Newcomb's Paradox: a Matter of Assumptions

Geert Jonker

September 2003

Abstract

One of the most intriguing decision paradoxes is without a doubt the paradox of Newcomb. It's been the subject of an ongoing debate since its first appearance in 1969 and the differences in views between the opponents have been far from settled. We introduce the paradox and summarize the different solutions that have been given until now. We then show that the solution to the paradox strongly depends on certain assumptions that are fundamental to the problem. We discuss the solutions to seven scenarios that are based on different assumptions.

1 Newcomb's Paradox

In 1960, the physicist William A. Newcomb thinks of an interesting paradox while meditating on a famous paradox of game theory called the prisoner's dilemma. A few years later, the paradox reaches the philosopher Robert Nozick, who can't resolve the problem but decides to write it up anyway. "It is not clear that I am entitled to present this paper," Nozick writes in the now famous article "Newcomb's Problem and Two Principles of Choice" [Noz69]. "It is a beautiful problem. I wish it were mine." We give the paradox as paraphrased by Martin Gardner [Gar86].

> There are two closed boxes on the table. The first box contains \$1,000. The second box contains either \$1 million or no money at all. You have a choice between two actions: 1) taking what is in both boxes; or 2) taking just what is in the second box.
>
> Imagine a Being that can predict your choices with high accuracy. You can think of this Being as a genie, or a superior intelligence from another planet, or a supercomputer that can

		Being predicts one box	Being predicts both boxes
you choose	second box	$1,000,000	$0
	both boxes	$1,001,000	$1,000

Figure 1: Payoff matrix for Newcomb's game

> scan your mind, or God. He has correctly predicted your choices in the past, and you have enormous confidence in his predictive powers. Yesterday, the Being made a prediction as to which choice you are about to make, and it is this prediction that determines the contents of the second box. If the Being predicted that you will take what is in both boxes, he put nothing in the second box. If he predicted that you will take only what is in the second box, he put $1 million in it. You know these facts, he knows you know them, etc. So, do you take both boxes, or only the second box?

To answer the question, one can use two plausible arguments, leading to two different decisions. To see how these arguments apply, consider the payoff matrix for Newcomb's game in figure 1.

1. The expected-utility argument. If you take what is in both boxes, the Being almost certainly will have predicted this and will not have put the $1 million in the second box. Almost certainly you will get only $1,000. If you take only what is in the second box, the Being almost certainly will have predicted this and put the money there. Almost certainly you will get $1 million. Suppose the probability of the Beings prediction being correct is 90%. The expected utility of the one-box strategy is

$$(.9 \times \$1,000,000) + (.1 \times \$0) = \$900,000.$$

The expected utility of the two-box strategy is

$$(.1 \times \$1,001,000) + (.9 \times \$1000) = \$101,000.$$

The one-box strategy has a higher expected utility. Therefore you should take only what is in the second box.

2. The dominance argument. The Being has already made his prediction and has either put the $1 million in the second box or has not. The money is either sitting in the second box or is not. The situation, whichever it is, is fixed and determined. If the Being put the $1 million in the second box, you will get $1,001,000 if you take both boxes and $1 million if you take only the second box. If the Being did not put the money in the second box, you will get $1,000 if you take both boxes and no money if you take only the second box. In either case you will do better by $1,000 if you use the two-box strategy rather than the one-box strategy.

"I have put this problem to a large number of people, both friends and students in class," writes Nozick. "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.

"Given two such compelling opposing arguments, it will not do to rest content with one's belief that one knows what to do. Nor will it do to just repeat one of the arguments, loudly and slowly. One must also disarm the opposing argument; explain away its force while showing it due respect."

The scientific debate on Newcomb's paradox has essentially followed the students example; Scientists are sharply divided on the problem. Despite Nozick's advices, more than once arguments have been repeated loudly and slowly. In the next section we will discuss the most important arguments that played a role in the debate.

2 Solutions

Since its first appearance in 1969, Newcomb's paradox has stirred up many reactions. Nozick alone received 148 letters of people reacting to a column in *Scientific American*. A great number of articles on the subject has been published since and the debate doesn't seem to be settled yet. Among the most important reactions are:

- If the Being was God, I would take the second box. If there was only the slightest chance of misprediction, I would take both boxes.
- Although it seems to be more intuitive to take only the second box, I cannot force myself to act irrationally, so I take both boxes.
- I will do something the Being won't expect.

