The two-envelope paradox

PHIL 20229
Jeff Speaks

March 26, 2008

1 A few versions of the paradox

1.1 The randomized open version

Suppose that you have a certain amount of money, say $20. I now put double that amount into one envelope, and half that amount into another envelope, and put the envelopes into a machine which randomly selects one. Suppose that I now give you the chance to trade the $20 for the envelope which comes out of the randomizer. Should you? How would you calculate the expected utility of switching?

1.2 The randomized closed version

Suppose that you have a certain amount of money in a closed envelope. You don't know how much; but you do know that there's some finite, nonzero amount of money in the envelope. I now put double that amount, whatever it is, into one envelope, and half that amount into another envelope, and put the envelopes into a machine which randomly selects one. Suppose that I now give you the chance to trade your envelope for the envelope which comes out of the randomizer. Should you? How would you calculate the expected utility of switching?

1.3 The probabilistic open version

Suppose that you have a certain amount of money, say $20. I have an envelope which has 1/2 chance of containing $40 and 1/2 chance of containing $10. Should you trade your $20 for my envelope?

1.4 The probabilistic closed version

Suppose that you have a finite, nonzero amount of money in a closed envelope, but you don't know how much. I have an envelope which has 1/2 chance of containing double that amount (whatever it is) and 1/2 chance of containing half that amount. Should you trade your envelope for mine?

So far, it might seem that these cases are pretty straightforward. But the obvious answers to the questions above lead us to say some strange things in (at least superficially) similar cases.

1.5 The choice open version

Suppose now that I have two envelopes, A and B, one of which contains twice the amount of money in the other. You choose one — suppose that it is A. You open it, and find $20 inside. Should you trade your $20 for envelope B? How should you reason about the expected utility of switching in this case?

This appears to be, in relevant respects, just the same as the probabilistic open version; so it appears that you should not only switching, but be willing to pay to switch.

But, on the other hand, the decision to switch in this case looks sort of odd. After all, you just chose A randomly; why should opening it give you a reason to think that you stand to gain by switching for the other envelope?

1.6 The choice open reverse version

Suppose now that I have two envelopes, A and B, one of which contains twice the amount of money in the other. You choose one — suppose that it is A. I now open envelope B,

which you did not choose, and find $20 inside. Should you trade envelope A for my $20?

Here the reasoning seems the reverse of the above; so it seems that you should want to, and indeed be willing to pay to, keep your own envelope.

But again, this seems odd. Why should my opening my envelope put you in a position where it would be rational for you to pay to keep your envelope?

Still, so far you might reasonably regard this sort of reasoning as surprising, but not quite paradoxical. After all, it might not seem as though we've really been able to deduce an absurd result yet.

1.7 The choice closed version

So suppose that again I have two envelopes, labeled A and B, and you know that one contains twice the amount in the other. Again, you choose envelope A. We open neither envelope. Should you exchange your envelope for mine?

On the one hand, you want to say 'Yes.' After all, we said 'yes' in the choice open version, and just as the randomized closed version seems just the same as the randomized open version the probabilistic closed versions seems just the same as the probabilistic open version, the choice closed version seems just the same as the choice open version.

However, there are two reasons for thinking that this cannot be correct:

1. Just as we can extend the reasoning from the choice open version to the choice closed version, so we can extend the reasoning from the choice open reverse version to the choice closed version. But this would tell us *not* to switch.
2. Suppose that we switch. Now I might ask you whether you want to switch again. Should you? Clearly not; you would just be trading back for the envelope you had in the first place, so any line of reasoning which led to the conclusion that you stood to gain from *both* switches must be flawed. But it seems that any line of reasoning which leads you to believe that you should switch the first time can be used to show that you should also switch a second time (and a third, and a fourth ...).

2 Problems with the set-up of the paradox

2.1 There's no such thing as 1/2 of $.01

The extension from the open versions of the case to the corresponding closed versions rests on the assumption that, no matter what value you found in your envelope, you would assign equal probabilities to the hypotheses that the other envelope contained double and that it contained half. But suppose you open your envelope, and there is only one cent inside. Then it seems that the other envelope cannot contain half that amount, since there is no such thing as one half of one cent.

Two responses to this problem: (i) this would make the argument for switching better, not worse; (ii) we can make sense of the idea of half of a cent.

However, nothing is lost in what follows if we just assume that the lowest amount of money that can be in an envelope is \$1, and that every possible value in the envelopes is some power of 2, so that possible values are \$1, \$2, \$4, ...

2.2 There's a finite amount of money in the world

Suppose that you know that there's a total of 100 billion dollars in the world, and you open your envelope to find 400 billion dollars inside. It doesn't seem that there could be double that amount in the other envelope, since their sum would then exceed the total amount of money in the world. So it seems that in this case you can be certain that you should not switch. As above, this seems to disrupt the argument which takes us from the 'open' versions of the example to the 'closed' versions.

