness of the choice" (Slezak 2006, p. 295).

Of course, problem rejection is a far less plausible strategy if the alleged regress, or the alleged logical incoherence, are removed from in Newcomb's problem. This is what the Gaifman–Koons paradox does (Gaifman 1983; Koons 1992). In a game that is structurally similar to Newcomb's problem,

> Adam is to play checkers against Adam* for the stake of 100 dollars. (In order to force a determinate outcome, assume that by not losing the game he will win the stake.) Before the game, Adam* tells Adam that he has decided to pay him [a bonus of] 1000 dollars if he, Adam, will behave during the game irrationally. How is Adam to behave? If he tries to 'behave irrationally' by playing a manifestly losing game, then, in view of his knowing that 'irrational behavior' will win him a much bigger sum than the 100 dollars he will lose [i.e., a 1000 dollars bonus minus his 100 dollars stake], his behavior becomes rational. But if, concluding that [given the expected bonus] playing to lose is rational, [so that by playing to lose] he plays to win then, again, this mode of behavior becomes rational if on its basis he stands to win the 1000 dollars. (Gaifman 1983, p. 150)

[2] (1) The intuition that it is wrong to choose *Box 1* and *Box 2*, given that your expected payoff is only $1,000, rather than choosing only *Box 2*, given that your expected payoff is $1 million. (2) The intuition that it is wrong to choose only *Box 2*, because your choice cannot affect its content (i.e., *Box 2*'s content does not dependent causally on your choice).

So, if Adam plays an instance of a manifestly *dominated* strategy, *S*, and thus plays prima facie irrationally, then this nevertheless is rational *qua* Adam knowing that playing *S* will result in a total payoff of $1,000 – $100 = $900 (see Koons 1992, p. 1). The Gaifman–Koons paradox thus removes the surreal nature of the "predictor," itself the target of easy criticism. There is no obvious regress either. The choice dilemma nevertheless seems to persist. Rather than reject Newcomb's problem, another strategy is to *clarify* it.

3.2 Problem clarification

One may study Newcomb's problem as a logical paradox, as a (surreal) construction of possible worlds, as a conflict between RCT-principles, or as some other rationality- concerning matter. In any case, if Newcomb's problem is a *genuine* dilemma, then it must entail some deficiency about RCT and BDT. Starting with Nozick (1969), this possibility has engendered continued reflection about such fundamental questions as: 'what is rationality?', or 'is a normative rational model/theory even possible?' This, in turn, has promoted the constructive development of RCT. The main tasks are to explain how Newcomb's choice dilemma arises and how to resolve it. Two common ways of addressing these tasks are to restrict the RCT-principles (3.2.1) and to specify the predictor's forecasting accuracy (3.2.2). A less common way is to allow for backward causation (3.2.3).

3.2.1 Restriction

Nozick (1969, p. 118; notation adapted) diagnoses the dilemma to arise from a conflict between the rationality principles ADS-P and MEU-P:

> ADS-P: If there is a partition of states of the world relative to which action A_1 weakly dominates action A_2, then the decision-making agent should perform A_1 rather than A_2.
> MEU-P: Among all available actions, the decision-making agent should perform the action that maximizes her expected utility (*eu*).

According to the standard model of calculating *eu*, the acts A_i ($i = 1, 2, \ldots, m - 1, m$) are open to the decision-maker, and the possible states of the world S_j ($j = 1, 2, \ldots, n - 1, n$) are mutually exclusive and jointly exhaustive (Jeffrey 1983 [1965]). Where O_{ij} denotes the outcome of action A_i under state S_j, and for each A_i and each state S_j, if the decision-maker performs action A_i and state S_j obtains, then the expected utility of A_i, $eu(A_i)$, is the product of the conditional probabilities, $P(S_j|A_i)$, and the outcome's utility, $u(O_{ij})$:

$$eu(A_i) = \sum_j P(S_j|A_i)u(O_{ij})$$

In Newcomb's problem, "two-boxing" (A_2) dominates "one-boxing" (A_1) relative to S_1 and S_2, because for any state, the payoff matrix (Table 1) promises an additional \$1,000. (Relative to the same state, the utility for the lower row in Table 1 always exceeds the utility for the upper row.) ADS-P thus dictates "two-boxing." Per the same payoff matrix, however, we find for A_1 and A_2 that $eu(A_1) = 1{,}000{,}000$ and $eu(A_2) = 1{,}000$. So, the *eu*-maximizing action is A_1. Thus, MEU-P instead dictates "one-boxing." This is where the apparent conflict lies.

Nozick (1969) observes (correctly) that,

> [i]f the actions or decisions to do the actions do not affect, help bring about, influence, etc., which state obtains, then whatever the conditional probabilities [are] (so long as they do not indicate an influence), one should perform the dominant action. (Nozick 1969, p. 131f.)

As an application condition for ADS-P, therefore, if the states S_j are *independent* of the acts A_1 and A_2—i.e., if the conditional probabilities are $P(S_j|A_1) = P(S_j|A_2)$—and *if* the dominant action is available, then "one should choose the dominant action and *ignore* the conditional probabilities which do not indicate an influence" (Nozick 1969, 133). In this *causally independent* version of Newcomb's problem, ADS-P and MEU-P obviously are no longer in conflict. (We explore this option more fully in Sect. 4)

Following Nozick's (1969) lead, Gibbard and Harper (1978) proposed a restricted version of ADS-P:

> ADS-P*: If the states S_j are causally independent of the acts A_i, then ADS-P holds.

