

# A robust resolution of Newcomb's paradox

Thomas A. Weber<sup>1</sup>

Published online: 4 March 2016

© The Author(s) 2016. This article is published with open access at Springerlink.com

**Abstract** Newcomb's problem is viewed as a dynamic game with an agent and a superior being as players. Depending on whether or not a risk-neutral agent's confidence in the superior being, as measured by a subjective probability assigned to the move order, exceeds a threshold or not, one obtains the one-box outcome or the two-box outcome, respectively. The findings are extended to an agent with arbitrary increasing utility, featuring in general two thresholds. All solutions require only minimal assumptions about the being's payoffs, and the being is always sure to predict the agent's choice in equilibrium. The relevant Nash equilibria are subgame-perfect, except for risk-seeking agents where for intermediate beliefs, the being may be unable to ensure perfect prediction without relying on noncredible threats. Lastly, analogies of Newcomb's problem to the commitment problem on a continuum are discussed.

**Keywords** Commitment · Decision theory · Rationality · Time consistency

**JEL Classification** C11 · C72 · D01 · D81 · D83 · G18

## 1 Introduction

Suppose a “superior being” claims to be able to predict an agent's action and, conditional on this prediction, influence his payoffs in a simple game of choice. In this game, there are two boxes, labelled I and II. Box I is transparent and contains a positive monetary reward  $r$ . Box II is opaque and, unbeknownst to the agent, the being puts in

---

✉ Thomas A. Weber  
thomas.weber@epfl.ch

<sup>1</sup> Ecole Polytechnique Fédérale de Lausanne, CDM-ODY 3.01, Station 5,  
1015 Lausanne, Switzerland

it *either* a monetary reward  $R$  (where  $R > r$ ) *or* no reward at all. The reward in box I remains untouched. The agent faces the choice of *either* taking both boxes *or* taking only box II; his payoff consists of the rewards in the boxes he takes. The being claims that box II will contain  $R$  if and only if the agent takes only box II. What should a rational agent do?

This problem, devised by William A. Newcomb in 1960 (Gardner 1973) and published by Nozick (1969), is now referred to as Newcomb's paradox because for both of the agent's feasible actions a priori reasonable justifications have been advanced (see also Gardner 1974). Taking both boxes (the "two-box strategy") seems plausible because of the following dominance argument: no matter what box II contains, adding the reward of box I increases the agent's payoff. Taking only box II (the "one-box strategy") may be plausible because a superior being by simulating the agent's decision process could anticipate the two-box strategy, which would therefore lead to no reward in box II and thus a payoff of  $r$ , which in turn is less than the reward  $R$  the agent could earn by only taking box II. Lastly, to discourage randomization on the part of the agent, the being claims to put no reward in box II whenever the agent's choice emanates from a mixed strategy, i.e., a randomization over his two pure strategies (Nozick 1969, pp. 115/143, endnote 1).

In this paper, we resolve Newcomb's paradox using game theory based on a belief the agent should reasonably form about the structure of the game, specifically about the probability  $p$  that the being is informed about the agent's action before placing the reward. We first consider a risk-neutral agent. Depending on whether the belief probability exceeds a certain threshold or not, the agent will prefer either the one-box strategy or the two-box strategy. If the being can observe this belief, then by anticipating the agent's choice the being can ensure perfect prediction on the equilibrium path, where, perhaps somewhat unexpectedly, an agent who thinks ex-post intervention of the being unlikely, ends up with the low reward  $r$  in the unique Nash equilibrium of the game. If the belief  $p$  belongs to the agent's private information, then the being can devise a mechanism that leads to full information revelation, guarantees perfect prediction, and implements the (first-best) one-box strategy, given the two additional assumptions that the agent's belief is always strictly positive and that the agent believes the being may be able to inflict a negative payoff to discourage deviations which therefore never occur in equilibrium. If the agent is not risk-neutral, his optimal decisions are generally described by two different thresholds, corresponding to whether, in equilibrium,  $R$  ends up in box II or not, should the being move first. Still, the being can solve Newcomb's problem with zero prediction error in equilibrium unless the agent is risk-seeking. Faced with a risk-seeking agent of intermediate belief, the being cannot maintain a zero prediction error in equilibrium without relying on noncredible threats.

## 1.1 Literature

Nozick (1969) formulates Newcomb's paradox and links it to decision theory without offering a resolution; he leans towards the two-box strategy on account of the "dominance principle" which he also defends in later discussions (Nozick 1993, 1997). In

the first account of the problem, rewards are set to  $r = \$1,000$  and  $R = \$1$  million, and these amounts have been largely maintained in the subsequent literature. Bar-Hillel and Margalit (1972) argue for the one-box strategy, stating that the “dominance principle loses its appeal when applied to situations where the states of the world (...) are affected by the decision maker’s actions” (p. 297). Their solution invokes game theory where the being acts in order to maximize the probability of being correct. However, they assume this probability is given exogenously, which makes it generally inconsistent with the proposed equilibrium of the game. Schlesinger (1974) points out that the one-box strategy might be interpreted as backward causation, questioning the possibility of free will, which he considers in turn inherently unpredictable.<sup>1</sup> He then attempts to rationalize the two-box strategy based on the agent’s following a well-meaning observer’s imaginary advice. Benditt and Ross (1976), Locke (1978) and Horgan (1981) give largely intuitive counterarguments. Sorenson (1983) and, more recently, Burgess (2004) defend the two-box strategy by separating in time the agent’s conclusion about the contents of box II and the eventual possibility of revising his decision to also take box I. Horwich (1985) argues for the “evidential principle” and thus the one-box strategy because “the choiceworthiness of an act depends solely upon its likelihood of being associated with desirable events” (p. 432), so the agent should maximize expected utility rather than following “causal decision theory” which relies on the additional assumption that his “act might be a *cause* of the desired outcome” (ibid.). The counterargument by Sobel (1988) invokes the possible endogeneity of the probability that the being predicts the agent’s action correctly, for the agent a priori cannot “be sure that he *cannot* falsify the prediction whatever it is” (p. 20).

In the early discussions of Newcomb’s paradox, the being’s accuracy in predicting the agent’s choice is justified by past observations of this predictive performance. Schlesinger (1974) argues that no amount of inductive evidence is able to increase the agent’s confidence in the being’s prediction performance. Brams (1975) discusses some analogies to the prisoner’s dilemma game,<sup>2</sup> which are again predicated upon assuming an exogenous probability of the being’s correctly predicting the agent’s choice which generally turns out to be inconsistent with the pure-strategy

<sup>1</sup> Simulating the agent’s full cognitive processes, as proposed by Schmidt (1998) in a realization of Newcomb’s paradox with “backward causation,” would amount to an artificial consciousness; see, e.g., Reggia (2013) for a recent survey. Libet et al. (1983) perform experiments that show that the conscious execution of a simple voluntary act is preceded (with 150–800 ms advance) by unconscious (nonrecallable) brain activity that increases a “readiness potential.” Langan (2002) notes that such evidence is not sufficient to decide on the predictability of voluntary actions, as the observations likely involved mere subtasks of a consciously decided supertask.