- The situation is 'isomorphic' with one in which I choose first and openly, after which the Being puts the money in the boxes.
- The second box doesn't actually contain either $1,000,000 or $0 until I have made my decision.
- The dominance argument fails because the choices are not probabilistically independent from the prediction.
- The strategy of taking the second box only is self-contradictory.
- The problem has two equally strong but contradictory solutions and is therefore logically impossible.
- The existence of a perfect predictor is logically impossible.
- The description of the problem is insufficient to enable me to make my decision.

The central principle to which most solutions are aimed is the seemingly cyclic relation between the prediction and the choice that is made. At one hand the choice I make seems to determine the prediction the Being has made, while on the other hand I would like to be able to act contrary to the prediction if that is beneficial to me. These two principles equally depend on each other, forming a cyclic relationship. Many attempts have been done to 'break' the cycle, of which we will give the three most important ones.

The dominance argument fails because the choices are not probabilistically independent from the states. This objection is based on the fact that the two states "The $1 million is in the second box" and "The $1 million is not in the second box" are not probabilistically independent of the actions "Take only the second box" and "Take both boxes". According to the objectors, if this is true, the dominance argument can not be applied. To model the states as being probabilistically independent of the choices, Ferejohn proposes to use the states "The Being predicts correctly" and "The Being predicts incorrectly" [Cra87]. In the corresponding matrix, shown in figure 2, the dominance argument no longer holds. It appears that the force of the dominance principle is undercut.

Nozick replies to this argument with a story showing that the dominance strategy can apply even if the states are probabilistically dependent of the choices [Noz86].

		Being predicts your choice correctly	Being predicts your choice incorrectly
you choose	second box	$1,000,000	$0
	both boxes	$1,000	$1,001,000

Figure 2: The Ferejohn payoff matrix.

Suppose there is a hypochondriac who knows that either man S or man T is his father, but he does not know which. S died of some very painful inherited disease that strikes in one's middle thirties, and T did not. The disease is genetically dominant. S carried only the dominant gene. T did not have the gene. If S is his father, the person will die of the dread disease. If T is his father, he will not. Furthermore, suppose there is a well-confirmed theory that states that a person who inherits this gene will also inherit a tendency towards behaviour that is characteristic of intellectuals and scholars. The person is now deciding whether to go to graduate school or to become a professional baseball player. He prefers (although not enormously) the life of an academic to that of a professional athlete. Regardless of whether or not he will die in his middle thirties, he would be happier as an academic. The choice of the academic life would thus appear to be his best choice.

Now suppose this hypochondriac reasons that if he decides to be an academic, the decision will show that he has such a tendency, and therefore it will be likely that he carries the gene for the disease and so will die in his middle thirties, whereas if he chooses to become a baseball player, it will be likely that T is his father; therefore he is not likely to die of the disease. Since he very much prefers not to die of the disease, he decides to pursue the career of an athlete. Surely everyone would agree that this reasoning is perfectly wild. It is true that the conditional probabilities of the states "S is his father" and "T is his father" are not independent of the actions "becoming an academic" and "becoming a professional athlete". If he does the first, it is very

> likely that S is his father and that he will die of the disease; if he does the second, it is very likely that T is his father and therefore unlikely that he will die of the disease. But who his father is cannot be changed. It is fixed and determined and has been for a long time. His choice of how to act legitimately affects our (and his) estimate of the probabilities of the two states, but which state obtains (which person is his father) does not depend on his action at all. By becoming a professional baseball player he is not making it less likely that S is his father; therefore he is not making it less likely that he will die of the disease.

Kiekeben observes the fact that in this story, there is a causal relation between the states and the choice which is absent in Newcomb's problem [Kie99]. While the presence of the gene makes the hypochondriac more likely to choose to become an academic, the prediction of the Being doesn't cause you to choose in a certain way.

In our view, the concept of genetic causation is misused in Nozick's story. The effect of carrying a gene should be understood as the hypochondriac being likely to become an academic if he doesn't know of the presence of the gene. Although not impossible, the effects that knowing or considering its presence have on his choice are clearly not meant to be part of the genetic causation[1]. In fact, this should rather be considered as a second causation, effecting the same event as the first one. The second causation can overrule the first one, as in the example where the hypochondriac decides to become a baseball player. Its effect could also be absent or to weak to overrule the first causation, which would be the case if the hypochondriac decides to forget his 'clever' reasoning and follow his heart. It is only in this second instance that the choice the hypochondriac makes will show whether he carries the gene or not. In the first instance, the first causation is overruled by the second one; the first causation is eliminated and only the second one is now in effect. Since the second causation is not caused by the presence of the gene but rather by a self-referential mental process, the presence of the gene and the decision have even become probabilistically independent. In short, as soon as the hypochondriac starts to *think* about what to decide, the the presence of the gene and the decision become probabilistically independent. This invalidates Nozick's counterexample.