Unlike Jeffrey ([1965] 1983), Gibbard and Harper (1978) maintain that *eu* for A_i ought to be calculated such that $P(S_j|A_i)$ is a *counterfactual* probability, $P(A_i \Box\!\!\rightarrow S_j)$,[3] rather than a conditional probability. (Lewis (1976) shows that $P(S_j|A_i) = P(A_i \Box\!\!\rightarrow S_j)$ is not a logical truth.) Gibbard and Harper (1978) claim that ADS-P* eliminates the conflict between ADS-P and MEU-P, but state no supporting argument. To explain, when ADS-P* is read as an application condition for ADS-P, then ADS-P* demands that S_j is causally independent of A_i. So, if ADS-P* does *not* hold in Newcomb's problem, then ADS-P does not hold either. Because this leaves MEU-P as the only applicable rational choice principle, a conflict cannot arise in the first place. Alternatively, if ADS-P* does hold—such that $P(A_i \Box\!\!\rightarrow S_j) = P(S_j)$, where i, $j = 1$, 2—then one can set $P(S_1) = x$, wherefore $P(S_2) = (1 - x)$. One now finds that:

$$eu(A_1) = 1,000,000x + 0(1 - x), \text{ whereas}$$

$$eu(A_2) = 1,001,000x + 1,000(1 - x).$$

[3] Gibbard and Harper (1978, p. 125) formalize sentences expressing a counterfactual content of the form "If I were to do *a*, then *c* would happen" as '$a \Box\!\!\rightarrow c$'. This makes '$P(a \Box\!\!\rightarrow c)$ the probabilistic version of the counterfactual conditional '$a \Box\!\!\rightarrow c$'.

Table 2 *Decision theoretic* payoff matrix for Newcomb's problem

	S_1 = the predictor forecasts *accurately*	S_2 = the predictor forecasts *inaccurately*
A_1 = choose only *Box 2*	1,000,000	0
A_2 = choose *both boxes*	1,000	1,001,000

Clearly, since $eu(A_2) > eu(A_1)$, MEU-P now dictates that the rational choice is to choose *both boxes*, just as ADS-P does. The conflict between ADS-P and MEU-P thus disappears.

In similar spirit, Horgan (1981, p. 348f.) distinguishes two versions of ADS-P*:

ADS-Pp: If the states S_j are *probabilistically* independent of the acts A_i, then ADS-P holds.

ADS-Pc: If the states S_j are *counterfactually* independent of the acts A_i, then ADS-P holds.

ADS-Pc is logically stronger than ADS-Pp, because counterfactual independence entails probabilistic independence, but not vice versa. Horgan (1981) determines *eu* from the conditional probability $P(S_j|A_i)$ *á la* Jeffrey (1983), and maintains that "neither ADS-Pp nor ADS-Pc sanctions taking both boxes […]" (Horgan 1981, p. 349; notation adapted). So the dilemma again disappears insofar as Newcomb's problem violates the application conditions of both ADS-Pp and ADS-Pc, because the states S_j ($j = 1, 2$) in Table 1 are neither probabilistically nor counterfactually independent of the acts A_i. A rational choice, therefore, can only rely on MEU-P. Hence, one ought to "two-box."

3.2.2 Forecasting accuracy

Brams (1975) takes the choice dilemma to arise because the payoff matrix (Table 1) is specified inappropriately. If Newcomb's problem is reformulated "as a decision-theoretic rather than as a game-theoretic problem," then "the apparent inconsistency between the two [RCT-]principles disappears" (ibid., 599). In this case, the relevant states are not the contents of each box, but rather the predictor's *forecasting accuracy*. With the predictor modelled as a proper game-player, the revised payoff matrix is given in Table 2 (cf. Table 1):

The conflict between ADS-P and MEU-P is now removed because the dominating action depends *only* on the forecasting accuracy. The boxes' contents play no role. If the forecast is accurate, then "one-boxing" maximizes *eu*; if the forecast is inaccurate, then "two-boxing" maximizes *eu*. Because the predictor's forecasting accuracy is a *counterfactual* probability ranging over $0 < P(A_i \Box\!\!\rightarrow S_j) < 1$, in order to identify the *rational* choice, it suffices to specify a cut-off point, x. If $x > 0.5005$, one ought to "one-box;" if $x < 0.5005$ one ought to "two-box," and if $x = 0.5005$, both actions make no difference to *eu* (Brams 1975, p. 600; Lewis 1979, p. 238f.; Ahmed 2014, p. 113f.).

To see this, it suffices to vary the final decimal. Thus, suppose that $x=0.5006$, and calculate *eu* as follows:

$$eu(A_1) = 1{,}000{,}000 \times 0.5006 + 0 \times 0.4994 = 500{,}600$$

$$eu(A_2) = 1{,}000 \times 0.5006 + 1{,}001{,}000 \times 0.4994 = 500{,}400$$

In this case, $eu(A_1) > eu(A_2)$. By contrast, suppose that $x=0.5004$, and calculate *eu* as:

$$eu(A_1) = 1{,}000{,}000 \times 0.5004 + 0 \times 0.4996 = 500{,}400$$

$$eu(A_2) = 1{,}000 \times 0.5004 + 1{,}001{,}000 \times 0.4996 = 500{,}600$$

In this case, $eu(A_1) < eu(A_2)$.

