The Lesson of Newcomb's Paradox

David H. Wolpert
Gregory Benford

SFI WORKING PAPER: 2011-08-032

SANTA FE INSTITUTE

The Lesson of Newcomb's Paradox

David H. Wolpert[1] & Gregory Benford[2]

1 - Santa Fe Institute, 1399 Hyde Park Road Santa Fe, NM 87501
Center for Nonlinear Studies, MS B258, LANL, Los Alamos, NM 87545
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

Abstract

In Newcomb's paradox you can choose to receive either the contents of a particular closed box, or the contents of both that closed box and another one. Before you choose though, an antagonist uses a prediction algorithm to accurately deduce your choice, and uses that deduction to fill the two boxes in a way that lessens the value of your choice. Newcomb's paradox is that game theory's expected utility and dominance principles appear to provide conflicting recommendations for what you should choose. Here we show that the conflicting recommendations assume different probabilistic structures relating your choice and the algorithm's prediction. This resolves the paradox: the reason there appears to be two conflicting recommendations is that the probabilistic structure relating the problem's random variables is open to two, conflicting interpretations. We then show that the accuracy of the prediction algorithm in Newcomb's paradox, the focus of much previous work, is irrelevant. We end by showing 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.

1 Introduction

1.1 Background

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 if a prediction algorithm used by the being had said that you would take only B. If instead the algorithm had predicted you would take both boxes, then *W* put nothing in box B.

Presume that due to determinism, there exists a perfectly accurate prediction algorithm, and assume that it is this perfect prediction algorithm that *W* uses. Suppose further that when you choose which boxes to take, you do not know the prediction of that algorithm. *W*hat should your choice be?

In Table 1 we present this question as a game theory matrix involving *W*'s prediction and your choice. Two seemingly logical answers to your question contradict each other. The Realist answer is that you should take both boxes, because your choice occurs after *W* has already made its prediction, and since you have free will, you are free to make whatever choice you want, independently of that prediction that *W* made. 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.

In contrast, the Fearful answer 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 the conventional formulation of Newcomb's Paradox, a famous logical riddle stated by William Newcomb in 1960 [**?**, **?**, **?**, **?**, **?**, **?**, **?**, **?**, **?**, **?**, **?**]. Newcomb never published the paradox, but had long conversations about it with 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. We often discussed the paradox, which Newcomb invented to test his own ideas. Newcomb 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" [**?**].

It was Nozick who pointed out that two accepted principles of game theory appear to conflict in Newcomb's problem. The expected-utility principle, considering the probability of each outcome, says you should take box B only. But the dominance principle argues that if one strategy is always better than the other strategies no matter what other players do, then you should pick that strategy [**?**, **?**, **?**].) No matter what box B contains, you are $1000 richer if you take both boxes than if you take B only. So the dominance principle says you should take both boxes.

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., for predictions made in the present to be perfectly determined by events in the future means that the future causes past events.[1]

But Nozick found a way to restate the problem specifically to exclude backward causation (and so time travel). To achieve this his formulation demands only that the predictions be of high accuracy, not perfect. So arguments about time travel cannot resolve the issue.

1.2 Our solution

To properly define a noncooperative game, there are several things one must specify. First, one must specify the set of statistical independencies relating the variables in the game.[2] Having done this leaves free the conditional distributions relating the non-independent variables. These conditional distributions are called the "player strategies", and are what the players of the game will set. So the second thing one must specify is which players of the game set which of those conditional distributions, and any restrictions on how they may do so. We refer to this pair of specifications as the "probabilistic structure" of the game.[3]

In this paper we show that Newcomb's scenario does not fully specify the

[1] 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 [**?**]. The issues of time travel and Newcomb-style paradoxes are intertwined.

[2] Formally, such a set of statistical independencies is known as a Bayes net [**?**].

[3] There are other things that must also be specified to fully specify a game, e.g., the utility functions. However for to elucidate the illusory nature of Newcomb's paradox it suffices to focus on the probabilistic structure. See [**?**, **?**] for full and formal discussions of how to define any noncooperative game by using a probabilistic structure.

probabilistic structure underlying the game you and W are playing. The two "conflicting principles of game theory" actually correspond to two different probabilistic structures, i.e., two different games. So there is no conflict of game theory principles in Newcomb's paradox — simply imprecision in specifying the probabilistic structure of the game you and W are playing. Once that probabilistic structure is fully specified, the game is fully specified. And once the game is fully specified, your optimal choice is perfectly well-defined, and the paradox is resolved.

After establishing this we go on to show that the accuracy of the prediction algorithm in Newcomb's paradox, the focus of much previous work, 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.