[1]The example wouldn't be relevant if the gene also caused the 'clever' reasoning of the hypochondriac. Suppose it did. Then, as soon as the hypochondriac becomes aware of his own reasoning, it would know that it carried the gene and further choices wouldn't matter anymore.

The content of the second box is caused by your decision. Advocates of this view claim that causality is independent of the direction of time. This enables the prediction to be caused by the choice you make. Since the contents of the second box are caused by the prediction that was made, your choice eventually causes the box to be filled or not. We will refer to this principle as *backward causality.*

A similar approach is the suggestion that the second box doesn't contain either \$0 or \$1,000,000, but that there is an amplitude that the box contains \$0 and an amplitude that the box contains \$1,000,000. These amplitudes interfere unless and until you make your move and open the box. To assert that 'either the second box contains \$1 million of else it is empty' is an intuitive argument for which there is no evidence unless you open the box[2].

Although the principle of backwards causality was ruled out in Nozick's original paper, both he and Gardner mention that the Being can just as well be identified as God. In this essay however, we consider the paradox both in the case where backward causation is in operation and where it is not. We will associate God with backward causality since we chose to categorize Him as someone existing out of time, which implies backwards causality.

To show that backwards causality is not operating, Nozick presents a variation of the scenario making the argument to take the two boxes even more vivid. Suppose the far side of the second box is transparent, so that you can't see its contents but a friend of you sitting opposite can look inside it. He will see all the money that's on the table. You realize that, although he says nothing, he wants you to take both boxes. He would not understand it if you only take the second box. As Schlesinger puts it, the one box strategy can't be the right one, since it would then not be in my best interest to follow the advice of a friend who advises me to do what is in my best interest, which is self-contradictory [Sch74].

In yet another variation of the game, both boxes are transparent. The first box contains the usual \$1,000, but the second box contains a piece of paper with a fairly large integer written on it. You do not know whether the number is prime or composite. If it proves to be prime (you must not test it, of course, until after you have made your choice), the you get \$1 million. The Being has chosen a prime number if he predicts you will take only the second box but has picked a composite number if he predicts you will take

[2] An even more extreme approach is that of solipsism, in which reality is only created by your observation. The physicist Fred A. Wolf defended this approach in [Wol81]. We leave this intriguing approach for what it is, since its implications seem to us too far stretching to cover in this essay.

both boxes.

These scenarios are also meant to disarm those who claim that backwards causality is in operation. Surely you cannot by an act of will make the large number change from prime to composite, or vice versa. The nature of the number is fixed for eternity. So why not take both boxes?

The advocates of backwards causality need not necessarily feel disarmed. They can point out that your choice affects what term the predictor writes down (or wrote down earlier), not whether the integer it designates is prime or composite. It is true that what term the predictor writes down causes you to see a prime or not, but at least your choice does not directly affect the number you see to be a prime or not. It is via the predictor that you have influence on the number you see. Of course it is still unintuitive to have, even indirectly, any influence on the nature of the number you see. This is one of the consequences one has to accept if one assumes backwards causality.

The existence of a perfect predictor is logically impossible. Gardner argues that a contradiction arises when a prediction is allowed to interact with the predicted event. Note that in this view, backward causality is not in operation. To see how the contradiction arises, consider the example of a psychic knowing whether you will take off your shoes before you go to bed next Thursday. As soon as he informs you of the prediction, you are able to do the opposite and thereby falsifying the prediction. Therefore the psychic cannot predict whether you will take off your shoes or not.

A second example involves a supercomputer that, given the current state of the world, is able to predict the future. If the computer is asked if a certain event will happen in the next three minutes, it will make a prediction of the event. If the prediction is no, it turns on a green light. If yes, it turns on a red light. The computer is now asked to predict whether the green light will go on. The reader will understand that the computer can't give an answer to this. By making the event part of the prediction, the computer is rendered logically impotent. Gardner goes as far as saying that the computer, like the barber who shaves those who do not shave themselves, cannot exist.

Gardner and Nozick arrive at the conclusion that a perfect predictor cannot exist. If confronted with the situation, they would therefore take both boxes. Although the one box strategy is more intuitive, they cannot force themselves to act irrationally. Funny enough, they do however regret the fact that they are likely to end up with only $1,000!