Of course, the value of the cut-off point x varies with the actions' payoffs. Given the payoffs are $m=1{,}000{,}000$ and $n=1{,}000$, $x=0.5005$ is merely an instance of the schema:

$$x = \frac{m+n}{2m}$$

To derive this schema, let O_{ij} denote the outcome of action A_i under state S_j (e.g., $O_{11}=A_1\&S_1$; see Sect. 3.2.1), and let m be the payoff of O_{11}, let n be the payoff of O_{21}, and let $m+n$ be the payoff of O_{22}, whereas the payoff of O_{12} is fixed as 0. One can now stipulate the *eu*-value of the focal actions as:

$$eu(A_1) = mx + 0(1-x) = mx$$

$$eu(A_2) = nx + (m+n)(1-x) = nx + m + n - (m+n)x$$

For the indifferent case, $eu(A_1)=eu(A_2)$, we have it that $mx=nx+m+n-(m+n)x$, and hence find that $mx=m+n-mx$, wherefore

$$x = \frac{m+n}{2m}.$$

Though the dominating action does depend *only* on the forecasting accuracy, notice that Newcomb's original problem-formulation *fails* to specify the forecasting accuracy numerically. That the predictor has in the past *rarely* forecasted choices inaccurately nevertheless warrants the assumption that the present forecast is almost certainly accurate. This suggests having almost full confidence, i.e., $x \gg 0.5005$. The decision-theoretic payoff matrix (Table 2) thus makes *one-boxing* the best strategy to maximize *eu*.

3.2.3 Backward causation

"One-boxing" thus depends on *past* evidence of the predictor's forecasting accuracy. As the forecast occurs *before* the agent's choice, however, "one-boxing" seems to rely on *backward causation*. Otherwise, how could the choice at time t_{i+1} influence the predictor's forecast at t_i? Recognition of two problem versions—one allowing for backward causation, and one that doesn't—provides another option of explaining why MEU-P and ADS-P seem to be in conflict.[4] Sainsbury (2009) points out that,

> [t]o the extent that we think of the case [Newcomb's problem] as involving backward causation, we are tempted by MEU-P. To the extent that we think of it as excluding backward causation we are tempted by ADS-P. What strikes us as conflicting views of the same case are really [two incompatible] views of different cases. (Sainsbury 2009, p. 75; notation adapted)

Finding "that backward causation in this [human-related] sense is possible," Schmidt (1998, p. 84), concludes that "the player should leave the $1000 on the table." So, the player should "one-box," as above.[5] Slezak (2013, p. 3) observes that "Schmidt relies on an equivocation on the notion of [actual vs. perceived] causation to establish his central claim that backward causation may be involved." Beyond committing a fallacy, moreover, what Schmidt (1998, p. 73) himself admits is a "strange but possible" scientific story relies on crucial assumptions that are very much *unlike* those underlying Newcomb's original problem.

When Dummett (1954) initiated a philosophical debate on backward causation, he merely defended its *metaphysical possibility*. Flew (1954) and Black (1956) quickly argued against this. Of course, backward causation remains counterintuitive. Instead, the temporal succession of cause and effect reflects a "normal" causal understanding. Indeed, most approaches to causality today assume that causes precede their effects in time (Russo 2010). This echoes the Humean view that "[w]e may define a cause to be an object *followed by another*, and where all the objects, similar to the first, are followed by objects similar to the second," which is to say that "[…] if the first object had not been, the second never had existed" (Hume, [1748] 2007, p. 56; *italics added*).

In the 1960s, physicists discussed the possibility of *tachyons*, hypothetical particles exceeding the speed of light. "If such particles existed, *and no observation indicates this*, [then] they would by some observers according to the theory of relativity be seen as if they were going backward in time" (Faye 2019, p. 136; *italics added*). Inspired by backward causation, "some philosophers argue that a perfect predictor implies a time machine, since with such a machine causality is reversed" (Wolpert and Benford 2013, p. 1639). Absent an actual time machine, of course, backward causation is less than a live possibility, wherefore many reject this attempt at

[4] This option implies rejecting that Newcomb's problem-formulation would already rule out backward causation by stipulation. When accepting this stipulation, of course, backward causation becomes a non-starter (Maitzen and Wilson 2003; Slezak 2006).

[5] Locke (1978, p. 18) maintains that backward causation is irrelevant: one ought to choose both boxes anyways. "[E]ven if reverse causation is involved […] the case for Choice Two (take both boxes) remains unaltered." Gallois (1979, p. 49) criticizes this firmly: "Locke gives no good reason for thinking that the possibility of reverse causation is irrelevant to the problem posed by Newcomb; second the argument for taking both boxes is […] misconceived."

clarifying Newcomb's problem. Mellor (1995, pp. 224–229) even rejects backward causation already on *a prior* grounds.

Since backward causation is at least as controversial as the character of Newcomb's problem, whether the former helps clarify the latter is unclear. Decisively, however, backward causation introduces a causal loop into Newcomb's problem. For this reason, Jeffrey (2004, p. 116) surmises that "Newcomb problems are like Escher's famous staircase on which an unbroken ascent takes you back where you started." Viewed from the predictor's perspective, after all, if one accepts backward causation, then the predictor *can in principle* know, and thus forecast, how the agent chooses. (Think of the predictor as accessing a "magic mirror" that displays the agent's future choice at the moment the predictor makes the forecast). Viewed from the agent's perspective, by contrast, agents seeking to maximize *eu cannot* choose freely between "one-boxing" and "two-boxing," because the *only* way of maximizing *eu* is "one-boxing." In this sense, fatalism might seem to loom (see Faye 2018, 2019). At any rate, insofar as allowing for backward causation reduces the agent's choices to a single option, namely "one-boxing," Newcomb's problem would simply ceases to be a genuine *decision* problem. *A fortiori*, it would cease to be choice dilemma.