<sup>2</sup> Lewis (1979) obtains a different prisoner’s dilemma where the being is a “replica” of the agent. Sorenson (1985) notes any finite repetition of this game still produces a two-box strategy in which both players end up with a low payoff. Cooperative solutions [also proposed by Hurley (1994) as a result of a “collective illusion of influence”] after infinite or an uncertain number of repetitions are justified by the folk theorem (Fudenberg and Maskin 1986; Schmidt-Petri 2004); our solution produces a resolution of the paradox without the need for any repetition and without the somewhat strained assumption of the superior being’s incarnation as a mere replica of the agent.

equilibrium.<sup>3</sup> The viewpoint in this paper is different, in that the being can in the end perfectly forecast the agent's choice, simply based on a robust assumption of common knowledge about the game that is being played (Lewis (1969); Aumann 1976; Brandenburger 1992). No performance data on repeated versions of the game or alternative justification of the superior being's predictive abilities are required other than the ability to determine the equilibrium of a game, which is then realized based on rational expectations (Muth 1961) on the part of the agent and the being.

## 1.2 Applications

Frydman et al. (1982) and Broome (1989) point out that the problem of a government committing to a policy which tries to implement measures contingent on agents' actions taken under the policy may be akin to Newcomb's problem. Indeed, when the agents anticipate the implemented policy measures, their actions may tend to undo the effect of these measures. Broome (1989) explains this using the example of monetary policy, where expanding the money supply leads to increased employment (ranked best) as long as this measure is unexpected, and otherwise to inflation (ranked third-best). If, on the other hand, the government does not expand the money supply, then provided an expansion is expected, it leads to a recession (ranked worst) and otherwise to no change at all (ranked second-best). As a result, assuming rational expectations on the part of the agents, the best the government can do is to obtain inflation, corresponding to the third-best outcome. As Frydman et al. (1982) point out, this dilemma relates to a problem discussed in a seminal paper by Kydland and Prescott (1977) where a government seeks to find an optimal dynamic policy. Because it cannot commit to its future actions, any time-consistent policy [which it has an incentive to adhere to at any future time; see Weber (2011), p. 191] is suboptimal compared to what could be achieved with full commitment, when agents believe that the initially announced policy remains unchanged.

Sugden (1991) points out that in dynamic games actions can be time-inconsistent, even though commitment may at the outset be in a player's best interest. He compares Newcomb's paradox to the "toxin" puzzle, where a player is promised a large reward based on his current intention to drink a nonlethal toxin after having received the reward. More importantly, Sugden (1991) also notes that in games, other players' actions are generally endogenous and cannot, in principle, be viewed as mere lotteries. Hence, a player's strategic choice is not reduced to a mere decision problem in the tradition of Savage (1954). For Newcomb's problem, we show that because of rational expectations the probability of the being's making correct predictions is endogenously determined in equilibrium and must be 100 % (except for one special case), provided

<sup>3</sup> Using a Bayesian updating argument in the decision-theoretic setting, Rapoport (1975) casts a first light on the problems associated with assuming an exogenous probability that the being makes correct predictions conditional on the agent's actions. Cargile (1975), Levi (1975) and Eells (1984) provide related arguments that the probability of a correct prediction must in principle depend on the agent's choice, and therefore is endogenous to the problem.

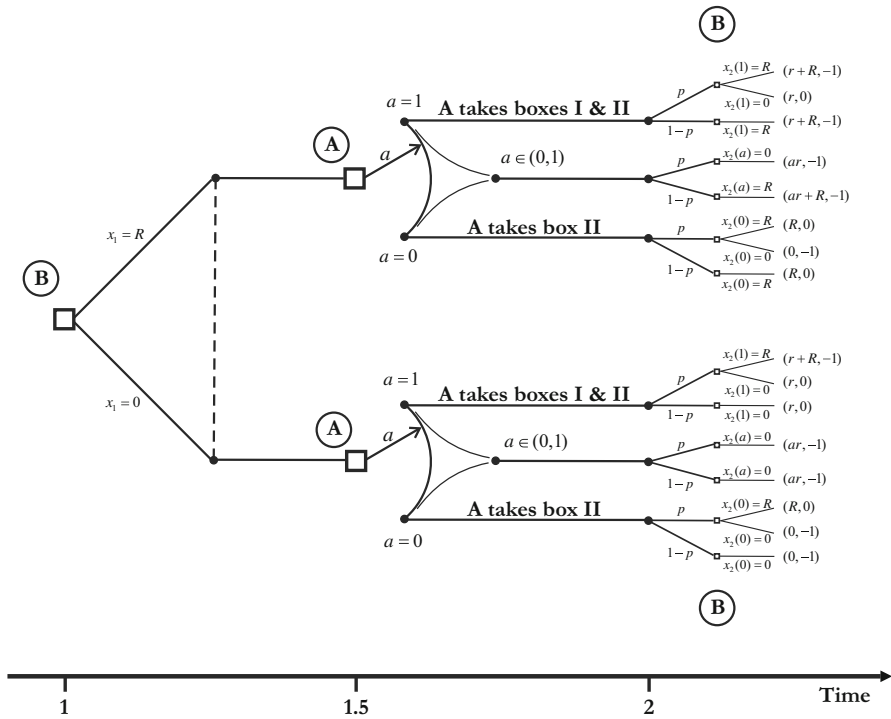


Fig. 1 Extensive-form representation of the dynamic game  $\mathcal{G}$ , for  $p \in [0, 1]$

the being, in the absence of other pursuits, has just the slightest interest in making correct predictions.

## 2 Results for a risk-neutral agent

A risk-neutral agent  $A$  and a superior being  $B$  play a dynamic prediction game  $\mathcal{G}$  over the times  $t \in \{1, 1.5, 2\}$ .<sup>4</sup> There are two boxes, I and II, where box I is transparent and box II is opaque. At the end of the game, box I contains an amount  $r > 0$  and box II contains an amount in  $\mathcal{X} = \{0, R\}$ , with  $R > r$ , which is under  $B$ 's control. An extensive-form representation of  $\mathcal{G}$  is given in Fig. 1; some equivalent implementations of this game are discussed in Remark 2.

- At time  $t = 1$ , the being chooses an amount  $x_1 \in \mathcal{X}$  to be placed in box II, and tells the agent that  $x_1 = R$  if and only if  $A$ 's choice (at time  $t = 1.5$ ) will be to take only box II, with probability 1.