Kiekeben claims that the interaction between prediction and event actually not exists in Newcomb's paradox [Kie96]. The Being doesn't inform

you of the prediction and therefore doesn't influence your decision.

We should add that on the other hand, the Being has informed you of the fact that he has made a prediction. That fact alone can influence your decision. Not to such an extend that you can adjust your decision to his prediction though, but still it could affect the the process of your decision. Also, interaction between the prediction and the decision is clearly the case in the prime number scenario. Although the Being hasn't informed you of the decision he's made, he has given you a piece of evidence of it, which meaning may be checked after the decision. Just as you might keep your shoes on if the Being has predicted you would take them off, you could accidentally make the decision of which the number you are looking at is not the evidence. This is the same kind of interaction which rendered the computers existence impossible.

In our view, what Gardner does is giving two examples of situations in which certain predictions cannot be made. We agree on his conclusion that the supercomputer cannot exists, since it cannot predict whether its green light will go on. We also agree on the fact that a psychic that always correctly predicts and tells you whether you will take your shoes of or not cannot exist. But we do not agree on the fact that a predictor cannot knowingly make a correct prediction.

Suppose the psychic knows that if he predicts that you will take off your shoes, you will actually do it, maybe because of your obedient nature. He can in that case safely make this particular prediction. In other words, this psychic knows when he can make a prediction and when not. If a prediction would cause the predicted event to be otherwise than predicted, this is simply a prediction that cannot be done.

Similarly, the Being in the prime number story could have predicted that if he would write down prime number x, and tell you the rules of the game, whether you would choose the second box or both boxes. If he predicted that in that case you would take the second box, he could safely write down prime number x and begin the game. If he predicted that you would choose both boxes, he knows number x is not an option for him. He could then instead go on to 'test' another number y, until he finds a number that works. Of course, it's not certain that he will find a number. But, as the story goes, we are faced with a Being claiming he has written down a number that is a prime if we choose the second box and a composite if we choose both. Apparently, he has found a number that works, with which the situation he puts you in leads you to make the corresponding choice.

Obviously, if a perfect prediction can be made in the prime number story, it can also be made in the original problem, where the interaction between

prediction and event is less strong. Therefore we agree with Gardner's and Nozick's reasoning, but feel compelled to precise their conclusion: a predictor that can predict the truth of any event in any given situation cannot exist. A predictor giving correct predictions when he knows he is able to, can exist.

We will shortly say something about some other solutions as well. One of the people responding to Nozick's column described that he would walk in with a device to scan the contents of the boxes, take the boxes with the money in them and never open an empty box. "He quite naturally succeeded in getting all the money, for the rule of bridge that one peek is worth two finesses applies here too... By introducing a choice which the Being has not anticipated, and is not permitted to take into account, he achieves a stunning victory for free will." This solution is ambiguous because it is not at all clear how much money he will gain by it. Besides, it clearly has a low esteem of the predicting abilities of the Being; he plans to do something else than what the Being predicted. In general, this and other forms of cheating (like doing nothing) are prohibited by the rules of the game, or in one case by armed guards who apparently will take appropriate measures if one tries to cheat [Pou88].

Another possible solution is the suggestion that the problem as described by Newcomb 'effectually the same as' one where you move first and an observer attempts to communicate with a 'mind reader' in the next room who then guesses your choice, using a payoff matrix identical with Newcomb's. It has even been suggested that the situation is 'isomorphic' with one in which the human moves first and openly. These suggestions effectively remove the sting of prediction/event interaction from the story, but it is not at all clear whether these scenarios are really equivalent. An advocate of this approach should convincingly prove this equivalence and since the problem involves so many issues, this might prove to be a hard task.

One of the earliest responses what that of Bar-Hillel and Margalit [BHM72]. They argue that, although backward causality is not operating, since there is a remarkably high correlation between the presence of the $1 million and the second box option, you should act as if backward causality did in fact exist. You must resign yourself to the fact that your best strategy is to behave *as* if the Being has made a correct prediction, even though you know there is a slight chance he has erred. Similarly they state that "the facts really imply that there is no free choice, but the illusion of free choice remains, and one has to behave as if free choice exists." As a result, Bar-Hillel and Margalit strongly urge you to "join the millionaires club" by taking only the second box.

The controversial relationship between the choice and the predication was identified by Levi as the culprit [Lev82]. He claimed that the problem laid out by Newcomb implied that

> (1) the probability is high that you will choose both boxes if the Being will so predict

but that this did not imply that

> (2) the probability is high that if you will choose both boxes, then the Being will so predict.

