The St. Petersburg paradox & the 2 envelope paradox

Consider the following bet:

The St. Petersburg

I am going to flip a fair coin until it comes up heads. If the first time it comes up heads is on the 1st toss, I will give you $2. If the first time it comes up heads is on the second toss, I will give you $4. If the first time it comes up heads is on the 3rd toss, I will give you $8. And in general, if the first time the coin comes up heads is on the *n*th toss, I will give you $2^n.

Would you pay $2 to take this bet? How about $4?

Suppose now I raise the price to $10,000. Should you be willing to pay that amount to play the game?

What is the expected utility of playing the game?

We can think about this using the following table:

Outcome	First heads is on toss #1	First heads is on toss #2	First heads is on toss #3	First heads is on toss #4	First heads is on toss #5	.....
Probability	$2	$4	$8	$16	$32	.....
Payoff	1/2	1/4	1/8	1/16	1/32	.....

The expected utility of playing = the sum of probability * payoff for each of the infinitely many possible outcomes. So, the expected utility of playing equals the sum of the infinite series

1+1+1+1+1+ 1+1+1+1+1+ 1+1+1+1+1+ 1+1+1+1+1+.....

But it follows from this result, plus the rule of expected utility, that **you would be rational to pay any finite amount of money to have the chance to play this game once**. But this seems clearly mistaken. What is going on here?

The St. Petersburg

I am going to flip a fair coin until it comes up heads. If the first time it comes up heads is on the 1st toss, I will give you $2. If the first time it comes up heads is on the second toss, I will give you $4. If the first time it comes up heads is on the 3rd toss, I will give you $8. And in general, if the first time the coin comes up heads is on the *n*th toss, I will give you $2^n.

The expected utility of playing = the sum of probability * payoff for each of the infinitely many possible outcomes. So, the expected utility of playing equals the sum of the infinite series

1+1+1+1+1+ 1+1+1+1+1+ 1+1+1+1+1+ 1+1+1+1+1+.....

But it follows from this result, plus the rule of expected utility, that **you would be rational to pay any finite amount of money to have the chance to play this game once**. But this seems clearly mistaken. What is going on here?

An initial thought is that this scenario depends, unrealistically, on its being possible for the person offering the bet to have no finite bound on the amount of money that they are able to disburse at the end of the bet. Suppose, after all, that someone offered the bet who had a net worth of $1 billion. Then, if the coin came up heads for the first time on the 30th flip (or later), he would be unable to pay up.

But this does not capture all that is puzzling about this paradox. For suppose that you were playing this game with a billionaire - the expected utility of playing would still be $30. And would you really pay $29 to play? If you think that it would not be rational to play, then this is still a counterexample to the rule of expected utility.

This paradox was discovered by the Swiss mathematician Nicolaus Bernoulli, who died in 1726 at the age of 31. He discussed the paradox with his younger brother Daniel, who based his theory of rational action in part on the paradox.

Bernoulli took the paradox to be a dramatic illustration of the phenomenon of **diminishing returns**: the fact that, in this instance, a gain of a certain amount of money is less valuable to the recipient if it comes after another large gain of money. How might this help to explain the (apparent) fact that it is irrational to pay what the rule of expected utility requires in this case?

The St. Petersburg

I am going to flip a fair coin until it comes up heads. If the first time it comes up heads is on the 1st toss, I will give you $2. If the first time it comes up heads is on the second toss, I will give you $4. If the first time it comes up heads is on the 3rd toss, I will give you $8. And in general, if the first time the coin comes up heads is on the *n*th toss, I will give you $2^n.

This paradox was discovered by the Swiss mathematician Nicolaus Bernoulli, who died in 1726 at the age of 31. He discussed the paradox with his younger brother Daniel, who based his theory of rational action in part on the paradox.

Bernoulli took the paradox to be a dramatic illustration of the phenomenon of **diminishing utilty**: the fact that, in this instance, a gain of a certain amount of money is less valuable to the recipient if it comes after another large gain of money. How might this help to explain the (apparent) fact that it is irrational to pay what the rule of expected utility requires in this case?