<sup>4</sup> The somewhat unorthodox choice of the time periods, which includes  $t = 1.5$  (instead of using times in  $\{0, 1, 2\}$  or  $\{1, 2, 3\}$ ), is motivated by allowing for the possibility to collapse the game  $\mathcal{G}$  into two periods [see Remark 2(i)] and also by the notational convenience that  $B$ 's first and second intervention happen at times  $t = 1$  and  $t = 2$ , respectively.

- At time  $t = 1.5$ , without having observed  $B$ 's action in the previous period, the agent decides to *either* take both boxes *or* to take box II alone. Allowing for randomization between the two possibilities, let  $a \in \mathcal{A} = [0, 1]$  denote the agent's chosen probability with which to take both boxes. Thus, with probability  $a$  the agent ends up with box I in addition to box II. When pondering his decision the agent believes that with probability  $p \in [0, 1]$ , the being can in the next period observe  $A$ 's strategy  $a$  and revise the amount placed in box II.
- At time  $t = 2$ , with probability  $p$  (according to  $A$ 's beliefs)  $B$  replaces  $x_1$  by an amount  $x_2(a) \in \mathcal{X}$  in box II, contingent on  $A$ 's strategy  $a$ . Otherwise, with probability  $1 - p$ , the amount  $x_1$  remains in box II. Immediately thereafter, both players' payoffs are realized.  $A$ 's payoff consists of the combined contents of the boxes he has taken at  $t = 1.5$ .  $B$ 's payoff is negative (say, equal to  $-1$ ) if the chosen actions are inconsistent with the statement at  $t = 1$  and zero otherwise. That is,  $B$ 's payoff is negative if  $x_2(a) \neq \mathbf{1}_{\{a=0\}}R$ , and it is zero otherwise.

To summarize,  $A$ 's strategy consists of the number  $a \in \mathcal{A}$  which describes the probability with which he ends up with box I in addition to box II.  $B$ 's strategy consists of the tuple  $x = (x_1, x_2(\cdot))$ , where  $x_1 \in \mathcal{X}$  is a number in  $\{0, R\}$  and  $x_2 : \mathcal{A} \rightarrow \mathcal{X}$  is a function. The agent's belief  $p$  is common knowledge.<sup>5</sup> At the end of the game,  $A$ 's payoff is the combined contents of all boxes, and  $B$ 's payoff is negative whenever  $x_1 \neq a$  and zero otherwise.<sup>6</sup> The agent's expected payoff is

$$\pi(a, x, p) = (1 - p)[ar + x_1] + p[ar + x_2(a)]. \quad (1)$$

The three-period dynamic game of incomplete information just described encapsulates Newcomb's problem in its standard form. As solution concept we use a subgame-perfect Nash equilibrium (Selten 1965).<sup>7</sup>

*Remark 1* (Does  $\mathcal{G}$  represent the Newcomb problem?) The main differences between the standard exposition of the Newcomb problem and the above game-theoretic formulation is that in  $\mathcal{G}$  the payoffs for the being and the beliefs for the agent are made explicit. Regarding the former, our conclusions depend only on the assumption that  $B$ 's payoffs are negative whenever its prediction is incorrect with positive probability. As far as  $A$ 's beliefs are concerned, they constitute in fact the "secret ingredient" which allows a resolution of the paradox. The agent's beliefs imply the being's ability to commit to the rules of the game.<sup>8</sup> Its value has nothing to do with  $B$ 's predictive accuracy of at least 99.9 % in the standard narrative (see also Appendix 1). In fact, on the equilibrium path of  $\mathcal{G}$  the being's predictive accuracy is 100 %, independent of  $p$ . The formulation of Newcomb's paradox as a dynamic game of the form  $\mathcal{G}$  recasts

<sup>5</sup> See Appendix 1 for the case of asymmetric information.

<sup>6</sup> Without loss of generality,  $B$  will never have an incentive to randomize allocations, since doing so would necessarily result in a negative expected payoff.

<sup>7</sup> One can also use a perfect Bayesian equilibrium as solution concept (see Appendix 1), but it turns out that the corresponding equilibrium refinement, introducing  $A$ 's beliefs about which node in his information set has been reached by  $B$ 's preceding action, does not change any of the results.

<sup>8</sup> One may also argue that in a fully specified game, no player can violate the rules of the game.

the problem effectively as a commitment problem (see, e.g., [Weber 2014](#)) where the degree of  $B$ 's commitment is determined by  $A$ 's beliefs.

*Remark 2* (Alternative versions of  $\mathcal{G}$ ) (i) By collapsing the periods  $t = 1$  and  $t = 1.5$  in  $\mathcal{G}$  into a single period at  $t = 1$ , the resulting two-period game  $\hat{\mathcal{G}}$  consists of a simultaneous-move stage game at time  $t = 1$ , followed by  $B$ 's decision at time  $t = 2$ . Because in  $\mathcal{G}$  the agent is at time  $t = 1.5$  unaware of  $B$ 's initial choice, the game  $\hat{\mathcal{G}}$  is equivalent to  $\mathcal{G}$ . (ii) Conceptually, one can interpret the last period of  $\mathcal{G}$  as a device that with probability  $p$  ensures the reward in box II is  $R$  if and only if  $a = 0$ , and with probability  $1 - p$  does nothing. However, while such a device would produce outcome equivalence with the unique subgame-perfect Nash equilibrium of  $\mathcal{G}$ , the resulting one-person game (against a machine) would not deliver a plausible explanation for the incentive compatibility of the equilibrium with the being's objectives. Instead, by hard-wiring  $B$ 's equilibrium strategy, this interpretation merely illustrates the fact that unilateral deviations from  $A$ 's equilibrium strategy are not in the agent's best interest. (iii) It is also possible to delete the first period in  $\mathcal{G}$  and think of  $B$ 's occurring as one of two types, the first being able to condition its action on  $A$ 's action, and the second moving without knowledge of  $A$ 's action. In that setting, the agent's common-knowledge prior belief about the distribution of  $B$ 's types is described by  $p$ . In that interpretation of the game, the corresponding Bayes-Nash equilibrium would specify  $B$ 's strategy at time  $t = 2$  as a mapping from types to actions, analogous to the branches before the last decision nodes in Fig. 1. (iv) To avoid an extensive form where the belief probability  $p$  is relevant for six different events, it is also possible to introduce Nature as a player that at time 0 randomizes over two possible move orders,<sup>9</sup> which induces the belief probability  $p$ .

Taking into account that the being explicitly discourages randomization by the agent ([Nozick 1969](#), pp. 115/143), one obtains<sup>10</sup>

$$a \in (0, 1) \Rightarrow x_2(a) = 0. \quad (2)$$

As a result, the agent can never strictly prefer randomization [when  $a \in (0, 1)$ ] to a pure strategy (when  $a \in \{0, 1\}$ ). In order to exclude noncredible threats,<sup>11</sup> we limit attention to subgame-perfect Nash equilibria, which can be found via backward induction, starting in the last period (which contains the only proper subgame of  $\mathcal{G}$ ).