Levi grants that we should choose the second box if (2) is true, and both if not. But since the original problem doesn't specify whether (2) holds, you are unable to decide on your strategy. Because of this Levi has been characterized as a 'no-boxer'. We think Levi's objection is formally correct, but far-fetched. Imagine the situation where the Being has already performed the test a thousand times before. Every time his prediction was a random choice between the one box and the two box choice. For some reason, the humans in the test every time chose both boxes. Indeed in this situation, (1) is true but (2) is not. This is not however in line with the way the paradox is usually interpreted. Horgan asserts that he construes the Newcomb situation to involve implicitly these conditions: (i) that almost all of those who have chosen both boxes in the past have received $1,000; (ii) that almost all of those who have chosen only the second box have received $1,000,000; and (iii) that the agent knows these facts [Hor85]. As Craig observes, this is the reasonable and natural way to construe the problem because only then do the paradoxical conflicts arise [Cra87]. Horgan does admit these conditions are not explicitly stated in the original problem. "Very well," he continues, "I hereby *stipulate* that the conditions are included".

While Gardner among others denies the existence of the predictor because of logical reasons, some others challenge his existence in order to hold on to the notion of free will. An exultant Isaac Asimov proclaims:

> I would, without hesitation, take both boxes... I am myself a determinist, but it is perfectly clear to me that any human being worthy of being considered a human being (including most certainly myself) would prefer free will, if such a thing could exist... Now, then, suppose you take both boxes and it turns out (as it almost certainly will) that God has foreseen this and placed nothing in the second box. You will then, at least, have expressed your willingness to gamble on his nonomniscience and

> on your own free will and will have willingly given up a million dollars for the sake of that willingness — itself a snap of the finger in the face of the Almighty and a vote, however futile, for free will... And, of course, if God has muffed and left a million dollars in the box, then not only will you have gained that million, but far more important you will have demonstrated God's nonomniscience.

3 Assumptions

In our view, most of the problems that have arisen in the debate on Newcomb's paradox are the product of a problematic interaction between mainly four principles: *determinism*, *predictability*, *backward causality* and *free will*. We have already introduced backward causality, the other three we will now introduce shortly.

Determinism is the theory that the state of the world at a certain moment of time is determined by the state at the moment just before it. According to the laws of nature, every state is caused by the previous state. Formally, there is a description of the state of the world p_n at any given time t_n and a finite set of laws $\{L_1, L_2, \ldots, L_k\}$. The description of the world p_{n+1} at time t_{n+1} derives from applying laws $\{L_1, L_2, \ldots, L_k\}$ to p_n.

Predictability is the notion that a certain fact about the state of the world at a certain time can be known in advance. Whether this fact is known by a human, an alien or a machine has, as we will see, consequences for its relationship with determinism and free will.

Free will is the notion that humans have an influence on the course of reality by the choices they make. These choices are not caused by deterministic laws and therefore called 'free choices'. What the cause for a free choice might be instead is hard to define. Sometimes a free choice is even defined as an uncaused choice.

According to these definitions determinism and free will are mutually exclusive and in fact that is what they are usually considered to be. The relation between determinism and predictability is more nuanced. The example with the supercomputer shows that not all events can be predicted by a predicting entity that is part of a deterministic system. When asked to predict the event of the green light turning on, the computer runs into a contradiction. However, we also showed that there are also events which may be safely predicted because the prediction does not stop the event from taking place. Furthermore, we stipulate that an entity outside a determin-

istic system can make predictions of the system if it knows the full state of the system and the laws that cause the next state. We can think of an alien which knows everything about the state of the earth and its laws, and is therefore able to predict our future.

The relation between free will and predictability is not so obvious. Predictability as a product of determinism is easier to conceive than predictability in a non-deterministic environment. From our human viewpoint, because of undetermined, free choices, it is impossible to know anything about the future. Keeping all the options open however, we will consider a prediction in a non-deterministic environment to be possible if information travels back in time. An example would be a psychic looking in his crystal ball, supernaturally receiving a piece of information from the future. Furthermore, we imagine God to be able to do predictions about a non-deterministic system as well. We consider Him to exist outside of time. From this extra-temporal vantage point, he considers all of existence – past present and future – as we observe the present. Because of this capability he can, from our viewpoint, make predictions about our future, including our free choices.

Determinism is incompatible with backward causation, since in determinism causation is forward by definition. Backward causation is one of the bases on which the possibility of prediction can be explained.