But diminishing utility does not seem to explain everything that is puzzling about this case. To say that this case is explained by diminishing utility is to say that this case is to be explained by the fact that dollar amounts and utility can come apart. But we can also set up the St. Petersburg in such a way as to get around this problem, by increasing payoffs by more than doubling them when sufficiently high amounts are in question.

A different response to the paradox is to try to explain it in terms of **risk aversion**: the tendency to value not losing something above gaining that same thing. But in what sense can risk aversion explain this paradox? To be sure, the fact that we **are** risk averse might well explain our unwillingness to take the bet; but the question is whether some sort of constraint about risk aversion can really be a rule on rational action. Why would risk aversion be rational (supposing, as above, that we are taking into account diminishing utility)?

One can bring this issue to the fore by focusing (as Clark suggests) on simpler cases than the St. Petersburg. Suppose that I offer you a chance to bet $10,000 on a one in a million chance to win 20 billion dollars. The expected utility of playing is $20,000-$10,000= $10,000 - so, if the rule of expected utility is correct, one should certain take the bet. But would you? Many people have a strong inclination to reject bets of this sort; but it seems that either the rule of expected utility is false (or limited in scope in some way) or our intuitive judgements about such cases are incorrect.

Issues which are in some ways related to these are raised by the two-envelope paradox.

Issues which are in some ways related to these are raised by the two-envelope paradox.

One forceful way to present this paradox is as a series of scenarios, the earliest of which are unproblematic, but which can be turned, by apparently innocuous variation, into genuinely paradoxical conclusions.

Let's begin with a simple one:

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?

It seems obvious that you should; and, if this seems clear, a similar verdict seems called for in the following case:

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?

But now consider the following variant on this example:

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?

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?

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?

It seems obvious that you should; and, if this seems clear, a similar verdict seems called for in the following case:

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?

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?

One way to bring out the weirdness here is by considering a scenario in which the other envelope is opened:

The choice open reverse version

Suppose again 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 $4 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? Why should which envelope is opened matter at all?

We can generate even stranger results by developing our initial case - the randomized open version - in another direction.

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?

We can generate even stranger results by developing our initial case - the randomized open version - in another direction.

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?

It does not seem that the fact that you don’t open the envelope should matter. You know that, whatever amount you find in your envelope, it will be rational for you to switch - so it seems clear that you should switch now, in advance of knowing that amount. And if this is clear, then the same can be said about the following case:

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?

But this is enough to set the stage for the most puzzling of the cases we will discuss.

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?

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?

We can generate even stranger results by developing our initial case - the randomized open version - in another direction.

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?

But this is enough to set the stage for the most puzzling of the cases we will discuss.

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?

Just as the probabilistic and choice open versions seem relevantly similar, so the probabilistic and choice closed versions seem relevantly similar. But that means - given that we should switch in the probabilistic closed version - we should also switch in the choice closed version. However, there are two reasons for thinking that this cannot possibly be right.

First, 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.

Second, 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…).

These results are odd enough that at this point it may be reasonable to think that there is something incoherent in the set-up of the examples. So let’s examine the assumptions needed to generate these cases a bit more explicitly.

These results are odd enough that at this point it may be reasonable to think that there is something incoherent in the set-up of the examples. So let's examine the assumptions needed to generate these cases a bit more explicitly.

First, the set-up assumes that, no matter what amount of money you find in your envelope, you should think that the probability that the other contains twice that amount is ½, and that the probability that it contains double that amount is ½. But suppose you find 1¢ in your envelope - then the other envelope could not contain ½ that amount, right?