At time  $t = 2$ , to avoid the possibility of a negative payoff,  $B$  has no choice but to set

$$x_2(0) = R \quad \text{and} \quad x_2(1) = 0. \quad (3)$$

<sup>9</sup> Representing the players' information structure turns out to be more complex for this setup because the world bifurcates at the beginning of time, instead of (equivalently) realizing at the end of time.

<sup>10</sup> Since in equilibrium players know the strategy profile,  $B$  would be aware of any randomization.

<sup>11</sup> In the absence of subgame perfection, the being can resort to noncredible threats, for example in the Nash equilibrium, where the agent always takes both boxes and the being always places nothing in box II (see Sect. 3).

Combining Eqs. (2) and (3),  $B$ 's only possible choice at time  $t = 2$  is

$$x_2(a) = \begin{cases} R, & \text{if } a = 0, \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

This corresponds to a forcing strategy, which rewards only perfect compliance by the agent using the strongest possible incentive scheme, independent of  $p$ .

At time  $t = 1.5$ , the agent chooses  $a$  to maximize his expected payoff in Eq. (1). Taking into account Eq. (4), for all  $a \in (0, 1)$  it is

$$\pi(1, x, p) - \pi(a, x, p) = (1 - a)r > 0,$$

so taking both boxes always strictly dominates any nondegenerate mixed strategy the agent might consider, independent of  $p$ . Furthermore,

$$\pi(0, x, p) \geq \pi(1, x, p)$$

if and only if

$$p \geq q = \frac{r}{R}.$$

Hence, as a function of his belief  $p$  the agent's equilibrium strategy is

$$a(p) \in \begin{cases} 1, & \text{if } p < q, \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

The agent's choice depends only on the sign of  $q - p$ ; it is independent of  $x_1$ .<sup>12</sup> The dashed black lines in Fig. 2 show  $A$ 's payoffs as a function of his belief  $p$ , for the two possible actions of  $B$  at time 1 (corresponding to  $x_1 \in \{0, R\}$ ), respectively. The upper envelope of each strategy (shaded in grey) shows  $A$ 's payoffs given his equilibrium strategy in Eq. (5), for both of  $B$ 's possible time-1 actions.

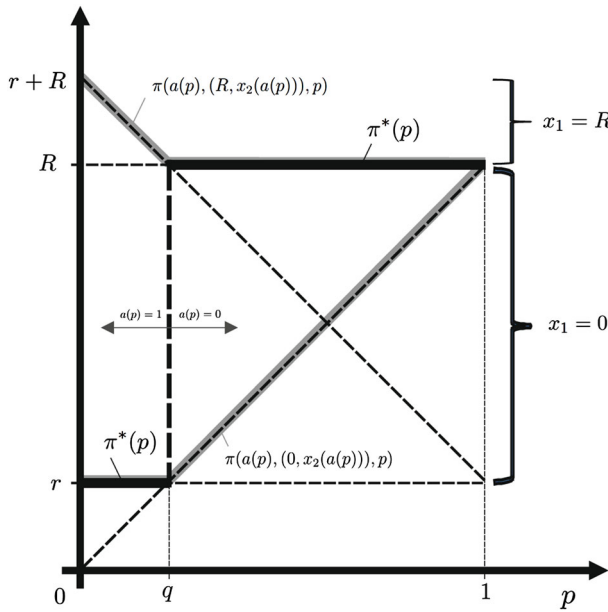
At time  $t = 1$ , the being uses the common knowledge about the agent's belief  $p$  [and  $A$ 's action  $a(p)$  as specified in Eq. (5)] to achieve a nonnegative payoff, by setting

$$x_1(p) = \begin{cases} R, & \text{if } p > q, \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

Thus if  $p \geq q$ , the being can implement  $a = 0$ . As already noted, the agent's choice will not depend on  $x_1$ . Yet, the being can guarantee a perfect prediction by choosing  $x_1 = R$ . If  $p < q$ , it is better for the agent to take both boxes, i.e., to choose  $a = 0$ . Hence, the being chooses  $x_1 = 0$ , and  $x_2(\cdot)$  applies to out-of-equilibrium behavior as well. The solid black line in Fig. 2 depicts  $A$ 's resulting equilibrium payoff.

<sup>12</sup> The potential indifference at  $p = q$  is broken, since by Eq. (4) the agent should rationally prefer the one-box strategy, i.e.,  $a(q) = 0$ , for in equilibrium it yields  $R$  (rather than  $r$  when using the two-box strategy).





**Fig. 2** Risk-neutral agent's expected payoff, for  $p \in [0, 1]$

**Proposition 1** For any risk-neutral agent with belief  $p \in [0, 1]$ , there is a unique subgame-perfect Nash equilibrium (SPNE)  $(a(p), x(p))$ , specified by Eqs. (4)–(6). In this equilibrium, the being makes perfect predictions, and the agent selects the two-box strategy if and only if  $p < q$ .

*Proof* By backward induction  $x_2(a)$  in Eq. (4) is  $B$ 's unique response at  $t = 2$  to any of  $A$ 's feasible actions  $a \in \mathcal{A}$  at  $t = 1.5$ . Anticipating this response,  $A$ 's unique optimal action (taking into account footnote 12) is given in Eq. (5). Thus, choosing  $x_1$  as in Eq. (6) is  $B$ 's only way to guarantee a perfect prediction (and thus a nonnegative payoff) in equilibrium.  $\square$

Consider the agent's expected payoff  $\pi^*(p) \equiv \pi(a(p), (x_1(p), x_2(a(p))), p)$  in equilibrium (also shown in Fig. 2). For  $p < q$ , the agent ends up with the low payoff  $r$ , analogous to the prisoner's dilemma solutions. The equilibrium in Proposition 1 does not rely on detailed assumptions about the being's payoffs, other than that  $B$  should have a preference for making correct predictions; in particular, no assumption inconsistent with the general narrative of Newcomb's problem is made. For  $p \geq q$ , the agent enjoys the sure payoff  $R$ , higher than any solution proposed thus far, except for the cooperative supergame solution in the repeated prisoner's dilemma.

### 3 Results for an agent with arbitrary risk attitude

The setting of the last section is generalized to an agent with an increasing utility function for money  $u(\cdot)$ , which represents preferences that are locally nonsatiated. So

far, Newcomb’s problem has not been considered for agents with a nonneutral risk attitude.<sup>13</sup> Without loss of generality one can select a cardinal representation of the agent’s preferences such that <sup>14</sup>

$$u(0) = 0 \quad \text{and} \quad u(R) = 1; \tag{7}$$

this simplifies some of the discussion below. The agent’s expected utility becomes

$$\bar{u}(a, x, p) = (1 - p)u(ar + x_1) + pu(ar + x_2(a)), \tag{8}$$

for any  $a \in \mathcal{A}$  and  $x = (x_1, x_2(\cdot))$  as specified in Sect. 2. Using the arguments in Sect. 2, Eqs. (2)–(3) continue to hold. Hence, at time  $t = 2$  the being sets  $x_2(\cdot)$  as in Eq. (4).