Many scientists, when telling Newcomb's paradox, mention the fact that the superior Being might just as well be imagined to be a highly superior intelligence from another planet, or God. The paradox remains just as strong, as Gardner claims. Nozick tells us that we may imagine a graduate student from another planet, checking a theory of terrestrial psychology, who first takes measurements of the state of our brains before making his predictions. Poundstone puts a psychic in the place of the Being, with an accuracy of 90%. He furthermore states that you are not interested in whether the psychic's powers are proven or not – your only motive is to leave with as much money as possible.

In our view, this indifference towards who we are actually facing in Newcomb's paradox has contributed to the nearly inextricable tangle of discussions it has aroused. By specifying more precisely who where dealing with, one inherently clarifies at least a number of assumptions that are crucial to the solution. Deciding on the assumptions first enables us to decide better which arguments to involve or exclude, keeping the discussion within the right boundaries. We will put this into practice by considering seven different interpretations of 'the Being', their corresponding assumptions and the solutions.

1. God. In this case predictions are assumed to be perfect because to God all information is known at any time. Backward causality is in operation. Free choices exist.

If we're up against God, there's no doubt about the outcome of the game. If we choose the second box, we're sure to find the $1 million. If we choose both boxes we're sure to find the second one empty. To us it appears like we are filling the boxes. But not to our friend, the spectator. He sees the money that's on the table. But this time, he's not urging us to take both boxes... What he sees is the result of the choice we're still going to make. If he sees the money in the second box, he can be happy because he knows we are going to take only the second box. If he sees the second box empty, he'll regret that we're going to choose both.

It remains astounding however that after the second box is filled or not, which is a result of our choice, we still have a *free* choice to make. Doesn't it appear like the choice is determined instead? We can make the case even stronger. Suppose God tells us we will take off our shoes by our own free will. Would we still make a free choice?

We can maintain the choice is still free. At the moment we make a free choice, we have more than one option and it is only we who make the choice, not somebody or something else. In our mind we can construct contradicting scenarios, like rebelliously keeping our shoes on if God told us we would take them off. These contradictory scenarios cannot take place — we arrived at the logical contradiction argument of Gardner. To this we responded by showing that although not all predictions can be made, some can. This implies that, if we are in a rebellious mood, God can not give us the prediction about whether we will take off our shoes. If we are in a compliant mood, he can. In other words, our free will cancels out possible scenarios or paths in time. The path in time that we do follow, is made possible by our free choices[3].

This brings up a question however: how can a free choice that is made at a certain moment cancel out paths in time that started before that moment? The only plausible answer to this is that free choice is not bound to time. Free choices are omni-temporal. This is the consequence we have to face if we assume divine and interacting prediction and free will.

2. A genuine psychic. In this case predictions are assumed to be perfect because information travels backwards through time. Backward causality is

[3]But what then if our free choices leave several time paths open? I leave that question to God — it was his idea to start predicting.

in operation. Free choices exist.

Let's play the prime number game again. We saw before that if God plays the prime number game with us, he will have thought of a number that, if shown to us, leads us to make the choice corresponding to it. But how about the psychic? It is hard to imagine that he could 'test' a number. However, if he looks in his crystal ball, he will see a fragment of the future. He sees which choice we are going to make. He can then go on and write down the number – prime if he saw us choose the second box, composite otherwise. In fact, as soon as he sees us make a choice, he can't act otherwise than to write down a correct number. He might try to write down a wrong number but then he will find himself to make a mistake and write down a correct number. He might try to not write down a number at all but he can't escape it. He saw in his crystal ball a piece of the future that could not have 'happened' if he was not going to write down a correct number.

We are back at Gardner's contradiction argument again. This time it is the psychic who sees his own future. Is he able to act against it? Again we use our argument of possible predictions. Suppose the psychic's crystal ball would remain black if the prediction about the future can not be made because of logical impossibility. If the event can safely be predicted, the ball will show it. In that case, we would know that everything will go as explained if the psychic tells us he's made a prediction. Apparently, the prediction was possible. Apparently, the psychic was not in a rebellious mood.

Of course we are still in the story as well. We could have ran away. But we didn't — we were not in a rebellious mood either.

3. A genuine psychic with some noise on the line. In this case predictions are not perfect because the psychic sometimes gets the information wrong, let's say in 10% of the cases. Backward causality is operating in the sense that if you make choice x, the probability that the psychic will predict x is 90%. The probability distribution is uniform. Free choices exist.

