

# What does Newcomb's paradox teach us?

David H. Wolpert<sup>1</sup> & Gregory Benford<sup>2</sup>

1 - NASA Ames Research Center, MS 269-1, Moffett Field, CA 94035-1000  
(650) 604-3362 (V), (650) 604-3594 (F), david.h.wolpert.nasa.gov

2 - Physics and Astronomy Department, University of California,  
Irvine, CA 92692

April 7, 2009

## Abstract

Newcomb's paradox highlights an apparent conflict involving the axioms of game theory. It concerns a game in which you choose to take either one or both of two closed boxes. However before you choose, a prediction algorithm deduces your choice, and fills the two boxes based on that deduction. The paradox is that game theory appears to provide two conflicting recommendations for what choice you should make in this situation. Here we analyze Newcomb's paradox using a recently introduced extension of game theory in which the players set conditional probability distributions in a Bayes net. Using this extended game theory, we show that the two game theory recommendations in Newcomb's scenario implicitly assume different Bayes nets relating the random variables of your choice and the algorithm's prediction. We resolve the paradox by proving that these two assumed Bayes nets are incompatible, i.e., the associated assumptions conflict. In doing this we show that the accuracy of the algorithm's prediction, which was the focus of much previous work on Newcomb's paradox, is irrelevant. We also show that Newcomb's paradox is time-reversal invariant; both the paradox and its resolution are unchanged if the algorithm makes its "prediction" *after* you make your choice rather than before.

**Keywords:** Newcomb's paradox, game theory, Bayes net, causality, determinism

# 1 Introduction

Suppose you meet a Wise being (*W*) who tells you it has put \$1,000 in box A, and either \$1 million or nothing in box B. This being tells you to either take the contents of box B only, or to take the contents of both A and B. Suppose further that the being had put the \$1 million in box B only if a prediction algorithm designed by the being had said that you would take only B. If the algorithm had predicted you would take both boxes, then the being put nothing in box B.

Presume that due to determinism, there exists a perfectly accurate prediction algorithm. Assuming *W* uses that algorithm, what choice should you make? In Table 1 we present this question as a game theory matrix involving *W*'s prediction and your choice. Two seemingly logical answers contradict each other. The Realist answer is that you should take both boxes, because you have free will, and your choice occurs after *W* has already made its prediction. More precisely, if *W* predicted you would take A along with B, then taking both gives you \$1,000 rather than nothing. If instead *W* predicted you would take only B, then taking both boxes yields \$1,001,000, which again is \$1000 better than taking only B. The Fearful answer, though, is that *W* designed a prediction algorithm whose answer will match what you do. So you can get \$1,000 by taking both boxes or get \$1 million by taking only box B. Therefore you should take only B.

This is Newcomb's Paradox, a famous logical riddle stated by William Newcomb in 1960 [Nozick(1969), Gardner(1974), Bar-Hillel and Margalit(1972), Campbell and Lanning(1985) Levi(1982), Collins(2001)]. Newcomb never published the paradox, but had long conversations about it with philosophers and physicists such as Robert Nozick and Martin Kruskal, along with Scientific American's Martin Gardner. Gardner said after his second Scientific American column on Newcomb's paradox appeared that it generated more mail than any other column.

One of us (Benford) worked with Newcomb, publishing several papers together, and was a friend until Newcomb died in 1999. We often discussed the paradox, which Newcomb thought would be his best remembered scientific accomplishment. Newcomb invented his paradox to test his own ideas, as a lapsed Catholic: How much faith do we place in the wise being's predictive power? Newcomb's said that he would just take B; why fight a God-like being? However Nozick said, "To almost everyone, it is perfectly clear and obvious what should be done. The difficulty is that these people seem to divide almost evenly on the problem, with large numbers thinking that the opposing half is just being silly" [Nozick(1969)].

Nozick also pointed out that two accepted principles of game theory conflict in Newcomb's problem. The expected-utility principle, proceeding from the probability of each outcome, says you should take box B only. But the dominance principle argues that if one strategy is always better, no matter what the circum-

stances, then you should pick it. No matter what box B contains, you are \$1000 richer if you take both boxes than if you take B only.

Is there really a contradiction? Some philosophers argue that a perfect predictor implies a time machine, since with such a machine causality is reversed, i.e., the future causes past events, allowing predictions to be perfect.<sup>1</sup> Faced with Newcomb's seemingly logical paradox, the conclusion must be that perfect prediction is impossible.