This sort of objection is not very serious, for two reasons. First, we can make sense of the idea of an envelope containing ½ of 1¢ - we could suppose that that envelope gave you a chance at a fair coin flip for a penny, for example. Second, the paradox is not weakened if we assume that there is a lower bound on the amount of money in the envelope. Suppose that the envelopes can only contain powers of $2, so that the permissible values are $1, $2, $4, $8, …. Then if one found $1 in one's envelope, this would strengthen rather than weaken the case for switching, which is the result we are trying to avoid.

Second, and more significantly, the argument assumes that there must be **no upper bound on the amount of money that can be in the envelopes** - and hence no upper bound on the amount of money available in the world. For suppose that there were $1 billion in the world, and you found $400 million in your envelope. Then you could be sure that the other envelope did not contain double that amount for, if it did, the sum of the envelopes would exceed the amount of money in the world.

It seems to me quite plausible that there must be something wrong with the arguments we just discussed other than the fact that it assumes that there is no finite bound on the amount of money in the world; it seems as though they must involve some simple logical flaw. Moreover, it seems that we can imagine a variant of the case in which there is an infinitely powerful being who is able to bestow any finite reward on you he wishes (this needn't, of course, be money). Then imagine that such a being places the relevant rewards in the envelopes. Wouldn't, for example, constantly trading envelopes in the choice closed version of the paradox still be absurd?

Finally, the arguments assume something quite striking about your views about the probabilities of various amounts being in the two envelopes. 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 ½ chance of containing half, and a ½ 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; hence, in the closed versions, it would not be true to say that **whatever amount you found in the envelope, you would be rational to switch**. So there must be no cases of this sort if the original paradox is going to be convincing.

So it seems that the paradox, as presented, requires you to think that **each of the infinitely many possible assignments of values to the envelopes is equally probable**. But this does not seem to make sense - what would the probability assigned to each be?

Finally, the arguments assume something quite striking about your views about the probabilities of various amounts being in the two envelopes. 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 ½ chance of containing half, and a ½ 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; hence, in the closed versions, it would not be true to say that **whatever amount you found in the envelope, you would be rational to switch**. So there must be no cases of this sort if the original paradox is going to be convincing.

So it seems that the paradox, as presented, requires you to think that **each of the infinitely many possible assignments of values to the envelopes is equally probable**. But this does not seem to make sense - what would the probability assigned to each be?

It can't be 0, since you think that any of the possible assignments has some chance of being the actual one. And it can't be some finite number, because, since you know that there is only one actual assignment of values to the envelopes, the probabilities must some to 1. And, for any finite number n, $n * \infty = \infty$

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 particular 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 might seem to be a devastating objection to the paradoxical arguments. But in fact, it is an objection which can be answered by complicating the relevant probability assignments.

Let's suppose that we are considering the version of the example discussed above, on which the values of the envelopes can be any power of 2.Then one might assign probabilities to the lower value of the two envelopes in the following manner:

$$\begin{aligned}
\Pr(\text{lower}{=}1{=}2^0) &= \tfrac{1}{4} * \tfrac{3}{4}^0 = \tfrac{1}{4} \\
\Pr(\text{lower}{=}2{=}2^1) &= \tfrac{1}{4} * \tfrac{3}{4}^1 = \tfrac{3}{16} \\
\Pr(\text{lower}{=}4{=}2^2) &= \tfrac{1}{4} * \tfrac{3}{4}^2 = \tfrac{9}{64} \\
\Pr(\text{lower}{=}8{=}2^3) &= \tfrac{1}{4} * \tfrac{3}{4}^3 = \tfrac{27}{256}
\end{aligned}$$

So it seems that the paradox, as presented, requires you to think that **each of the infinitely many possible assignments of values to the envelopes is equally probable**. But this does not seem to make sense - what would the probability assigned to each be?

It can't be 0, since you think that any of the possible assignments has some chance of being the actual one. And it can't be some finite number, because, since you know that there is only one actual assignment of values to the envelopes, the probabilities must some to 1. And, for any finite number n, n * ∞ = ∞

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 particular 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 might seem to be a devastating objection to the paradoxical arguments. But in fact, it is an objection which can be answered by complicating the relevant probability assignments.