At time  $t = 1.5$ , when maximizing his expected utility in Eq. (8), the agent cares about the value of  $x_1$ . For  $x_1 = 0$ , the agent’s optimal action is

$$a_0(p) = \begin{cases} 1, & \text{if } p < q_0, \\ 0, & \text{otherwise;} \end{cases} \tag{9}$$

the decision threshold in that case is given by

$$q_0 = \frac{u(r) - u(0)}{u(R) - u(0)} = u(r) \in (0, 1).$$

Similarly, if  $x_1 = R$ , the agent’s optimal action becomes

$$a_1(p) = \begin{cases} 1, & \text{if } p < q_1, \\ 0, & \text{otherwise,} \end{cases} \tag{10}$$

with corresponding decision threshold

$$q_1 = \frac{u(r + R) - u(R)}{u(r + R) - u(r)} = \frac{u(r + R) - 1}{u(r + R) - u(r)} \in (0, 1).$$

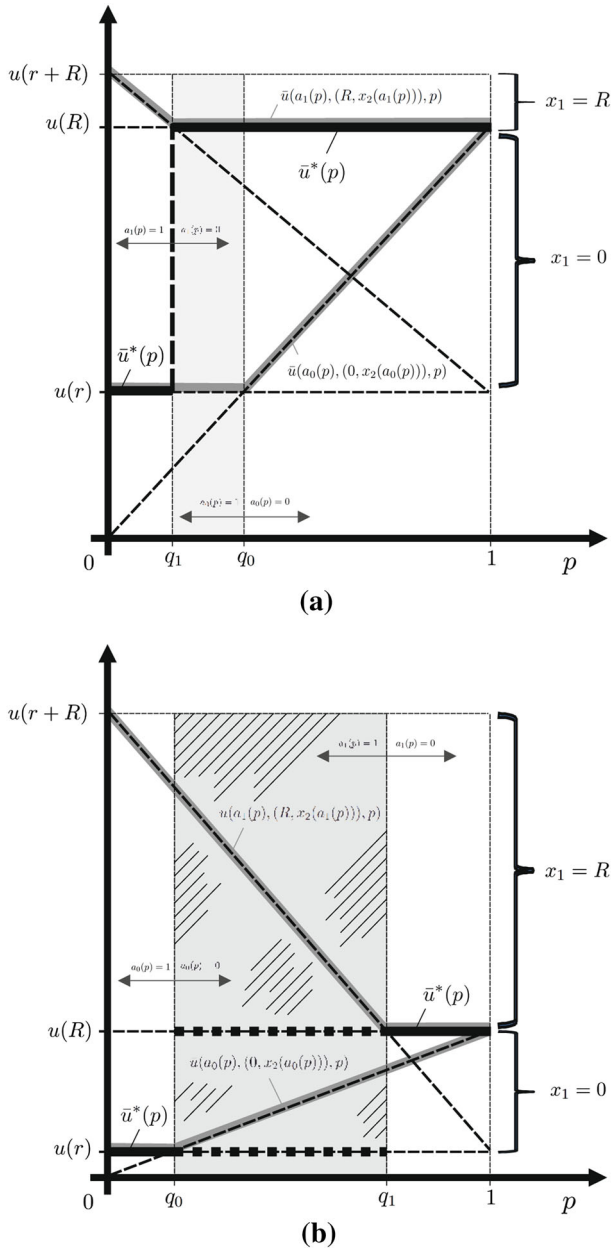
Fig. 3 shows the agent’s expected utility and optimal actions for different utility functions; the line types follow the same logic as in Fig. 2. If  $x_1(p) \in \{0, R\}$  can be chosen in such a way that  $A$ ’s optimal action is

$$a(p) = 1 - (x_1(p)/R), \tag{11}$$

then  $B$  is guaranteed a perfect prediction [in conjunction with Eq. (4)], and thus a nonnegative payoff. Otherwise a perfect prediction is not possible given the said

<sup>13</sup> Contrary to what Rapoport (1975, p. 613) claims, a rescaling of the payoffs cannot reproduce the effects discussed here.

<sup>14</sup> For any increasing utility of money  $\hat{u}(\cdot)$ , the equivalent representation  $u = c_1\hat{u} + c_2$  satisfies Eq. (7) with  $c_1 = (\hat{u}(r) - \hat{u}(0))^{-1} > 0$  and  $c_2 = -\hat{u}(0)/(\hat{u}(r) - \hat{u}(0))$ .



**Fig. 3** Expected payoffs when agent is risk-averse (a) or risk-seeking (b), for  $p \in [0, 1]$

specification of  $x_2(\cdot)$  in Eq. (4), which was obtained via backward induction. We now show that the existence of a subgame-perfect Nash equilibrium with perfect prediction, where

$$x_1(p) = \begin{cases} R, & \text{if } p > q_1, \\ 0, & \text{otherwise,} \end{cases} \quad (12)$$

depends on the relative magnitudes of the decision thresholds  $q_0$  and  $q_1$ .

**Proposition 2** Consider an agent with belief  $p \in [0, 1]$  and arbitrary risk attitude.

- (i) If  $q_1 \leq q_0$ , then there is a unique subgame-perfect Nash equilibrium  $(a(p), x(p))$ , specified by Eqs. (4), (11) and (12).
- (ii) If  $q_0 < q_1$  and  $p \notin (q_0, q_1)$ , then there is a unique pure-strategy Nash equilibrium specified as in part (i).
- (iii) If  $q_0 < p < q_1$ , then there is a unique pure-strategy Nash equilibrium. In this equilibrium with  $a(p) = 1$  and  $x(p) = 0$ , the being relies on the noncredible threat of never placing the reward, even when it could be placed contingent on the agent's choice.