This case is exactly similar to the genuine psychic with perfect predictions, except for the fact that free choices cancel out paths in time with a certain probability. Therefore, if we choose the second box, there's a 90% chance we will find the $1 million there and still a 10% chance our choice has been incorrectly transmitted to find an empty box. If we play this game many times, the average payoff will become that of the expected utility argument; $900,000 for the one-box strategy, $101,000 for the two-box strategy. Maximizing our expected utility, we choose only the second box.

4. A superior alien. In this case our world is deterministic. The superior alien has the ability to perfectly observe everything observable in our world. Since he knows all of the laws that are in effect, he can make perfect predictions about our future. Because of determinism, free will and backward causation are not in operation.

First we have to realize that our world is actually not perfectly deterministic in this case. In addition to the set of laws $\{L_1, L_2, \ldots, L_k\}$, the alien, who is not part of our world, has an effect on the state of our world as well. We will however consider the case where the interaction of the alien with our world does not have any effect on the predicted event. For this we need the following assumptions. (i) The prediction is done solely in the alien's mind. (ii) The only interaction between the alien and the world is the placement of the money in the box. (iii) The placement of the money doesn't have any effect on our world until the box is opened. We can imagine (ii) being true if instead of the alien, human beings have set up the experiment and told us the rules. Also, his observations have no effect on our world. We can imagine (iii) being true if the alien has a supernatural way of filling the second box without it being noticed by anybody. Furthermore, we allow ourselves the luxury of neglecting the effect the placement has on the air molecules in the box. The effects of the alien's interaction only take place when we open the box and the contents are revealed.

In this situation, all the normal deterministic laws apply until the moment of opening the box. It is these laws that enable the alien to predict which box we will open. When we ponder about which choice we should make, we are allowed to realize the comforting fact that our process of pondering and the resulting choice are simply caused by the deterministic laws. Therefore, whichever choice we make, it was determined and therefore predictable by the alien. If we choose our pondering to result in taking the second box only, that will be what was predicted. We are not worried about the self-referentiality of our argument — there is no escape to the mighty laws of determinism. Let's take the second box.

In the case of the prime number game, things are different. (ii) could be reformulated to (ii) The only interaction between the alien and the world is the writing down of the number. However, (iii) can not be assumed in this example. Clearly, the interaction of the alien takes place immediately after the prediction has been made, before the choice is going to be made. Therefore, the alien is no longer predicting a deterministic system from an external viewpoint but rather the larger system of our world augmented with a number of laws and predicates accounting for his influence. In other words, the alien has become part of the system he is predicting. This is a

case we separately discuss in the next paragraph. The prime number game falls outside this category.

5. A supercomputer. In this case our reality is again deterministic. The supercomputer, like the alien, has the ability to perfectly observe everything observable in our world and know the laws. It can therefore make perfect predictions. Unlike the alien however the computer itself is part of the deterministic system. Because of determinism, free will and backward causation are not in operation.

We've already seen that a supercomputer predicting the truth of any future event cannot exist. We will therefore assume the computer to give a prediction only if it is a possible prediction. As Kiekeben points out, it is problematic to assume the existence of a computer with full knowledge of the world it is part of [Kie99]. Suppose the computer knows fact q. To have full knowledge of the world, the computer should then also know that he knows fact q. This again is a fact he should know. This leads to an infinite number of facts the computer should know. Kiekeben claims that the supercomputer therefore cannot exist. We will however assume that there exists a supercomputer that is able to know an infinite number of facts and therefore has full knowledge of the world. Also, this computer is able to derive from this infinite number of facts the next state the world will be in. It is thereby able to perfectly predict the future, given that it will only make possible predictions.

We will assume that the supercomputer has made the prediction, and that a human has prepared the boxes and told us the rules of the game. With these assumptions, the question of which box to take becomes straight forward. We know that if the computer gives a prediction, it will be based on the determined future and it will be a possible one, i.e. one that doesn't lead to a logical contradiction. So we know that whatever was put in the boxes was determined and whatever we choose is determined, and that these two events don't lead to a logical contradiction. We can conclude that if we take the second box, the $1 million will be in there. Therefore we will take the second box.

The same reasoning applies to the prime number example, which is under these assumptions equivalent. We don't need to wonder whether we can choose contrary to the number being prime or composite. The fact that the computer exists and has made a prediction tells us that we are in a noncontradictory scenario. We will take the second box.

6. A skillful psychologist. Whether our world is deterministic or free will exists is undecided. We only know that the person facing us has no supernatural powers, but has somehow managed to make the correct prediction in 90% of the many previous times he has performed.