Let's suppose that we are considering the version of the example discussed above, on which the values of the envelopes can be any power of 2.Then one might assign probabilities to the lower value of the two envelopes in the following manner:

$$\Pr(\text{lower}{=}1{=}2^0) = \tfrac{1}{4} * \tfrac{3}{4}^0 = \tfrac{1}{4}$$
$$\Pr(\text{lower}{=}2{=}2^1) = \tfrac{1}{4} * \tfrac{3}{4}^1 = \tfrac{3}{16}$$
$$\Pr(\text{lower}{=}4{=}2^2) = \tfrac{1}{4} * \tfrac{3}{4}^2 = \tfrac{9}{64}$$
$$\Pr(\text{lower}{=}8{=}2^3) = \tfrac{1}{4} * \tfrac{3}{4}^3 = \tfrac{27}{256}$$

That is, one might assign probability ¼ to the claim that the envelopes contain $1/$2, 3/16 to the claim that they contain $2/$4, and so on.

This probability assignment has two crucial features. First, the probabilities assigned form an infinite series which sums to 1. Second, despite the fact that higher assignments are now regarded as less probable, the probabilities are still close enough to generate the paradoxical arguments above.

For example, suppose that you find $2 in your envelope. Then you know that the assignment must either be $1/$2 or $2/$4. Given the above probability assignment, you go in regarding the former possibility as being more likely than the latter - more exactly, you think that the odds are 4 to 3 in favor of the former. So you think that there is a 4/7 chance that the other envelope contains $1, and only a 3/7 chance that it contains $4. So the expected utility of switching is 3/7*$4 + 4/7 *$1 = $16/7 = (roughly) $2.29 - so the argument in favor of switching still goes through. This will be true no matter what value you find in your envelope.

So it seems that the paradox, as presented, requires you to think that **each of the infinitely many possible assignments of values to the envelopes is equally probable**. But this does not seem to make sense - what would the probability assigned to each be?

This might seem to be a devastating objection to the paradoxical arguments. But in fact, it is an objection which can be answered by complicating the relevant probability assignments.

Let's suppose that we are considering the version of the example discussed above, on which the values of the envelopes can be any power of 2.Then one might assign probabilities to the lower value of the two envelopes in the following manner:

$$\Pr(\text{lower}{=}1{=}2^0) = \tfrac{1}{4} * \tfrac{3}{4}^0 = \tfrac{1}{4}$$
$$\Pr(\text{lower}{=}2{=}2^1) = \tfrac{1}{4} * \tfrac{3}{4}^1 = \tfrac{3}{16}$$
$$\Pr(\text{lower}{=}4{=}2^2) = \tfrac{1}{4} * \tfrac{3}{4}^2 = \tfrac{9}{64}$$
$$\Pr(\text{lower}{=}8{=}2^3) = \tfrac{1}{4} * \tfrac{3}{4}^3 = \tfrac{27}{256}$$

That is, one might assign probability ¼ to the claim that the envelopes contain $1/$2, 3/16 to the claim that they contain $2/$4, and so on.

This probability assignment has two crucial features. First, the probabilities assigned form an infinite series which sums to 1. Second, despite the fact that higher assignments are now regarded as less probable, the probabilities are still close enough to generate the paradoxical arguments above.

For example, suppose that you find $2 in your envelope. Then you know that the assignment must either be $1/$2 or $2/$4. Given the above probability assignment, you go in regarding the former possibility as being more likely than the latter - more exactly, you think that the odds are 4 to 3 in favor of the former. So you think that there is a 4/7 chance that the other envelope contains $1, and only a 3/7 chance that it contains $4. So the expected utility of switching is 3/7*$4 + 4/7 *$1 = $16/7 = (roughly) $2.29 - so the argument in favor of switching still goes through. This will be true no matter what value you find in your envelope.