*Proof* (i) Assume  $q_1 \leq q_0$ . If  $B$  sets  $x_1(p) = 0$  for  $p < q_1$  and  $x_1(p) = R$  for  $p \geq q_1$ , then by Eq. (9) it is  $a(p) = a_0(p) = 0$  for  $p < q_1$ . By Eq. (10) it is  $a(p) = a_1(p) = 1$  for  $p \geq q_1$ . Moreover, since  $a_1(p) = a_0(p) = 0$  for  $p < q_1$ , we obtain  $A$ 's optimal action at time  $t = 1.5$ ,

$$a(p) = a_1(p),$$

for all  $p \in [0, 1]$ . Lastly, at time  $t = 1$  anticipating  $a(p)$  and  $x_2(a(p))$ , the being expects a nonnegative payoff and would therefore not want to deviate, so Eqs. (4), (11) and (12) specify indeed an SPNE. This equilibrium is unique, since  $a(p)$  and  $x_2(\cdot)$  are unique, analogous to the proof of Proposition 1 and any deviation by  $B$  at  $t = 1$  leads to a violation of Eq. (11) and thus to an expected negative payoff due to possibly imperfect prediction. (ii) Assume  $q_0 < q_1$  and  $p \notin (q_0, q_1)$ . If  $p < q_0$  (resp.  $p \geq q_1$ ), then  $A$ 's optimal action is  $a_0(p) = 1$  (resp.  $a_1(p) = 0$ ). Hence, we obtain  $a(p) = a_1(p)$  as in part (i) and therefore the existence of a unique SPNE with the same specifications as before, for these extreme beliefs  $p \notin (q_0, q_1)$ . (iii) Assume  $q_0 < p < q_1$ . If  $x_1(p) = 0$ , then  $A$ 's optimal action is  $a(p) = a_0(p) = 0$  which violates Eq. (9). Otherwise, if  $x_1(p) = R$ , then  $A$ 's optimal action  $a(p) = a_1(p) = 1$  again violates Eq. (9). Hence, no matter what  $B$ 's allocational choice at time  $t = 1$ , there is a positive probability of an incorrect prediction, resulting in a negative expected payoff. To improve the probability  $1 - p$  of failing to choose a correct allocation using pure strategies, the being would want to randomize in equilibrium, eliminating thus the possibility of a pure-strategy SPNE. On the other hand, a pure-strategy subgame-imperfect Nash equilibrium does exist, with  $B$  choosing  $x(p) = 0$ , i.e., the allocation  $x_t = 0$  for  $t \in \{1, 2\}$ , to which the agent finds it (weakly) optimal to respond with  $a(p) = 0$ , which satisfies Eq. (11) and thus ensures a perfect prediction.  $B$ 's threat of an unconditional zero allocation is not credible, because were the agent to deviate and take only box II ( $a = 0$ ), then  $B$  would find it beneficial to respond with  $x_2 = R$  in order to avoid an allocation error.  $\square$

The critical condition  $q_1 \leq q_0$  in Proposition 2 is equivalent to<sup>15</sup>

$$u(r + R) - u(R) \leq u(r) - u(0). \quad (13)$$

In other words, the utility difference from owning the additional reward  $r$  does not go up when the default reward increases from 0 to  $R$ . Since  $u(0) = 0$  by Eq. (7), this holds for all  $r, R \geq 0$  if and only if the utility is subadditive, so  $u(r + R) \leq u(r) + u(R)$ . A sufficient condition for inequality (13) to hold is that  $u$  is concave, i.e., the agent is risk-averse. In the special case where the agent is risk-neutral, the thresholds are identical,  $q_0 = q_1 = q$ , and the same results as in Sect. 2 obtain. When the agent is risk-seeking, that is, his utility function is strictly convex, the inequality (13) becomes false for all positive  $r, R$  with  $r < R$ , that is  $q_1 > q_0$ , which according to part (iii) of Proposition 2 implies a region of beliefs for which a pure-strategy SPNE does not exist. The expected equilibrium utility,  $\bar{u}^*(p) = \bar{u}(a(p), (x_1(p), x_2(a(p))), p)$  together with the agent's optimal actions are shown in Fig. 3a for the "risk-averse" case where  $q_0 > q_1$  and in Fig. 3b for the "risk-seeking" case where  $q_1 > q_0$  (where the dashed bold line in the shaded area indicates the support of payoffs in a mixed-strategy equilibrium).

## 4 Discussion

Beginning with Nozick (1969) the discussion of Newcomb's problem has centered on trying to adjudicate which of two alternative views of the world is correct, a deterministic ("causal") framework in which future outcomes are fully deductible from past states or an indeterministic ("evidentiary") framework in which future outcomes cannot be predicted with certainty and depend inductively on past states.<sup>16</sup> Depending on which of the two is chosen, solutions to the problem were proposed via a dominance argument (supporting the two-box strategy) or an expected-utility-maximization argument (supporting the one-box strategy), respectively, with no possible resolution between these two solutions. Gardner (1973) noted a third approach by Immanuel Kant which "accepts both sides as being equally true but incommensurable ways of viewing human behavior" (p. 104). In physics, a version of this third approach was formulated by Niels Bohr in his "principle of complementarity" which states, roughly speaking, that reality must depend on the observer (Bohr 1958/1961; Plotnitsky 2012).

Much in the same way as Bohr's complementarity principle, our proposed solution depends on the agent, specifically his beliefs in the being's ability to commit. And the being, knowing these beliefs, will react to them. For an agent who thinks it unlikely that the being can condition the content of box II on his action, the two-box strategy prevails, essentially because of dominance. On the other hand, an agent who thinks that the being is likely to condition the content of box II on his own action, prefers the one-box strategy. In this context, the parameter  $p$  can also be viewed as continuously indexing the being's ability to commit to an allocation at time  $t = 1$ , from  $p = 0$  for

<sup>15</sup> It is  $q_1 \leq q_0 \Leftrightarrow u(r) \geq (u(r + R) - 1)/(u(r + R) - u(r)) \Leftrightarrow u(r + R) - u(r) \leq 1 = u(R) - u(0)$ .

<sup>16</sup> Nozick (1993, pp. 44–45) acknowledges that decisions in Newcomb's problem may depend on the relative magnitude of the reward  $R$  compared to  $r$  and that for larger  $q = r/R$  the dominance ("causal") argument is less plausible.

full commitment to  $p = 1$  for no commitment. At the ends of the spectrum one obtains the two-box outcome (at low reward  $r$ ) and the one-box outcome (at high reward  $R$ ), respectively. For intermediate values of  $p$ , the being's perceived ability depends on the agent's perception of reality, in particular his risk preference. In this interpretation, Newcomb's problem is related to commitment measured on a continuum because an initial action may at least be partially revocable (Henkel 2002; Weber 2014).<sup>17</sup> Instead of a belief, the parameter  $p$  then indexes an objective ability by the being to revise the earlier allocation after the agent selected his strategy. In either interpretation, as in the Coase problem (see footnote 17), the agent benefits from the other party's inability to commit. As much as the being lacks the ability to commit, the agent's own action becomes influential in determining the final payoffs of the game.

The solution to Newcomb's problem presented in Sects. 2 and 3 is robust in the sense that almost no assumptions were made about the being's intentions, other than a preference for making correct predictions.<sup>18</sup> The being also does not need very complex distributional assumptions about the agent's potential preferences other than the decision threshold  $q$  (resp.  $q_1$  in the case with general risk attitude; see Sect. 3). Even if the being operates in the complete absence of any information about the agent (see Appendix 1), there is a high chance, given the small  $q = r/R = 0.001$  (using the standard numbers), that the being's decisions would lead to correct predictions, exactly as in the usual narrative of Newcomb's problem.

**Acknowledgments** The author would like to thank several anonymous referees, Paul B. Kantor, and participants of the 2015 INFORMS Annual Meeting in Philadelphia, Pennsylvania, for valuable comments and suggestions.

**Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

## Appendix 1: Asymmetric information

Suppose that information about a risk-neutral agent's belief is private, or equivalently, that the being faces a population of different agents with beliefs  $\theta \in \Theta = [\underline{\theta}, \bar{\theta}]$ , where  $0 \leq \underline{\theta} \leq \bar{\theta} \leq 1$ .  $B$  encounters one of these agents at random.<sup>19</sup> From the preceding analysis we have seen that what matters for the agent's binary decision is whether his beliefs are above or below the threshold  $q = r/R$ . Hence, when design-

<sup>17</sup> The continuum of commitment appears in discussions about the robustness of the Coase conjecture (Coase 1972) which depends on zero commitment. The *degree* of commitment available to a monopolist in the Coase setup varies continuously with the length of time that passes until the price for a durable good can be changed (Gul et al. 1986); an analogous situation is obtained with varying degrees of perishability of goods to be sold (Cho 2007).

<sup>18</sup> In virtually all extant game-theoretical approaches, the "superior being" is assumed a "replica" of the agent (see also footnote 2), which amounts to an introspective approach that is difficult to justify for any concrete instantiation of the game, the original description of which, by Nozick (1969), is asymmetric to begin with.

<sup>19</sup> The situation in Sect. 2 obtains for  $\underline{\theta} = \bar{\theta} = p$ , so  $\Theta = \{p\}$ .

ing a mechanism  $(\mathcal{M}, \xi)$  to reveal the agent’s information,  $B$  can restrict attention to a binary message space  $\mathcal{M} = \{L, H\}$ , where  $L$  is a message that (if truthful) indicates a low type ( $p < q$ ), and  $H$  is a message that (if truthful) indicates a high type ( $p \geq q$ ). For each action-message tuple  $(a, m) \in \mathcal{A} \times \mathcal{M}$ , the being chooses the allocation  $\xi(a, m) = (\xi_1(m), \xi_2(a, m))$ . To minimize the prediction error on the equilibrium path,  $B$  does not randomize allocations, so  $\xi(a, m) \in \{(0, 0), (v, R), (R, v), (R, R)\}$  for all  $(a, m) \in \mathcal{A} \times \mathcal{M}$ , where  $v$  is a real number to be determined below. Analogous to the equilibrium concept chosen in the main text, we limit attention to perfect Bayesian Nash equilibria.

**Proposition 3** *When the information about the belief  $p = \theta$  is private to the agent and  $\underline{\theta} > 0$ , there is a revelation mechanism under which the agent does not object to reporting his beliefs truthfully and the being makes no mistakes.*

*Proof* Given a mechanism  $(\mathcal{M}, \xi)$ , let  $(\alpha, \mu)$  with  $\alpha : \Theta \rightarrow \mathcal{A}$  and  $\mu : \Theta \rightarrow \mathcal{M}$  be the agent’s strategy. Then

$$(\alpha(\theta), \mu(\theta)) \in \arg \max_{(a,m) \in \mathcal{A} \times \mathcal{M}} \{ar + (1 - \theta)\xi_1(m) + \theta\xi_2(a, m)\}.$$

To avoid mistakes,  $B$  has no choice but to discourage  $a \in (0, 1)$  by leaving box II empty, that is  $\xi_2(a, m) = 0$  irrespective of  $m$ ; therefore randomization is never optimal for the agent. Consider an allocation function  $\xi = (\xi_1, \xi_2)$  with  $\xi_1 : \mathcal{M} \rightarrow \mathbb{R}$  and  $\xi_2 : \mathcal{A} \times \mathcal{M} \rightarrow \mathbb{R}$  with

$$\xi_1(L) = \xi_1(H) = R, \tag{14}$$

and

$$\xi_2(a, m) = \begin{cases} R, & \text{if } a = 0, \\ v, & \text{if } (a, m) = (1, L), \\ 0, & \text{otherwise,} \end{cases} \tag{15}$$

where the real number  $v$  is determined below. We know for  $\theta < q$  the agent prefers  $(1, L)$  and for  $\theta \geq q$  he prefers  $(0, H)$  in the set of deterministic actions  $(a, m)$  in  $\{(0, L), (1, L), (0, R), (1, R)\}$ . That is, any agent reports truthfully whether his belief is above or below the threshold and otherwise chooses the one-box strategy.

1.  $\theta \in [q, \bar{\theta}]$ : The agent prefers  $(0, H)$  to  $(1, H)$ , since by Eqs. (14)–(15):

$$\begin{aligned} (1 - \theta)\xi_1(H) + \theta\xi_2(0, H) &= R > (1 - \theta)R \\ &= (1 - \theta)\xi_1(H) + \theta\xi_2(1, H). \end{aligned} \tag{16}$$

Furthermore,  $(0, H)$  is (weakly) preferred to  $(0, L)$  since either action yields a sure payoff of  $R$ . Lastly,  $(0, H)$  is also preferred to  $(1, L)$  because

$$\begin{aligned} (1 - \theta)\xi_1(H) + \theta\xi_2(0, H) &= R \geq r + (1 - \theta)\xi_1(L) + \theta\xi_2(1, L) \\ &= r + (1 - \theta)R + \theta\xi_2(1, L), \end{aligned}$$

for all  $\theta \geq q = r/R$ , which is equivalent to  $\xi_2(1, L) \leq 0$ .

2.  $\theta \in [\underline{\theta}, q)$ : By Eqs. (14)–(15) the agent prefers  $(0, L)$  to  $(1, L)$  and to  $(1, H)$  if and only if

$$R \geq r + (1 - \theta)R + \theta v.$$

This inequality is satisfied for all  $\theta \in [\underline{\theta}, q)$  if  $B$  sets

$$v \equiv R - \frac{r}{\underline{\theta}} = - \left( \frac{q}{\underline{\theta}} - 1 \right) R.$$

Finally, the agent (weakly) prefers  $(0, L)$  to  $(0, H)$ , for he is guaranteed a certain payoff of  $R$ .

This concludes our proof. □

The mechanism specified in Proposition 3 depends on the being's ability to deter the agent from deviations with negative payoffs. The being therefore needs to be able to credibly threaten the agent, saying something to the effect that if  $B$  gets to choose at time  $t = 2$  and finds that  $a \neq x_1 = 0$ , then box II is likely to “explode” when the agent opens it. Without this threat, the only other possibility for the being is to revert to the Nash equilibrium where  $B$  never puts any reward in box II, relying on the noncredible threat discussed at the end of the proof of Proposition 2.