How should we get around this problem?

2.3 Infinite probability distributions

For the reasons given above, it is clear that there cannot be a finite upper bound on the amount of money in the envelopes. So, there are infinitely many possible values of the envelopes. Furthermore, it seems that it cannot be the case that some of these values are less likely than others to be in one of the envelopes; otherwise, there are some values you might find in your envelope which are such that you would not be rational to believe that the other envelope had a 1/2 chance of containing half, and a 1/2 chance of containing double. (For example, imagine that it is much more likely that an envelope will contain \$2 than that it will contain \$8, and you find \$4 in the envelope you select. It would not be rational for you to switch. So there must be no cases of this sort if the original paradox is going to be convincing.)

So there must be infinitely many equiprobable possible values of the envelopes. But this does not seem possible. What would the probability of each possible value be? (There's an analogy here with Zeno's arguments against the possibility of motion in a world in which space and time are continuous; even if we can perform infinitely many tasks in a finite time, there's still a problem with performing infinitely many tasks each of which takes some finite amount of time t in a finite time. Even if an infinite series can sum to 1, an infinite series of equal finite numbers cannot.)

This raises a genuine problem for the paradox. However, it is a problem that can be solved. For example, suppose that the probabilities of values of the lower envelope are as follows:

$$\Pr(\text{lower}{=}1{=}2^0) = \tfrac{1}{4} * \tfrac{3}{4}^0 = \tfrac{1}{4}$$
$$\Pr(\text{lower}{=}2{=}2^1) = \tfrac{1}{4} * \tfrac{3}{4}^1 = \tfrac{3}{16}$$
$$\Pr(\text{lower}{=}4{=}2^2) = \tfrac{1}{4} * \tfrac{3}{4}^2 = \tfrac{9}{64}$$

$\Pr(\text{lower}=8=2^3) = \frac{1}{4} * \frac{3}{4}^3 = \frac{27}{256}$

...

This probability distribution raises none of the problems raised by the simpler one mentioned above — since the series of probabilities of the possible lower values is an infinite series of ever decreasing finite values which sums to 1 — and yet it still supports switching. For any amount you find in your envelope, there is a greater chance that the other envelope contains less than that it contains more, but the difference in probabilities is more than offset by the difference between what you would gain if the other envelope were higher and what you would lose if the other envelope were lower.

3 Solutions to the paradox

3.1 'The Two-Envelope Paradox and Infinite Expectations'

Antzenius & McCarthy (in 'The Two-Envelope Paradox and Infinite Expectations') offer a solution to the paradox which raises the following doubt about the argument for switching. Supposing as above that the possible values in the envelopes are all powers of 2, then the amounts that you can stand to gain and lose are all also powers of 2. So consider, for example, \$4. What are the chances that switching will gain you \$4, versus the chances that switching will lose you \$4? The first happens if there is \$4 in envelope A, and \$8 in envelope B; the second happens if there is \$8 in envelope A, and \$4 in envelope B. But the chances of these being the values in the envelopes is the same; so the chances of switching gaining you \$4 is the same as the chances of switching losing you \$4.

Once you see that, you can see that this argument generalizes. The chances of you gaining \$1 is the same as the chances of you losing \$1, the chances of you gaining \$64 is the same as the chances of you losing \$64, and so on for every power of 2.

This makes it clear that you do not stand to gain by switching in the closed case. But we already knew that. It seems that you might still wonder (i) whether you stand to gain by switching in the open case, and (ii) what was wrong with the initial line of reasoning which seemed to support switching in the closed case.

3.2 Dominance reasoning and inference from an unknown

It will be useful to take a closer look at that reasoning. Recall that in our discussion of Newcomb's problem and the prisoner's dilemma our use of the following 'dominance principle':

> Suppose that you are choosing between two actions, act 1 and act 2. It is always rational to do act 2 if the following is the case: whatever happens, doing act 2 will never make you worse off than doing act 1; and in some cases, doing act 2 will make you better off than doing act 1.

This principle does not straightforwardly apply to the two-envelope paradox, but a principle closely related to it does:

> Suppose that you are choosing between two actions, act 1 and act 2. It is always rational to do act 2 if the following is the case: there is some piece of information about the case which you lack but which is such that, were you acquire that piece of information (no matter what it turns out to be), it would be rational for you to do act 2.

Call this principle *inference from an unknown*. This is the principle which seems to lead us from the open versions of the case to the closed versions; so, one possible view of the paradox is that in the open versions, you are rational to believe that you stand to gain by switching, and in the open choice reverse version you are rational to believe that you stand to gain by not switching, but that in the closed choice version you have no argument that you stand to gain either way. If this is correct, then the case is one in which inference from an unknown leads us astray.