3.3 Upshot

None of the forgoing attempts at pacifying Newcomb's problem are particularly persuasive. Rejecting Newcomb's problem as an RCT-application is unprincipled. Clarifying the problem by restricting the scope of the two RCT principles ultimately relies on reformulating Newcomb's original problem as a distinct problem. So, what one treats no longer is Newcomb's *original* problem. Finally, allowing for backward causation appears metaphysically dubious. This provides a sufficient reason to look for an alternative resolution of Newcomb's problem in a probability approach.

4 Probability structures

4.1 Causal graphs

The causal relations in Newcomb's original problem are not only formulated imprecisely, they are also mutually intertwined. The use of probabilities makes imprecise causal relations tolerable in principle. Indeed, modelling causal relations as probabilistic relation seems natural to us. As Pearl (2009) points out,

> causal utterances are often used in situations that are plagued with uncertainty. We say, for example, 'reckless driving causes accidents' […] Any theory of causality that aims at accommodating such utterances must therefore be cast in a language that distinguishes various shades of likelihood—namely, the language of probabilities. (Pearl 2009, p. 1)

Table 3 Utility matrix

	$S_1 = b_2$	$S_2 = b_1b_2$
$A_1 = B_2$	1,000,000	0
$A_2 = B_1B_2$	1,001,000	1,000

Table 4 Probability matrix

	$S_1 = b_2$	$S_2 = b_1b_2$
$A_1 = B_2$	$P(B_2, b_2)$	$P(B_2, b_1b_2)$
$A_2 = B_1B_2$	$P(B_1B_2, b_2)$	$P(B_1B_2, b_1b_2)$

An effective tool to probabilistically model causal relations are causal graphs. Bayesian networks represent causal variables and their conditional dependences as directed acyclic graphs, "provid[ing] convenient means of expressing substantive assumptions; to facilitate economical representation of joint probability functions; and to facilitate efficient inferences from observations" (Pearl 2009, p. 13). A graph's nodes represent variables, edges represent the causal relations between variables.

Corresponding to the decision-maker and the predictor agents in Newcomb's problem are two game variables: 'the predictor's forecast' and 'the decision-maker's choice'. Let 'b_1b_2' denote the *prediction*: 'the decision-maker chooses both boxes', and let 'b_2' denote: 'the decision-maker chooses only *Box 2*'. Moreover, let 'B_1B_2' denote the *evidence report*: 'the decision-maker did choose both boxes;" and let 'B_2' denote: 'the decision-maker did choose only *Box 2*.' The utility matrix and the probability matrix for Newcomb's problem then are as in Tables 3 and 4.

According to expected utility theory (von Neumann and Morgenstern 1944), and with 'u' standing for the utility function of the outcomes, '$\mathbb{S}$' for the set of states $\{S_1, S_2\}$, '$\mathbb{A}$' for the set of actions $\{A_1, A_2\}$, and with the variable C (for 'choice') as the element of $\mathbb{A}$ and the variable F (for 'forecast') as the element of $\mathbb{S}$, we have it that C and F jointly determine eu of the decision-maker's choice as follows:

$$eu(C) = \sum [P(C, F) \times u(C, F)] \quad \text{where} \quad C \in \mathbb{A},\ F \in \mathbb{S}$$

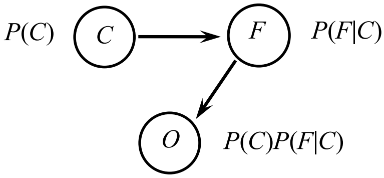

Fig. 1 Bayesian network for "one-boxing"

Given the matrixes in Tables 3 and 4, one calculates *eu*[6] thus:

$$eu(A_1) = 1,000,000\left[P\left(\mathrm{B}_2, \mathrm{b}_2\right)\right] + 0\left[P\left(\mathrm{B}_2, \mathrm{b}_1\mathrm{b}_2\right)\right] \tag{1}$$

$$eu(A_2) = 1,001,000\left[P\left(\mathrm{B}_1\mathrm{B}_2, \mathrm{b}_2\right)\right] + 1,000\left[P\left(\mathrm{B}_1\mathrm{B}_2, \mathrm{b}_1\mathrm{b}_2\right)\right] \tag{1$'$}$$

Determining *eu* in (1) and (1′) thus requires specifying the probability matrix, which then specifies the joint probability *P*(*C*,*F*). In standard probabilistic models, however, *P*(*C*,*F*) lacks a *unique* decomposition, and specific decompositions give rise to different rational choices. We now turn to this.

4.2 Probability structure for "one-boxing"

"One-boxers," as we saw, treat 'choose only *Box 2*' as the sole rational choice. As we also saw, Nozick takes ADS-P to dictate this choice. Analysis of the joint probability shows that the decision-maker would in this case adopt the following decomposition:

$$P(C, F) = P(C)P(F|C) \quad \text{If}\, P(C) = 0, \text{ then}\, P(C)P(F|C) = 0 \tag{2}$$

We can now model the causal relations that (2) expresses as the Bayesian network in Fig. 1 (Li 2017). Variable *C*, the "parent node," has probability *P*(*C*). Variable *F*, the "child node," has probability *P*(*F*|*C*). The arrows between nodes *C* and *F* represents the *dependence* of *F* on *C*. In this way, *P*(*C*) and *P*(*F*|*C*) jointly determine the probability of the decision result, *O*, i.e. *P*(*C)P(F*|*C).*