Since we cannot base our reasoning on determinism or backwards causation, we are left with our common sense, psychological skills and intuition. We might wonder why the psychologist has done so many correct predictions. Most likely, he has a very good sense of which choice people will make. Based on his psychological judgment of the one opposite him, he predicts whether it is a one-boxer or a two-boxer. Because of his skills, his prediction is quite accurate. He can feel when somebody is going to use the expected utility argument. He can feel when somebody is going to use the dominant strategy argument. He might even know when somebody favours both boxes but doesn't want to be predicted as a two-boxer and takes the second box only.

However, all this wondering is futile. We can reason about our strategy, about his prediction, about our predictionability, about his reasoning about our reasoning, about whether he knows we've written this essay, but all this will not change the prediction that has already been made. We've been classified already, no smart mind-tricks can change that. The money has been put on the table. We'll take it all. We'll take both boxes.

7. A suspiciously accurate human predictor. This case is similar to the psychologist, except for the fact that this man has a remarkably high prediction accuracy: 100%. He's done the test many times before, say a thousand times, and he was right every time.

In our view, this is the scenario closest to the one in which Gardner and Nozick eventually decide to take both boxes, mournful of the almost certain absence of the $1 million. Indeed this absence is almost certain, since the score of the predictor is so high that it can not be explained by any plausible reason we know of. Rather there must be an unknown reason for the predictors perfect score, unknown but real. Gardner suspects treason. But there is no treason, the predictor is perfect. Levi hesitates. Bar-Hillel and Wolf go for the money. "Just look at the correlation!" Bar-Hillel exclaims. "Put the money in there!" Wolf urges. And look, two more happy one-boxers walk away, counting their $1 million... Gardner suffers... He can't get himself to act against his reason. Levi grins. He's found a loophole. They open the boxes... faces as stone. Two more two-boxers perplexed... and not rich. Asimov comes in. Without hesitation, he tears open his two boxes. Nothing... of course. But no regret. "At least

I care about my freedom!" he shouts at us.

In comes Jonker. "An unknown reason..." he ponders. "Nothing wrong with that. Since when do we know everything?" He opens the second box. A million dollars is smiling at him. "And that," he adds, "I did by my own free will!"

4 Conclusion

We have shown that four assumptions are fundamental when giving a solution to Newcomb's paradox: determinism, predictability, backward causality and free will. We have looked at seven scenarios involving different sets of assumptions. We showed how in each scenario we were able to decide on which strategy to use. When the scenario involves backward causality, determinism or perfect prediction one should take the second box only. Only against the skillful psychologist should one pick both boxes. Against the suspiciously accurate predictor we know there must be an unknown reason for his accuracy. For this reason, we take the second box only.

References

[BHM72] Maya Bar-Hillel and Avishai Margalit. Newcomb's paradox revisited. *The British Journal For the Philosophy of Science*, 23(4):295–304, November 1972.

[Cra87] William L. Craig. Divine foreknowledge and Newcomb's paradox. *Philosophia*, 17:331–350, 1987.

[Gar86] Martin Gardner. *Knotted Doughnuts and Other Mathematical Entertainments*, chapter 13. Freeman, New York, 1986.

[Hor85] Terence Horgan. Newcomb's problem: A stalemate. In *Paradoxes of Rationality and Cooperation*, pages 223–234. The University of British Columbia Press, Vancouver, 1985.

[Kie96] Franz Kiekeben. Newcomb's paradox. `http://members.aol.com/kiekeben/newcomb.html`, 1996.

[Kie99] Franz Kiekeben. Does determinism imply absolute predictability? *De Philosophia*, XV(1), Spring/Summer 1999.

[Lev82] Isaac Levi. A note on Newcombmania. *The Journal of Philosophy*, 74(6):337–342, 1982.

[Noz69] Robert Nozick. Newcomb's problem and two principles of choice. *Essays in Honor of Carl G. Hempel*, 1969.

[Noz86] Robert Nozick. Reflections on Newcomb's paradox. In Martin Gardner, editor, *Knotted Doughnuts and Other Mathematical Entertainments*, chapter 14. Freeman, New York, 1986. Addendum by Martin Gardner.

[Pou88] William Poundstone. *Labyrinths of Reason*. Anchor Press / Doubleday, New York, 1988.

[Sch74] George Schlesinger. The unpredictability of free choices. *The British Journal for the Philosophy of Science*, 25:209–21, 1974.

[Wol81] Fred A. Wolf. *Taking the Quantum Leap*. Harper and Row, 1981.