But Nozick stated the problem specifically to exclude backward causation (and so time travel), because his formulation demands only that the predictions be of high accuracy, not certain. So this line of reasoning cannot resolve the issue. Worse still, Nozick's reformulation seems to imply that the (in)fallibility of  $W$ 's prediction provides yet another conundrum, in addition to the one underlying Newcomb's paradox.

## 2 Game theory over Bayes nets

Central to Newcomb's scenario is a prediction process, and its (in)fallibility. Recent work has revealed deep formal connections between prediction and observation. Amongst other things, this work proves that any given prediction algorithm must fail on at least one prediction task [Binder(2008), Wolpert(2008)]. Unfortunately, that result doesn't directly resolve Newcomb's paradox. However its proof requires an extension of game theory. And as we demonstrate below, that extension can be used to resolve Newcomb's paradox.

In game theory there are several "players", each with their own preferences over the values of an underlying set of game variables,  $\{X_j\}$ . Every player has their own "move set", where each move is a probability distribution relating some of the variables  $\{X_j\}$ . To play the game, the players all independently choose a move (i.e., choose a distribution) from their respective move sets. The moves sets are carefully designed so that every such joint move by the players uniquely specifies a legal joint probability distribution relating the game's variables [Fudenberg and Tirole(1991), Myerson(1991), Osborne and Rubenstein(1994), Koller and Milch(2003)].

A richer mathematics arises if we expand the move sets of the players, so that some joint moves would violate the laws of probability, and therefore are impossible. It is this mathematics that is used to prove the fallibility of prediction in [Binder(2008), Wolpert(2008)].

---

<sup>1</sup>Interestingly, near when Newcomb devised the paradox, he also coauthored a paper proving that a tachyonic time machine could not be reinterpreted in a way that precludes such paradoxes [Benford et al.(1970)Benford, Book, and Newcomb]. The issues of time travel and paradoxes are intertwined.

What happens if we apply this mathematics to Newcomb’s paradox? There are two game variables that are central to Newcomb’s paradox: the God-like being  $W$ ’s prediction,  $g$ , and the choice you actually make,  $y$ . So the player moves will involve the distribution relating those variables. Since there are only two variables, there are two ways to decompose that joint probability. These two decompositions turn out to correspond to the two recommendations for how to answer Newcomb’s question, one matching the reasoning of Realist and one matching Fearful.

The first way to decompose the joint probability is

$$P(y, g) = P(g | y)P(y) \tag{1}$$

(where we define the right-hand side to equal 0 for any  $y$  such that  $P(y) = 0$ ). Such a decomposition is known as a “Bayes net” having two “nodes” [Pearl(2000)]. The unconditioned distribution,  $P(y)$  is identified with the first, “parent” node, and the conditional distribution,  $P(g | y)$ , is identified with the second, “child” node.

This Bayes net can be used to express Fearful’s reasoning. Fearful interprets the statement that “ $W$  designed a perfectly accurate prediction algorithm” to imply that  $W$  has the power to set the conditional distribution in the child node of the Bayes net,  $P(g | y)$ , to anything it wants (for all  $y$  such that  $P(y) \neq 0$ ). More precisely, since the algorithm is “perfectly accurate”, Fearful presumes that  $W$  chooses to set  $P(g | y) = \delta_{g,y}$ , the distribution that equals 1 if  $g = y$ , zero otherwise. So Fearful presumes that there is nothing you can do that can affect the values of  $P(g | y)$  (for all  $y$  such that  $P(y) \neq 0$ ). Instead, you get to choose the unconditioned distribution in the parent node of the Bayes net,  $P(y)$ . Intuitively, this choice constitutes your “free will”.

Fearful’s interpretation of Newcomb’s paradox specifies what aspect of  $P(y, g)$  you can choose, and what aspect is instead chosen by  $W$ . Those choices —  $P(y)$  and  $P(g | y)$ , respectively — are the “moves” that you and  $W$  make. It is important to note that these moves by you and  $W$  do *not* directly specify the two variables  $y$  and  $g$ . Rather the moves you and  $W$  make specify two different distributions which, taken together, specify the full joint distribution over  $y$  and  $g$  [Koller and Milch(2003)]. This kind of move contrasts with the kind considered in decision theory [Berger(1985)] or causal nets [Pearl(2000)], where the moves are direct specifications of the variables (which here are  $g$  and  $y$ ).