[6] Contrary to what Wolpert and Benford (2013, Sect. 2) claim, their formula (1) for *eu* breaks with the von Neumann-Morgenstern version of expected utility theory, making the calculation unnatural, too complicated if the forecasting accuracy falls below 100% (i.e., if $p(F|C) \neq 1$), and generally hard to understand. The issue is this: for decision problems whose probability matrix is given, Wolpert & Benford's formula (1) states only the *sum* of the actions' *eu*-values, whereas the *eu*-value for *each individual action* feeding into that sum remains opaque, and so fails to guide the decision. To see this, consider Jeffrey's (1983, p. 8f) "nuclear disarmament" example, where the focal actions "arming" and "disarming" take the *eu*-values: $eu(\text{arming}) = (-100) \times 0.1 + 0 \times 0.9 = -10$; $eu(\text{disarming}) = (-50) \times 0.8 + 50 \times 0.2 = -30$. According to formula (1), we find $(-100) \times 0.1 + 0 \times 0.9 + (-50) \times 0.8 + 50 \times 0.2 = -40$. However, -40 is the sum of the *eu*-values *for both actions*. To a decision-maker seeking to maximize *eu*, therefore, this result is uninformative. By contrast, RCT demands that the agent maximize *eu* as follows: (1) calculate each action's *eu* by using the states' probability and the outcomes' utility relative to the actions; (2) compare all the *eu*s to determine which one is the largest; (3) choose the action with the highest *eu*. Our own proof does exactly this.

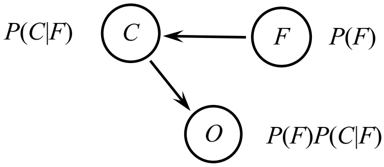

Fig. 2 Bayesian network for "two-boxing"

Here, $P(F|C)$ quantifies the predictor's forecasting accuracy.[7] Crucially, "one-boxers" assume that the value of $P(F|C)$ is large, because the decision-maker's choice is predicted with high accuracy, i.e., a very good prediction algorithm is at hand. The predictor would thus treat $P(F|C)$ as the Kronecker δ function:

$$P(F|C) = \delta_{F,C} \quad \text{if}\, F = C, \text{ then set}\, P(F|C) = 1; \quad \text{otherwise}\, P(F|C) = 0 \tag{3}$$

This means that, at the child node F, given that $P(C) \neq 0$, the predictor can assign the conditional probability $P(F|C)$ *arbitrarily*. For all C such that $P(C) \neq 0$, therefore, the decision-maker cannot affect the value of $P(F|C)$. The one value the decision-maker can control is that of $P(C)$ for node C. Properly understood, then, the decision-maker's discretion in assigning $P(C)$ presupposes the ability to choose freely at node C (cf. our Sect. 3.2.3)

Expressing the probability structure of "one-boxers" as a Bayesian network yields a crucial advantage. One can separate the aspects of the joint probability $P(C,F)$ that the decision-maker determines, on one hand, from the aspects that the predictor determines, on the other. This defines the *decision-maker's strategic space* as $P(C)$, whereas the *predictor's strategic space* is $P(F|C)$. Because the decision-maker can affect only $P(C)$, she can maximize *eu* in (4) or (4′) *only* by assigning an appropriate value to $P(C)$. For the prediction algorithm describing the δ function, moreover, it holds in (4) and (4′) that $P(\mathrm{b}_1\mathrm{b}_2|\mathrm{B}_1\mathrm{B}_2) = 1$ and $P(\mathrm{b}_2|\mathrm{B}_2) = 1$, whereas $P(\mathrm{b}_2|\mathrm{B}_1\mathrm{B}_2) = 0$ and $P(\mathrm{b}_1\mathrm{b}_2|\mathrm{B}_2) = 0$. Given this probability structure, the decision-maker's *eu* now is:

$$\begin{aligned} eu(A_1) &= 1,000,000\left[P\left(\mathrm{B}_2\right)P\left(\mathrm{b}_2|\mathrm{B}_2\right)\right] + 0\left[P\left(\mathrm{B}_2\right)P\left(\mathrm{b}_1\mathrm{b}_2|\mathrm{B}_2\right)\right] \\ &= 1,000,000\left[P\left(\mathrm{B}_2\right)\right] \end{aligned} \tag{4}$$

$$\begin{aligned} eu(A_2) &= 1,001,000\left[P\left(\mathrm{B}_1\mathrm{B}_2\right)P\left(\mathrm{b}_2|\mathrm{B}_1\mathrm{B}_2\right)\right] \\ &\quad + 1,000\left[P\left(\mathrm{B}_1\mathrm{B}_2\right)P\left(\mathrm{b}_1\mathrm{b}_2|\mathrm{B}_1\mathrm{B}_2\right)\right] \\ &= 1,000\left[P\left(\mathrm{B}_1\mathrm{B}_2\right)\right] \end{aligned} \tag{4′}$$

According to the probability axioms, $0 \leq P(\mathrm{B}_2) \leq 1$ and $0 \leq P(\mathrm{B}_1\mathrm{B}_2) \leq 1$. In order to maximize *eu*, the decision-maker should therefore assign $P(\mathrm{B}_1\mathrm{B}_2) = 0$ and $P(\mathrm{B}_2) = 1$; for this maximizes $eu(A_1)$. Hence, she ought to choose A_1, i.e., "one-box" in all cases.

[7] 'C' denotes the decision-maker's choice and '$P(F|C)$' denotes the conditional probability of F given C. If the predictor's forecast is very accuracy, then $P(F|C)$ will be very close to 1.