When the agent has a general risk attitude,  $B$  is left with a similar dichotomy and thus needs to decide whether case (iii) in Proposition 2 applies or not. If negative payoffs can be implemented, then a mechanism analogous to the one in Proposition 3 can be found. Otherwise,  $B$  needs to form beliefs about the population of agents. For example, if in the risk-neutral case, agents' types (beliefs) are uniformly distributed on  $\Theta = [0, 1]$ , then for  $q = 0.001$  the being in the equilibrium specified in Proposition 1, assuming  $p \geq q$ , would make a correct prediction 99.9 % of the time, consistent with the standard narrative of Newcomb's problem. Note also that in our resolution of Newcomb's problem, the agent does not need to make any assumptions (other than the structure of the game, which we have tacitly assumed to be common knowledge), but rather just act according to his beliefs.

## References

- Aumann, R. J. (1976). Agreeing to disagree. *Annals of Statistics*, 4(6), 1236–1239.
- Bar-Hillel, M., & Margalit, M. (1972). Newcomb's paradox revisited. *British Journal for the Philosophy of Science*, 23(4), 295–304.
- Benditt, T. M., & Ross, D. J. (1976). Newcomb's paradox. *British Journal for the Philosophy of Science*, 27(2), 161–164.
- Bohr, N. (1958/1961). Quantum physics and philosophy: causality and complementarity. In R. Klamburg (Ed.), *Philosophy in the mid-century: a survey, logic and philosophy of science*, 2nd edn, vol. I (pp. 308–314). Florence: Nuova Italia.
- Brams, S. J. (1975). Newcomb's problem and prisoners dilemma. *Journal of Conflict Resolution*, 19(4), 596–612.
- Brandenburger, A. (1992). Knowledge and equilibrium in Games. *Journal of Economic Perspectives*, 6(4), 83–101.



- Broome, J. (1989). An economic Newcomb problem. *Analysis*, 49(4), 220–222.
- Burgess, S. (2004). The Newcomb problem: An unqualified resolution. *Synthese*, 138(2), 261–287.
- Cargile, J. (1975). Newcomb's paradox. *British Journal for the Philosophy of Science*, 26(3), 234–239.
- Cho, I.-K. (2007). Perishable durable goods. Working paper, Department of Economics, University of Illinois at Urbana, Champaign.
- Coase, R. (1972). Durability and monopoly. *Journal of Law and Economics*, 15(1), 143–149.
- Eells, E. (1984). Newcomb's many solutions. *Theory and Decision*, 16(1), 59–105.
- Fudenberg, D., & Maskin, E. (1986). The Folk theorem in repeated games with discounting or with imperfect public information. *Econometrica*, 54(3), 533–554.
- Frydman, R., O'Driscoll, G. P., & Schotter, A. (1982). Rational expectations of government policy: An application of Newcomb's problem. *Southern Economic Journal*, 49(2), 311–319.
- Gardner, M. (1973) Free will revisited, with a mind-bending prediction Paradox by William Newcomb. In *Scientific American*, vol. 229(1) (pp. 104–109) [Expanded and reprinted in: Gardner, M. (2001) *The Colossal Book of Mathematics*, Norton, New York, NY, Chapter 44, pp. 580–591].
- Gardner, M. (1974). Reflections on Newcomb's problem: A prediction and free-will dilemma. *Scientific American*, 230(3), 102–109.
- Gul, F., Sonnenschein, H., & Wilson, R. B. (1986). Foundations of dynamic monopoly and the Coase conjecture. *Journal of Economic Theory*, 39, 155–190.
- Henkel, J. (2002). The 1.5th mover advantage. *RAND Journal of Economics*, 33(1), 156–170.
- 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.
- Hurley, S. L. (1994). A new take from Nozick on Newcomb's problem and prisoners dilemma. *Analysis*, 54(2), 65–72.
- Kydland, F. E., & Prescott, E. C. (1977). Rules rather than discretion: The inconsistency of optimal plans. *Journal of Political Economy*, 85(3), 473–492.
- Langan, C. (2002). *The art of knowing: Expositions on free will and select essays*. Eastport: Mega Foundation Press.
- Levi, I. (1975). Newcomb's many problems. *Theory and Decision*, 6(2), 161–175.
- Lewis, D. (1969). *Convention: A philosophical study*. Cambridge: Harvard University Press.
- Lewis, D. (1979). Prisoners dilemma is a Newcomb problem. *Philosophy and Public Affairs*, 8(3), 235–240.
- Libet, B., Gleason, C. A., Wright, E. W., & Pearl, D. K. (1983). Time of conscious intention to act in relation to onset of cerebral activity (readiness potential)—the unconscious initiation of a freely voluntary act. *Brain*, 106(Pt. 3), 623–642.
- Locke, D. (1978). How to make a Newcomb choice. *Analysis*, 38(1), 17–23.
- Muth, J. F. (1961). Rational expectations and the theory of price movements. *Econometrica*, 29(3), 315–335.
- 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.
- Nozick, R. (1993). *The nature of rationality*. Princeton: Princeton University Press.
- Nozick, R. (1997). *Socratic puzzles*. Cambridge: Harvard University Press.
- Plotnitsky, A. (2012). *Niels Bohr and complementarity*. New York: Springer.
- Rapoport, A. (1975). Comment on Bram's discussion of Newcomb's paradox. *Journal of Conflict Resolution*, 19(4), 613–619.
- Reggia, J. A. (2013). The rise of machine consciousness: Studying consciousness with computational models. *Neural Networks*, 44, 112–131.
- Savage, L. J. (1954). *The foundations of statistics*. New York: Wiley (expanded edition published in 1972 by Dover, New York, NY).
- Schlesinger, G. (1974). The unpredictability of free choices. *British Journal for the Philosophy of Science*, 25(3), 209–221.
- Schmidt, J. H. (1998). Newcomb's paradox realized with backward causation. *British Journal for the Philosophy of Science*, 49(1), 67–87.
- Selten, R. (1965). Spieltheoretische Behandlung eines Oligopolmodells mit Nachfrageträgheit. *Zeitschrift für die Gesamte Staatswissenschaft*, 121, 301–324, 667–689.
- Schmidt-Petri, C. (2004). Newcomb's problem and repeated prisoners' dilemma. In *Philosophy of science (Proceedings of the 2004 Biennial meeting of the Philosophy of Science Association)*, vol. 72(2) (pp. 1160–1173).
- Sobel, J. H. (1988). Infallible predictors. *Philosophical Review*, 97(1), 3–24.

- Sorenson, R. A. (1983). Newcomb's problem: Recalculations for the one-boxer. *Theory and Decision*, *15*(4), 399–404.
- Sorenson, R. A. (1985). The iterated versions of Newcomb's problem and the prisoner's dilemma. *Synthese*, *63*(2), 157–166.
- Sugden, R. (1991). Rational choice: A survey of contributions from economics and philosophy. *Economic Journal*, *101*(407), 751–785.
- Weber, T. A. (2011). *Optimal control theory with applications in economics*. Cambridge: MIT Press.
- Weber, T. A. (2014). A continuum of commitment. *Economics Letters*, *124*(1), 67–73.