In game theory, your task is to make the move that maximizes your expected payoff under the associated joint distribution. For Fearful, this means choosing the  $P(y)$  that maximizes your expected payoff under the  $P(y, g)$  associated with that choice. Given Fearful’s presumption that the Bayes net of Eq. 1 underlies the game and that you get to set the distribution at the first node, for you to maximize expected payoff you should choose  $P(y) = \delta_{y,B}$ , i.e., you should make choice

$B$  with probability 1. Your doing so results in the joint distribution  $P(y, g) = \delta_{g,y} \delta_{y,B} = \delta_{g,B} \delta_{y,B}$ , with payoff 1,000,000. This is the formal justification of Fearful’s recommendation.

The second way to decompose the joint probability is

$$P(y, g) = P(y | g)P(g) \tag{2}$$

(where we define the right-hand side to equal 0 for any  $g$  such that  $P(g) = 0$ ). In the Bayes net of Eq. 2, the unconditioned distribution identified with the parent node is  $P(g)$ , and the conditioned distribution identified with the child node is  $P(y | g)$ . This Bayes net can be used to express Realist’s reasoning. Realist interprets the statement that “your choice occurs after  $W$  has already made its prediction” to mean that you can choose any distribution  $h(y)$  and then set  $P(y | g)$  to equal  $h(y)$  (for all  $g$  such that  $P(g) \neq 0$ ). This is how Realist interprets your having “free will”. (Note that this is a different interpretation of “free will” from the one made by Fearful.) Under this interpretation,  $W$  has no power to affect  $P(y | g)$ . Rather  $W$  gets to set the parent node in the Bayes net,  $P(g)$ . For Realist, this is the distribution that you cannot affect. (In contrast, in Fearful’s reasoning, you set a non-conditional distribution, and it is the conditional distribution that you cannot affect.)

Realist’s interpretation of Newcomb’s paradox specifies what it is you can fix concerning  $P(y, g)$ , and what is fixed by  $W$ . Just like under Fearful’s reasoning, under Realist’s reasoning the “moves” you and  $W$  make do not directly specify the variables  $g$  and  $y$ . Rather the moves by you and  $W$  specify two distributions which, taken together, specify the full joint distribution. As before, your task is to choose your move — which now is  $h(y)$  — to maximize your expected payoff under the associated  $P(y, g)$ . Given Realist’s presumption that the Bayes net of Eq. 2 underlies the game and that you get to set  $h$ , you should choose  $h(y) = P(y | g) = \delta_{y,AB}$ , i.e., you should make choice  $AB$  with probability 1. Doing this results in the expected payoff  $1,000 P(g = AB) + 1,001,000 P(g = B)$ , which is your maximum expected payoff no matter what the values of  $P(g = AB)$  and  $P(g = B)$  are. This is the formal justification of Realist’s recommendation.<sup>2</sup>

What happens if we try to merge the Bayes net that Fearful presumes to underlie the game with the Bayes net that Realist presumes to underlie the game? More

---

<sup>2</sup>In Realist’s Bayes net, given the associated restricted possible form of  $P(y | g)$ ,  $g$  and  $y$  are “causally independent”, to use the language of causal nets [Pearl(2000)]. This is consistent with interpreting Newcomb’s scenario as the game in Table 1. In contrast, in Fearful’s Bayes net,  $y$  “causally influences”  $g$ . To cast this kind of causal influence in terms of conventional game theory, we would have to replace the game in Table 1 with an extensive form game in which you first set  $y$ , and *then*  $W$  moves, having observed  $y$ . This alternative game is incompatible with Newcomb’s stipulation that  $W$  moves before you do, not after. This is one of the reasons why it is necessary to use extended game theory rather than conventional game theory to formalize Fearful’s reasoning.

formally, what game arises if we combine your move set under Fearful’s presumption of the underlying Bayes net with your move set under Realist’s presumption, and do the same for  $W$ ? As we now know, combining move sets this way gives an “extended game” of the sort considered in [Wolpert(2008)], with the same kind of impossibility result as the extended game in [Wolpert(2008)].

First, if  $W$ ’s move sets  $P(g | y)$ , as under Fearful’s presumption, then some of your moves under Realist’s presumption become impossible. (This is true for almost any  $P(g | y)$  that  $W$  might choose, and in particular even if  $W$  does not predict perfectly.) More precisely, if  $P(g | y)$  is set by  $W$ , then the only way that  $P(y | g)$  can be  $g$ -independent is if it is one of the two delta functions,  $\delta_{y,AB}$  or  $\delta_{y,B}$ . (See the appendix for a formal proof.) This contradicts Realist’s presumption that you can set  $P(y | g)$  to any  $h(y)$  you desire.<sup>3</sup>

Similarly, if  $P(g | y)$  is fixed by  $W$ , as under Fearful’s presumption, then your (Realist) choice of  $h$  affects  $P(g)$ . In fact, your choice of  $h$  fully specifies  $P(g)$ .<sup>4</sup> This contradicts Realist’s presumption that it is  $W$ ’s move that sets  $P(g)$ , independent of you.