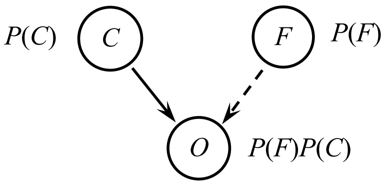

Fig. 3 Revised Bayesian network for "two-boxing"

4.3 Probability structure for "two-boxing"

"Two-boxers" instead base the decision on MEU-P. Analysis of the joint probability shows that "two-boxers" presuppose an *alternative* decomposition of $P(C,F)$:

$$P(C,F) = P(F)P(C|F) \quad \text{if}\, P(F) = 0, \text{ then}\, P(F)P(C|F) = 0 \tag{5}$$

The causal relations corresponding to this decomposition are modelled as the Bayesian network in Fig. 2 (Li 2017). Variable F, the "parent node," has probability $P(F)$, and variable C, the "child node," has probability $P(C|F)$.

"Two-boxers," as we saw, assume that the decision-maker is *ignorant* of the predictor's forecast, and that she exercises her choice *after* the predictor has forecasted it. The decision-maker can therefore assign *any* probability $0 \leq h(C) \leq 1$ at the child node C, i.e., $P(C|F) = h(C)$, where $P(F) \neq 0$. After all, the predictor can assign any probability $P(F)$ to the parent node F, but cannot affect the probability assignment $P(C|F)$. By contrast, the decision-maker can assign any probability $P(C|F)$ at the child node C, but cannot affect the probability assignment of the parent node F. "Two-boxers" therefore presuppose this decomposition of $P(C,F)$:

$$P(C,F) = P(F)P(C|F) = P(F)h(C) \tag{6}$$

Because $0 \leq h(C) \leq 1$, one can set $h(C) = P(C)$, and thus obtain:

$$P(C,F) = P(F)h(C) = P(F)P(C) \tag{7}$$

Insofar as "two-boxers" assume that F and C are *causally independent*, it follows that $P(F)$ directly affects *not* node C, but the decision result, O. Accordingly, one must replace the arrow leading from C to F in Fig. 2 with a new arrow (dashed) leading from F to O, yielding the network in Fig. 3.

Figure 3 shows that "two-boxers" treat Newcomb's problem as a single-stage game, where $P(C)$ is the decision-maker's strategic space and $P(F)$ is the predictor's strategic space. In this probability structure, one calculates *eu* thus:

$$eu(A_1) = 1{,}000{,}000\left[P\left(\mathrm{b}_2\right)P\left(\mathrm{B}_2\right)\right] + 0\left[P\left(\mathrm{b}_1\mathrm{b}_2\right)P\left(\mathrm{B}_2\right)\right] = 1{,}000{,}000\left[P\left(\mathrm{b}_2\right)P\left(\mathrm{B}_2\right)\right] \tag{8}$$

$$eu(A_2) = 1{,}001{,}000\left[P\left(\mathrm{b}_2\right)P\left(\mathrm{B}_1\mathrm{B}_2\right)\right] + 1{,}000\left[P\left(\mathrm{b}_1\mathrm{b}_2\right)P\left(\mathrm{B}_1\mathrm{B}_2\right)\right] \tag{8$'$}$$

Accordingly, no matter which values the predictor has assigned to $P(F)$, i.e. $P(\mathrm{b}_2)$ in (8) and (8′), in order to maximize *eu* the decision-maker ought to set the probability of 'choose only *Box 2*' to 0, i.e., $P(\mathrm{B}_2) = 0$, and set the probability of 'choose *both boxes*' to 1, i.e. $P(\mathrm{B}_1\mathrm{B}_2) = 1$. In all cases, then, "two-boxing" is the rational choice.

5 What dilemma?

A probabilistic approach suggests a specific reason why disagreement about the rational choice in Newcomb's problem persists between "one-boxers" and "two-boxers": each assigns a distinct probability structures, reflecting a distinct causal structure. A "one-boxer" treats the result of the prediction algorithm, $P(F|C)$, as highly accurate, and views the choice, C, to *depend probabilistically* on the forecast, F. A "one-boxer's" decision therefore *must* consider the forecast by conditionalizing on it. By contrast, a "two-boxer" treats the decision-maker's choice as *causally independent* of the forecast, which she can therefore *ignore*.

We can only agree with Wolpert and Benford (2013, p. 1642): "[t]he simple fact that those two decompositions differ is what underlies the resolution of the paradox." Yet, which decomposition—thus which probability structure—one *should* adopt becomes a less pressing question, if one adopts just one structure. Indeed the (near-trivial) point is this: if one implicitly applies *both decompositions at once*, then one cannot expect *one* rational choice. Put more upbeat: if a rational choice seems impossible, check the available decompositions!

To appreciate this more fully, consider the difference in the probabilities that a "two-boxer" and a "one-boxer" assign *arbitrarily*. A "two-boxer" takes the *decision-maker* to assign $P(C|F)$ arbitrarily; a "one-boxer" takes the *predictor* to assign $P(F|C)$ arbitrarily. Each assignment *is* arbitrary, of course. In both cases, these assignments nevertheless affect the decision-maker's probability structure, because "one-boxers" (see line 1, below) and "two-boxers" (see line 2) start from *incompatible* assumptions regarding the joint probability $P(C,F)$.

$$P(C,F) = P(C) \times P(F|C) \quad \left[\text{joint probability decomposition, "one-boxer''}\right] \tag{1}$$

$$P(C,F) = P(F) \times P(C|F) \quad \left[\text{joint probability decomposition, }\}\}\text{two-boxer''}\right] \tag{2}$$

(1) and (2) together generate a contradiction, by lines (5) and (8′), as follows:

$$P(C) \times P(F|C) = P(F) \times P(C|F) \quad \text{from (1) \& (2), by identity} \tag{3}$$

$$P(F|C) = P(F) \textit{ iff } P(C|F) = P(C) \quad \text{from (3), by rearrangement} \tag{4}$$

Because "two-boxers" commit to the choice and the forecast being causally independent, $P(C)$ does not matter to $P(C|F)$, wherefore:

$$P(C|F) = P(C) \quad \left[\text{the choice is causally independent of the forecast}\right], \tag{5}$$

whereas "one-boxers" commit to the forecast predicting the choice very accurately, thus letting $P(F)$ matter to $P(F|C)$:

$$P(F|C) \neq P(F) \quad \left[\text{the forecast is very accurate}\right], \tag{6}$$

and since, by (4),

$$P(C|F) = P(C) \text{ implies } P(F|C) = P(F), \tag{7}$$

it follows, from (6) and (7), by *modus tollens*, that

$$P(C|F) \neq P(C), \tag{8}$$

and it thus follows, from (8), that

$$\neg[P(C|F) = P(C)], \tag{8$'$}$$

so that (5) and (8′) are contradictory. QED.