The analysis of Newcomb's paradox we present here only involves decision theory, without any explicit consideration of game theoretic issues. However much of the literature considers Newcomb's paradox from the point of view of game theory. In particular, the way Newcomb's question is usually phrased suggests that somehow the two possible probabilistic structures underlying Newcomb's scenario, corresponding to the two kinds of reasoning, can be combined into one probabilistic structure, i.e., into one game. While not required, our analysis can be naturally embedded in a broader game theoretic framework to analyze combination of games involving multiple probabilistic structures [**?**]. Doing this shows that any combining of the two particular probabilistic structures that underlie Newcomb's scenario is mathematically impossible.

More generally, *however* one formalizes Newcomb's scenario in terms of a single probabilistic structure, once one has made that formalization, there is no room for paradox. Bayesian decision theory fully specifies the optimal decision for any properly specified single set of conditional independencies. The only issue that can be "debated" is how to translate the vague English with which Newcomb phrased his scenario into mathematics; there is no formal paradox.

In the next section we present a simplified version of our full argument. In the subsequent section we provide our fully detailed resolution of Newcomb's paradox. We end with a discussion.

2 Probabilistic Structures Assumed in the Reasoning of Fearful and Realist

There are two players in Newcomb's paradox: you, and the wise being W. In addition, there are two game variables that are central to the paradox: W's prediction, g, and the choice you actually make, y. So the probabilistic structure will involve the joint probability distribution relating those two variables. Since there are only two variables, there are only two ways to decompose that joint probability distribution. 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.

Define ab as the event that W predicts you will take both boxes, and b as the event that W predicts you will only take box B. Similarly define AB as the event that you actually take both boxes, and B as the event that you actually take only box B. So $P(g = ab \mid y = AB) \equiv P(ab \mid AB)$ is the probability that W predicts correctly, given that you choose AB. Similarly $P(b \mid B)$ is the probability that W predicts correctly given that you choose only B.

Von Neumann-Morgenstern expected utility theory says that whatever the probabilistic structure underlying the game, your goal is to maximize

$$\begin{aligned} 1000[P(ab, AB)] \;&+\; 1001000[P(b, AB)] \\ &+\; 0[P(ab, B)] \;+\; 1000000[P(b, B)] \end{aligned} \tag{1}$$

Where Fearful and Realist differ is how they decompose those joint probabilities. We discuss those two decompositions next.

2.1 The Reasoning of Fearful

The first decomposition of the joint probability distribution over g and y is

$$P(y, g) = P(g \mid y)P(y)$$

where we define the right-hand side to equal 0 for any y such that $P(y) = 0$.

This decomposition can be used to express Fearful's reasoning. $P(g \mid y)$ is the accuracy of W's prediction algorithm. Fearful interprets the statement that 'W designed a perfectly accurate prediction algorithm' to imply that W has the power to set the conditional distribution $P(g \mid y)$, to anything it wants (for all y

such that $P(y) \neq 0$). More precisely, since the algorithm is 'perfectly accurate', Fearful assumes that W chooses to set $P(g \mid y) = \delta_{g,y}$, the Kronecker delta function that equals 1 if $g = y$, zero otherwise. So Fearful assumes that there is nothing you can do that can affect the values of $P(g \mid y)$ (for all y such that $P(y) \neq 0$). Instead, Fearful assumes that you get to choose the unconditioned distribution $P(y)$. (Intuitively, this choice constitutes your 'free will'.) This is the probabilistic structure of Fearful.

Decision theory says that you should make your choice so as maximize the associated expected utility. For the probabilistic structure of Fearful, this means that you should set $P(AB)$ (and therefore $P(B) = 1 - P(AB)$) so as to maximize

$$\begin{aligned} 1000[P(ab \mid AB)P(AB)] \; &+ \; 1001000[P(b \mid AB)P(AB)] \\ &+ \; 0[P(ab \mid B)P(B)] \; + \; 1000000[P(b \mid B)P(B)] \end{aligned} \tag{2}$$

Provided that the associated distributions $P(ab \mid AB)$ and $P(b \mid B)$ are large enough — provided W's prediction algorithm is accurate enough — to achieve this maximization you should set $P(B) = 1$. In other words, you should choose to take only B. This expresses Fearful's "expected utility" reasoning.

2.2 The Reasoning of Realist

The second way to decompose the joint probability is

$$P(y, g) \;\; = \;\; P(y \mid g)P(g) \tag{3}$$

where we define the right-hand side to equal 0 for any g such that $P(g) = 0$.