Conversely, if you can set  $P(y | g)$  to be an arbitrary  $g$ -independent distribution (as Realist presumes), then what you set it to may affect  $P(g | y)$  (in violation of Fearful’s presumption that  $P(g | y)$  is set exclusively by  $W$ ). In other words, if your having “free will” means what it does to Realist, then you have the power to change the prediction accuracy of  $W$  (!). As an example, if you set  $P(y = AB | g) = 3/4$  for all  $g$ ’s such that  $P(g) \neq 0$ , then  $P(g | y)$  cannot equal  $\delta_{g,y}$ .

The resolution of Newcomb’s paradox is now immediate: You can be free to set  $P(y)$  however you want, with  $P(g | y)$  set by  $W$ , as Fearful presumes, *or*, as Realist presumes, you can be free to set  $P(y | g)$  to whatever distribution  $h(y)$  you want, with  $P(g)$  set by  $W$ . It is not possible to play both games simultaneously.<sup>5</sup>

We emphasize that this impossibility arises for almost any  $P(g | y)$  choice by  $W$ , i.e., no matter how accurately  $W$  predicts. This means that the stipulation in Newcomb’s paradox that  $W$  predicts perfectly is a red herring. (Interestingly, Newcomb himself did not insist on such perfect prediction in his formulation of

---

<sup>3</sup>Note that of the two  $\delta$  functions you can choose in this variant of Newcomb’s scenario, it is better for you to choose  $h(y) = \delta_{y,B}$ , resulting in a payoff of 1,000,000. So your optimal response to Newcomb’s question for this variant is the same as if you were Fearful.

<sup>4</sup>For example, if you set  $h(y) = \delta_{y,AB}$ , then  $P(g) = \delta_{g,AB}$ , and if you set  $h(y) = \delta_{y,B}$ , then  $P(g) = \delta_{g,B}$ .

<sup>5</sup>In a variant of Newcomb’s question, you first choose one of these two presumptions, and then set the associated distribution. If the pre-fixed distribution  $P(g | y)$  arising in the first presumption is  $\delta_{g,y}$ , then your optimal responses depend on the pre-fixed distribution  $P(g)$  arising in the second presumption— a distribution that is not specified in Newcomb’s question. If  $P(g)$  obeys  $P(g = B) > .999$ , then your optimal pair of choices are first to choose to set the distribution  $P(y | g)$  to some  $h(y)$ , and then to set  $h(y) = \delta_{y,AB}$ . If this condition is not met, you should first choose to set  $P(y)$ , and then set it to  $\delta_{y,AB}$ .

the paradox, perhaps to avoid the time paradox problems.) The crucial impossibility implicit in Newcomb’s question is the idea that at the same time you can arbitrarily specify “your” distribution  $P(y | g)$  and  $W$  can arbitrarily specify “his” distribution  $P(g | y)$ . In fact, neither of you two can set your distribution without possibly affecting the other’s distribution; you and  $W$  are inextricably coupled.

Note also that no time variable occurs in our analysis of Newcomb’s paradox. So that analysis is time-reversal invariant. This means that both the paradox and its resolution are unchanged if the prediction occurs *after* your choice rather than before it. This is even the case if the “prediction” algorithm directly observes your choice. See [Wolpert(2008)] for more on the equivalence of observation and prediction and the time-reversal invariance of both.

Newcomb’s paradox has been so vexing that it has led some to resort to non-Bayesian probability theory in their attempt to understand it [Gibbard and Harper(1978), Hunter and Richter(1978)], some to presume that payoff must somehow depend on your beliefs as well as what’s under the boxes [Geanakoplos(1997)], and has even even led some to claim that quantum mechanics is crucial to understanding the paradox [Piotrowski and Sladkowski(2002)]. This is all in addition to work on the paradox based on now-discredited formulations of causality [Jacobi(1993)].

Our analysis shows that the resolution of Newcomb’s paradox is in fact quite simple. Newcomb’s paradox takes two incompatible interpretations of a question, with two different answers, and makes it seem as though they are the same interpretation. The lesson of Newcomb’s paradox is just the ancient verity that one must carefully define all one’s terms.