That "one-boxers" and "two-boxers" make incompatible assumption about the joint probability $P(C,F)$ thus "creates" the dilemma that Newcomb's problem was meant to be. To pacify the dilemma, it suffices to recognize that Newcomb's problem is meaningful *either* as a "one-boxer" *or* as a "two-boxer" understands it. Each understanding entails a distinct decision-making schema with a distinct probability structure. Crucially, each schema leads to a rational choice *in its own right*: "two-boxing" is consistent with ADS-P; "one-boxing" is consistent with MEU-P. The pragmatic fact thus emerges that "one-boxing" and "two-boxing" are incompatible actions that one cannot perform jointly at once.

This has implications for the eliminative approaches to Newcomb's dilemma in Sect. 3, all of which question BDT's normalization property. Similar approaches are reasonable only if one accepts Newcomb's original problem-formulation. Conversely, treating the problem-formulation as underspecified—which it is—*fails* to let a well-defined dilemma arise. We already saw that Newcomb's original problem-formulation lacks causal information needed to fully specify the probability structure.[8] Specifically, the original problem-formulation leaves the details of decomposing the joint probability $P(C, F)$ open. For this very reason, indeed, a probabilistic approach can in the first place *show* that Newcomb's problem arises from conflicting probability structures based on mutually incompatible assumptions regarding $P(C, F)$, themselves motivated by two different understanding of the problem's causal structure.

A resolution of Newcomb's problem thus arises from using the desirably clear language of probability to specify the causal structures the problem admits. In a second step, opting for *either* this *or* that probability structure keeps the decision-maker from interpreting Newcomb's problem as if it described two different games at once. Once specified, Newcomb's problem admits of a *single* rational choice, because BDT "fully specifies the optimal decision for any properly specified single set of

[8] McKay (2004, p. 118) rightly remarks that "the right way to approach the Newcomb problem is to attempt to work out the underlying causal structure," and that "the right choice depends on extra information about the actions." Although authors such as Levi (1975, 1982) and (Eells 1982) have sought to supply this structure, their approaches failed to supply extra information of the right kind (see Slezak 2006). Slezak (2006) himself sees "no grounds for insisting on a plausible causal structure for a science-fiction story" (ibid., 295) such as Newcomb's, thus suspending the need to "wonder about how such a predictor could possibly accomplish his success" (ibid., 283) in the first place.

conditional independencies" (Wolpert and Benford 2013, p. 1640). As the dilemma disappears, the rational choice thus (re-)appears.

Compared to eliminative approaches, the probabilistic approach to Newcomb's problem offers key advantages. Since the need to change the problem-formulation disappears, what one treats in fact *is* Newcomb's problem. This avoids negating the problem itself, and retains the normativity of BDT in guiding rational choices. Moreover, dubious concepts such as backward causation, or hypothetical devices such as a time machine, are not needed, thus respecting Occam's razor. *A fortiori*, appealing to psychological mechanisms in rational decision-making seems entirely misplaced.

6 Conclusion

Having reviewed the main approaches to Newcomb's problem, we saw why it is *not* an RCT-dilemma that questions BDT's normalization property. Specifically, it is false that Newcomb's problem creates a conflict between the decision principles ADS-P and MEU-P. Instead, if a conflict arises, then it arises from assigning two different probability structures. Analyzing these structures showed that causal information is *missing* in the original problem-formulation, and that adding this information results in a RCT-solution to a choice dilemma that Newcomb's problem is emphatically not. Rather, if Newcomb's problem is specified fully, then calculating expected utility suffices to identify the rational choice.

The scholarly value of tackling Newcomb's problems and its variants lies in the continuous development of BDT it has brought about. Yet this development has never deviated from the axiomatic model pioneered by Savage (1954). The rationality assumption, as well as consideration of instrumental rationality, have remained BDT's most important foundations. This continues to reflect Leibniz's dream: "whenever controversies arise, there will be no need of more disputation than what occurs between two philosophers or calculators. It will be sufficient to pick up their pens, sit down at the desks and say to each other: let us calculate" (Leibniz 1688 [2006], p. 266).

Acknowledgements We thank Peter Slezak and a second (anonymous) reviewer for constructive comments that helped improve an earlier version of this manuscript. We also thank audience members at ISRAC 2019, held 12–13 October 2019 at SWUFE, Chengdu, PRC, as well as Dr. Zuojun Xiong, particularly for discussion of the formal proofs. Z.L. acknowledges support from the National Social Science Fund of China (No. 19BZX134). F.Z. acknowledges support from the European Union (COST Action No. 17132).

References

Ahmed, A. (2014). *Evidence, decision and causality*. Cambridge: Cambridge University Press.

Arnauld, A., & Nicole, P. (1996[1662]). *Logic or the Art of Thinking* (Translated by Jill V. Buroker). Cambridge, UK: Cambridge University Press.