This decomposition can be used to express Realist's reasoning. Realist interprets the statements that 'your choice occurs after W has already made its prediction' and 'when you have to make your choice, you do not know what that prediction is' to mean that you can choose any distribution $h(y)$ and then set $P(y \mid g)$ 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 slightly different interpretation of 'free will" from the one made by Fearful, which instead concerns your setting the distribution $P(y)$.) Under this interpretation, W has no power to affect $P(y \mid g)$. Rather W gets to set $P(g)$. 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.)

Formally, in Realist's reasoning we have

$$\begin{aligned} P(y, g) &= P(y \mid g)P(g) \\ &= h(y)P(g) \end{aligned}$$

Accordingly, $P(y) = h(y)$, and therefore $P(y, g) = P(y)P(g)$. For this probabilistic structure, decision theory says that you should choose $P(AB) = h(AB)$ to maximize

$$\begin{aligned} 1000[P(ab)P(AB)] &+ 1001000[P(b)P(AB)] \\ &+ 0[P(ab)P(B)] + 1000000[P(b)P(B)]. \end{aligned}$$

which is achieved by setting $P(B) = 0$, no matter what $P(g)$ is. This conclusion expresses the dominance principle of game theory.

Note that by Bayes' theorem,

$$P(y \mid g) = \frac{P(y, g)}{\sum_{y'} P(y', g)}$$

whereas $P(y) = \sum_{g'} P(y, g')$. Combining this with the equations of the previous subsection, it is straight-forward to verify that under Fearful's reasoning, using a prediction algorithm $P(g \mid y)$ that is at least moderately accurate, $P(y \mid g) \neq P(y)$. This means that your choice y is statistically dependent on W's prediction g in Fearful's reasoning. On the other hand, under Realist's reasoning, $P(y \mid g) = P(y)$, and you can make your decision entirely independently of W's prediction

This foregoing analysis illustrates how the reasoning of Fearful and Realist correspond to different decompositions of the joint probability of your choice and W's choice. The simple fact that those two decompositions differ is what underlies the resolution of the paradox — you can state Newcomb's scenario in terms of Fearful's reasoning, or in terms of Realist's reasoning, but not both.[4]

3 Discussion

In this section we cursorily mention some implications of our analysis.

[4] We are grateful to an anonymous referee for emphasizing to us the underlying simplicity of our analysis of Newcomb's scenario.

3.1 Irrelevance of Accuracy of Prediction Device

Consider the probabilistic structure underlying Fearful's reasoning, with the modification that the algorithm $P(g \mid y)$ used by the wise being is not perfect, but instead has an error rate α. Such error would mean is that the values in Eq. 2 would change, and therefore *if one adopts Fearful's probabilistic structure*, so might the optimal decision. But no paradox of any sort would ensue; the optimal decision is perfectly well-defined.

On the other hand, $P(g \mid y)$ does not even arise in the probabilistic structure underlying Realist's reasoning. (Under Realist's reasoning, y and g are statistically independent, and you should always choose $h(y) = \delta_{y,AB}$.) So changing the accuracy of the prediction algorithm does not affect Realist's reasoning. Again, no paradox ensues by introducing error into the algorithm used by the wise being.

This shows that just as when the wise being's algorithm is perfect, when that algorithm is noisy, the only "issue" is whether to formalize Newcomb's scenario with Fearful's or Realist's probabilistic structure. Once the formalization is made, decision theory gives a single, well-defined optimal choice. So 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 the paradox, perhaps to avoid the time paradox problems.)

3.2 Irrelevance of Causality and Free Will

We emphasize that vague concepts (e.g., 'free will') or controversial ones (e.g., 'causality') are not relevant to the resolution of Newcomb's paradox; it is not necessary to introduce mathematizations of those concepts to resolve Newcomb's paradox. The only mathematics needed is standard probability theory, together with the axioms of decision theory.

3.3 Lack of Conflict between Dominance and Expected Utility Principles

Indeed, our using those axioms shows that Newcomb's scenario does not demonstrate a conflict between game theory's dominance principle and its expected utility principle, as some have suggested. The key behind our avoiding a conflict between those two principles is our taking care to specify what probabilistic structure underlies the game. Realist's reasoning *appears* to follow the dominance principle, and Fearful's to follow the principle of maximizing expected utility. (Hence the conflicting answers of Realist and Fearful appear to illustrate a conflict between those two principles.) However Realist is actually following the principle of maximizing expected utility *for that probabilistic structure for which Realist's answer is correct*. In contrast, Realist's reasoning is an unjustified violation of that principle for the probabilistic structure game in which Realist's answer is incorrect. (In particular, Realist's answer does *not* follow from the dominance principle in that probabilistic structure.) It is only by being sloppy in specifying the underlying probabilistic structure that it appears that there is a conflict between the expected utility principle and the dominance principle.