Obviously, thinking of the original arguments in terms of this probability distribution makes things much, much more complicated - so I will return to thinking of things in terms of the simpler "half chance that it contains double, and half that it contains half" formulation. But I think that the above is enough to show that there is no fundamental problem with the initial probabilities which the paradox requires.

Let's now think about what it would take to **solve** this paradox.

Let's now think about what it would take to **solve** this paradox.

Perhaps the most problematic example we discussed was the choice closed version:

> **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?

One problem here is that we seem to have a compelling argument for switching - a compelling argument for thinking that it would be rational to believe that you stand to gain by switching (and switching again, and again, and again …).

One important thing to see is that this paradox is not solved by giving an argument for the **opposite conclusion**. That is very easy to do; for example: "Whatever the amount of money in the envelopes, there is some difference between them - call it D. If you have the lesser amount, you stand to gain D, and if you have the greater amount, you stand to lose D. But you have the same chance (given that your choice was at random) of having the lesser and the greater; hence you have the same chance (½) of gaining and losing D. Hence the expected utility of switching is 0."

This is a good argument against switching in this case - but we already knew that one should not switch in this case. What we want to solve the paradox is not an argument against switching, **but an explanation of what went wrong in the argument for switching**. One does not respond to Zeno's paradoxes of motion simply by saying "Yes, but I walked to school today" - this would be a perfectly good argument for the reality of motion, but does not give us a response to Zeno's arguments for its impossibility.

So let's return to the argument for switching in the choice closed version. Remember that it went via the argument for switching in the choice open version:

> **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?

The idea was that the argument for switching in that case did not depend on the particular value found in the envelope - it would be rational to switch **no matter what value was found**.

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?

So let's return to the argument for switching in the choice closed version. Remember that it went via the argument for switching in the choice open version:

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?

The idea was that the argument for switching in that case did not depend on the particular value found in the envelope - it would be rational to switch **no matter what value was found**.

Why does this support switching in the choice closed version? It seems that we are relying on some principle of the following sort:

Inference from an unknown

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 truth about the situation which you do not know but which is such that, were you to come to know it, it would be rational for you to do act 2.

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.

However, this sort of response to the choice closed version leaves us with two residual puzzles.

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?

However, this sort of response to the choice closed version leaves us with two residual puzzles.

First, it leaves unresolved our puzzles about the choice open version and the choice open reverse version. We gave arguments for switching in the former case and not switching in the latter which did not involve use of inference from an unknown.

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?

The choice open reverse version

Suppose again 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?

If the rejection of inference from an unknown were the whole story, then we would be stuck with these results. But this seems very odd. How could opening an envelope make it rational to pay someone to switch envelopes? How could the other person opening their envelope make it rational for you to pay them not to switch?

Another way to see the oddness here is to 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?

So one worry is that we have at least near-paradoxical conclusions with the open versions alone. A second worry is that it is just hard to see **how** inference from an unknown could be false.

So one worry is that we have at least near-paradoxical conclusions with the open versions alone. A second worry is that it is just hard to see **how** inference from an unknown could be false.

> ***Inference from an unknown***
>
> 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 truth about the situation which you do not know but which is such that, were you to come to know it, it would be rational for you to do act 2.

If there really is some truth knowledge of which would make it rational to perform some action, why isn't knowing that there is such a truth by itself enough to make it rational to perform the action? Suppose we were at the horse races, and I said: "I have some information about horse #2. Trust me - if you knew it, you would be rational to bet on him to win." **Isn't that enough to make it rational for you to bet on #2 now, even though I have not told you what I know?**

Perhaps inference from an unknown is usually OK, but can lead one astray in very special cases. Perhaps 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 conclusion that you ought to believe that you stand to gain by 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.

One thing you might want to think about is: are there any other cases in which unrestricted use of inference form an unknown would lead to contradictions of this sort? That is, are there other cases in which we know that there are two truths about our situation such that, if we were to learn the first but not the second, it would be rational to do one thing, and if we were to learn the second but not the first, it would be rational to do the other?