Bernoulli, D. (1954[1738]). Exposition of a new theory on the measurement of risk. *Econometrica*, *22*(1), 23–36.

Black, M. (1956). Why cannot an effect precede its cause? *Analysis, 16*(3), 49–58.

Brams, S. J. (1975). Newcomb's problem and prisoners' dilemma. *Journal of Conflict Resolution, 19*(4), 596–612.

Dummett, A. E. (1954). Can an effect precede its cause? *Proceedings of Aristotelian Society (Supplement), 28,* 27–44.

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

Elliot, E. (2019). Normative decision theory. *Analysis, 79*(4), 755–772.

Faye, J. (2018). Backward Causation. *The Stanford Encyclopedia of Philosophy* (Summer 2018 Edition), https://plato.stanford.edu/archives/sum2018/entries/causation-backwards/.

Faye, J. (2019). Backward causation. In R. Poli (Ed.), *Handbook of anticipation* (pp. 121–136). Cham: Springer.

Fischer, J. M. (1994). *The metaphysics of free will*. Oxford: Blackwell.

Flew, A. (1954). Can an effect precede its cause? *Proceedings of the Aristotelian Society (supplement), 28,* 45–62.

Gaifman, H. (1983). Paradoxes of infinity and self-applications, I. *Erkenntnis, 20*(2), 131–155.

Gallois, A. (1979). How not to make a Newcomb choice. *Analysis, 39*(1), 49–53.

Gibbard, A., & Harper, W. (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. 1, pp. 125–162). Dordrecht: Reidel.

Horgan, T. (1981). Counterfactuals and Newcomb's problem. *Journal of Philosophy, 78*(6), 331–356.

Horwich, P. (1985). Decision theory in light of Newcomb's problem. *Philosophy of Science, 52*(3), 431–450.

Hume, D. (2007). *An enquiry concerning human understanding (edited with an introduction and notes by Peter Millican)*. Oxford: Oxford University Press.

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

Jeffrey, R. (1988). How to probabilize a Newcomb problem. In J. Fetzer (Ed.), *Probability and causality* (pp. 241–251). Dordrecht: D. Reidel Publishing Company.

Jeffrey, R. (1993). Causality in the logic of decision. *Philosophical Topics, 21*(1), 139–151.

Jeffrey, R. (2004). *Subjective probability: The real thing*. New York: Cambridge University Press.

Koons, R. C. (1992). *Paradoxes of belief and strategic rationality*. Cambridge: Cambridge University Press.

Leibniz, G. W. (2006[1688]). *The art of controversies (translated and edited, with an introductory essays and notes by Marcelo Dascal)*. Dordrecht: Springer.

Levi, I. (1975). Newcomb's many problems. *Theory and Decision, 6*(2), 161–175.

Levi, I. (1982). A note on newcombmania. *Journal of Philosophy, 79*(6), 337–342.

Lewis, D. (1976). The paradoxes of time travel. *American Philosophical Quarterly, 13*(2), 145–152.

Lewis, D. (1979). Prisoners' dilemma is a Newcomb problem. *Philosophy & Public Affairs, 8*(3), 235–240.

Lewis, D. (1981). Why ain'cha rich. *Nous, 15,* 377–380.

Li, Z. (2017). The solution based on probability structure for Newcomb's problem. *Studies in Dialectics of Nature, 33*(9), 3–8.

Locke, D. (1978). How to make a newcomb choice. *Analysis, 38*(1), 17–23.

Maitzen, S., & Wilson, G. (2003). Newcomb's hidden regress. *Theory and Decision, 54*(2), 151–162.

McKay, P. (2004). Newcomb's problem: The causalists get rich. *Analysis, 64*(2), 187–189.

Mellor, D. H. (1995). *The facts of causation*. London: Routledge.

Nozick, R. (1969). Newcomb's problem and two principles of choice. In N. Rescher (Ed.), *Essays in honor of Carl G. Hempel* (pp. 114–146). Dordrecht: D. Reidel Publishing Company.

Nozick, R. (1993). *The nature of rationality*. New Jersey: Princeton University Press.

Pearl, J. (2009). *Causality: Models, reasoning, and inference* (2nd ed.). Cambridge: Cambridge University Press.

Price, H. (1986). Against causal decision theory. *Synthese, 67*(2), 195–212.

Priest, G. (2002). Rational dilemmas. *Analysis, 62*(1), 11–16.

Russo, F. (2010). *Causality and causal modelling in the social sciences*. Dordrecht: Springer.

Sainsbury, R. (2009). *Paradoxes* (3rd ed.). New York: Cambridge University Press.

Savage, L. J. (1954). *The foundations of statistics*. New York: Wiley.

Schmidt, J. H. (1998). Newcomb's paradox realized with backward causation. *The British Journal for the Philosophy of Science, 49*(1), 67–87.

Skyrms, B. (1980). *Causal necessity: A pragmatic investigation of the necessity of laws*. New Haven: Yale University Press.

Slezak, P. (2006). Demons, deceivers and liars: Newcomb's malin genie. *Theory and Decision, 61*(3), 277–303.

Slezak, P. (2013). Realizing Newcomb's Problem. http://philsci-archive.pitt.edu/9634/. Accessed December 12, 2020.

Spohn, W. (2012). Reversing 30 years of discussion: why causal decision theorists should one-box. *Synthese, 187*(1), 95–122.

von Neumann, J., & Morgenstern, O. (1944). *Theory of games and economic behavior*. Princeton: Princeton University Press.

Wolpert, D., & Benford, G. (2013). The lesson of Newcomb's paradox. *Synthese, 190*(9), 1637–1646.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.