But this just leads to another question: how *could* inference from an unknown fail? One reply is that it can fail when applications of the principle in a single case lead to contradictory results for different choices of the 'unknown.' This is illustrated by the move from the open choice case to the switching in the closed choice case, and the move from the open reverse choice case to the conclusion that you ought to believe that you stand to gain by not switching in the closed choice case. These arguments both rely on inference from an unknown, but the relevant unknown in the first case is the value of envelope A, and in the second case it is the value of envelope B.

Is it really true that we should believe that we stand to gain by switching in the open case? Suppose that each players in the game look inside their envelopes, but don't tell the other person what they've seen. On the above sort of solution, each would be willing to pay the other to switch. Does this indicate that something has gone wrong?

4 Problems with infinite expectations

There's another sort of problem raised by the two-envelope paradox, which concerns how we should reason about situations in which the expected gain of some action is infinite. A famous example of such a case is the St. Petersburg paradox. Clark describes the case as follows:

> This needs some explanation. Suppose you are the player: if heads comes up on the first throw you get £2, if it comes up on the second throw you get £4, on the third £8, and so on. For each successive throw the payout doubles. The chance of heads on the first throw is ½; the chance that heads comes up first on the second throw is the chance of tails on the first × ½, that is, ¼; the chance that heads comes up first on the third throw is ½ × ½ × ½, and so on. For each successive throw the chance halves, just as the payout doubles. The expected gain from the first toss is £2 × the probability that it is heads (½), that is, £1; from the second throw it is £4 × ¼ = £1; from the third £8 × ⅛ = £1, and in general for the *n*th throw it is £2^n × $½^n$ = £1. Since there is no limit to the possible number of throws before heads comes up, the sum for the expected gains, 1 + 1 + 1 + . . . goes on for ever and the expectation is infinite. Yet would you pay *any* sum, however large, to participate?

Is this sort of case fully explained by diminishing returns and risk aversion? In other words, if we 'valued each dollar equally' and had did not take a loss of $10 to be worse than a win of $10 is good, is it really true that we would be rational to pay any amount of money to play this game?

McCarthy and Arntzenius offer an interesting version of the case in their 'paradox of heaven and hell':

> "One day you wake up in Purgatory, and you are about to discover what you have long suspected, that God does not much care for rational people like yourself. First, God reliably informs you that you are immortal, but this does not make you revise your temporal neutrality: you care just as much about how well off you will be on some day in the distant future as you do about tomorrow. Then God gives you a guided tour of Heaven and Hell and asks you what you think. You decide that a day in Heaven is as good as a day in Hell is bad, and you would be indifferent between a day in Heaven followed by a day in Hell versus two days in Purgatory. Furthermore, you decide that how many or few days you have spent in the past or expect to spend in the future in either Heaven, Hell, or Purgatory, does not affect how much you would enjoy or hate any day in the present in any of those places. So we can represent your preferences as follows. The values of one day in Heaven, Purgatory, and Hell are, respectively 1, 0, and -1. And the value of any gamble over days in these places is equal to the expected number of days in Heaven minus the expected number of days in Hell, at least when these are finite. . . .
>
> God then offers you a St. Petersburg gamble where the payoffs are days in Heaven: a probability of a half of the next day in Heaven, and then back to Purgatory; a probability of a quarter of the next two days in Heaven, then

> back to Purgatory; and so on. Great! You accept, and as was inevitable, you win some nite number of days in Heaven, then back to Purgatory. But early the next morning at the entrance to Heaven you meet God, and he makes you a deal: if you abandon all the days in Heaven you have won, and spend today in Hell, hell give you another shot at the St. Petersburg gamble. But if you decline, after your nite stay in Heaven youll spend the rest of your days in Purgatory. From your preferences it seems rational for you to accept the deal, so you do, and as was inevitable, you win another nite number of days in Heaven, but you have to spend today in Hell. Early the next morning, rather beleaguered after your day in Hell, you meet God again at the entrance to Heaven. He makes you exactly the same deal as he did the day before, and given your preferences, it seems rational for you to accept, so you do. Off you go back to Hell, and rational person that you are, you are beginning to suspect that you have an unending life in Hell to look forward to. What's gone wrong?"

This paradox relies on the fact that although the expected utility of playing a St. Petersburg game is infinite, the actual payoff is always finite. So in every St. Petersburg game, the actual payoff is less than the expected utility of playing; that's why it seems that the rational person will always give up their winnings for a day in hell and the chance to play again; the expected utility of playing again will always be greater than the days won + the day in hell (which, by hypothesis, is as bad as one day in heaven is good).

What's wrong with the line of reasoning which leads to spending eternity in hell?