**ACKNOWLEDGEMENTS:** We would like to thank Mark Wilber for helpful comments.

**APPENDIX:**

In the text, it is claimed that if  $P(g | y)$  is pre-fixed, then the only way that  $P(y | g)$  can be  $g$ -independent is if it is one of the two delta functions,  $\delta_{y,A}$  or  $\delta_{y,B}$ . To see why this is true, combine Eq.’s 1 and 2 of the text to get

$$P(y | g)P(g) = P(g | y)P(y).$$

If for all  $g$  such that  $P(g) \neq 0$ ,  $P(y | g) = h(y)$  for some distribution  $h$ , then we can sum both sides over the two values of  $g$ , getting  $P(y) = h(y)$ . Plugging this back in shows that for any  $y$  such that  $h(y) \neq 0$ ,  $P(g | y)$  must equal  $P(g)$ . If there were two such  $y$ ’s in the support of  $h$ , then  $P(g | y)$  would have to be the same distribution over  $g$  for both of those  $y$ ’s. This is not the case for a perfectly accurate  $P(g | y)$

though (for which  $P(g | y) = \delta_{g,y}$ ), nor is it the case for almost all other  $P(g | y)$ 's. The only way to avoid this contradiction is for you to set  $h(y)$  so that it equals 0 for one of the two  $y$ 's. **QED.**

## References

- [Bar-Hillel and Margalit(1972)] M. Bar-Hillel and A. Margalit. Newcomb's paradox revisited. *British Journal of Philosophy of Science*, 23:295–304, 1972.
- [Benford et al.(1970)Benford, Book, and Newcomb] G. Benford, D. Book, and W. Newcomb. *Physical Review D*, 2:263, 1970.
- [Berger(1985)] J. M. Berger. *Statistical Decision theory and Bayesian Analysis*. Springer-Verlag, 1985.
- [Binder(2008)] P. Binder. Theories of almost everything. *Nature*, 455:884–885, 2008.
- [Campbell and Lanning(1985)] R. Campbell and S. Lanning. *Paradoxes of Rationality and Cooperation: Prisoners' Dilemma and Newcomb's Problem*. University of British Columbia Press, 1985.
- [Collins(2001)] J. Collins. Newcomb's problem. In *International Encyclopedia of the Social and Behavioral Sciences*. Elsevier Science, 2001.
- [Fudenberg and Tirole(1991)] D. Fudenberg and J. Tirole. *Game Theory*. MIT Press, Cambridge, MA, 1991.
- [Gardner(1974)] M. Gardner. Mathematical games. *Scientific American*, page 102, 1974.
- [Geanakoplos(1997)] J. Geanakoplos. The hangman's paradox and newcomb's paradox as psychological games. Yale Cowles Foundation paper, 1128, 1997.
- [Gibbard and Harper(1978)] A. Gibbard and W. Harper. Counterfactuals and two kinds of expected utility. In *Foundations and applications of decision theory*. D. Reidel Publishing, 1978.
- [Hunter and Richter(1978)] D. Hunter and R. Richter. Counterfactuals and newcomb's paradox. *Synthese*, 39:249–261, 1978.
- [Jacobi(1993)] N. Jacobi. Newcomb's paradox: a realist resolution. *Theory and Decision*, 35:1–17, 1993.

- [Koller and Milch(2003)] D. Koller and B. Milch. Multi-agent influence diagrams for representing and solving games. *Games and Economic Behavior*, 45:181–221, 2003.
- [Levi(1982)] I. Levi. A note on newcombmania. *Journal of Philosophy*, 79: 337–342, 1982.
- [Myerson(1991)] Roger B. Myerson. *Game theory: Analysis of Conflict*. Harvard University Press, 1991.
- [Nozick(1969)] R. Nozick. Newcomb’s problem and two principles of choice. In *Essays in Honor of Carl G. Hempel*, page 115. Synthese: Dordrecht, the Netherland, 1969.
- [Osborne and Rubenstein(1994)] M. Osborne and A. Rubenstein. *A Course in Game Theory*. MIT Press, Cambridge, MA, 1994.
- [Pearl(2000)] J. Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2000.
- [Piotrowski and Sladkowski(2002)] E. W. Piotrowski and J. Sladkowski. Quantum solution to the newcomb’s paradox. <http://ideas.repec.org/p/sla/eakjkl/10.html>, 2002.
- [Wolpert(2008)] D. H. Wolpert. Physical limits of inference. *Physica D*, 237: 1257–1281, 2008. More recent version at <http://arxiv.org/abs/0708.1362>.

	<u>Choose AB</u>	<u>Choose B</u>
<b>Predict AB:</b>	1000	0
<b>Predict B:</b>	1,001,000	1,000,000

**Table 1: The payoff to you for the four combinations of your choice and  $W$ 's prediction.**

**Short title:** Newcomb's paradox resolved