3.4 Irrelevance of Temporal Sequence of Decisions

Another note-worthy point is that no time variable occurs in our analysis of Newcomb's paradox. Concretely, nothing in our analysis using probabilistic structures requires that W's prediction occur before your choice. Statistical (in)dependence is an atemporal concept. Indeed, for any given probabilistic structure over a set of random variables, there are multiple ways of trying to assign a temporal sequence to the variables [**?**]. Broadly speaking, this lack of a temporal aspect to our analysis is a manifestation of the fact that statistical correlation (which for us means the probabilistic structure) does not imply causation (which for us means a temporal sequence of the variables).

This lack of an inherent temporal aspect to our analysis means we can assign times to the events in our analysis any way we please, and the analysis still holds; the analysis is time-reversal invariant. We can use this invariance to clarify the differences in the assumptions made by Realist and Fearful.

The temporal invariance means that both the formal statement of the paradox and its resolution are unchanged if the prediction occurs *after* your choice rather than (as is conventional) before your choice. In other words, W could use data accumulated up to a time *after* you make your choice to 'retroactively predict' what your choice was. In the extreme case, the 'prediction' algorithm could even directly observe your choice. All of the mathematics introduced above concerning probabilistic structures, and possible contradictions still holds.

In particular, in this time-reversed version of Newcomb's scenario, Fearful would be concerned that W can *observe* his choice with high accuracy. (That's what it means to have $P(g \mid y)$ be a delta function whose argument is set by the value of y.) Formally, this is exactly the same as the concern of Fearful in the conventional Newcomb's scenario that W can *predict* his choice with high accuracy.

In contrast, in the time-reversed version of Newcomb's scenario, Realist would believe that he can guarantee that his choice is independent of what W says he chooses. (That's what it means to have $P(y \mid g)$ equal some $h(y)$, independent of g.) In essence, he assumes that you can completely hide your choice from W, so that W can only guess randomly what choice you made. Formally, this assumption is exactly the same as the 'free will' belief of Realist in the conventional Newcomb's scenario that you can force W's prediction to be independent of your choice.

In this time-reversed version of Newcomb's scenario, the differences between the assumptions of Realist and Fearful are far starker than in the conventional form of Newcomb's scenario, as is the fact that those assumptions are inconsistent. One could use this time-reversed version to try to argue in favor of one set of assumptions or the other (i.e., argue in favor of one probabilistic structure game or the other). Our intent instead is to clarify that Newcomb's question does not specify which probabilistic structure game is played, and therefore is not properly posed. As soon as one or the other of the probabilistic structure games is specified, then Newcomb's question has a unique, correct answer.

3.5 Extensions of our Framework

Some readers might object that we have not proven that there are no possible formalizations of Newcomb's scenario in terms of probabilistic structures and associated games beyond the two formalizations we have analyzed. One might suspect that it is possible to "merge" those two formalizations somehow. However it turns out that this is impossible. You cannot arbitrarily specify 'your' distribution $P(y \mid g)$, as in Realist's reasoning while W arbitrarily specifies 'his" distribution $P(g \mid y)$ as in Fearful's reasoning. In fact, neither of you two can set your distribution arbitrarily, without possibly affecting the other's distribution; you and W are inextricably coupled. (For a detailed discussion of this see [**?**], which analyzes Newcomb's scenario using a broad framework concerning games played over probabilistic structures.)

More generally still, our analysis of prediction / observation in Newcomb's scenario is a special case of a more general and elaborate analysis of such processes and how they can transpire in the real world [**?**, **?**, **?**]. This analysis shows that there are unavoidable fallibilities in both processes. (Those fallibilities can be viewed as variants of the uncertainty relation of quantum mechanics, variants that are both somewhat weaker and more broadly applicable.) However none of that more elaborate analysis is needed to resolve Newcomb's paradox.

4 Conclusion

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 [**?**, **?**], some to argue that payoff must somehow depend on your beliefs as well as what's under the boxes [**?**], and has even even led some to claim that quantum mechanics is crucial to understanding the paradox [**?**]. This is all in addition to work on the paradox based on now-discredited formulations of causality [**?**].

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 *W*e would like to thank Mark *W*ilber, Charlie Strauss and an anonymous referee for helpful comments.

	Choose *AB*	**Choose *B***
Predict *ab*:	1000	0
Predict *b*:	$1,001,000$	$1,000,000$

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