The two envelope paradox and infinite expectations

Frank Arntzenius & David McCarthy

Someone has written two cheques, one for twice as much as the other, and placed them randomly in two identical looking envelopes, Left and Right. You are allowed to choose an envelope, and keep the cheque it contains. For no particular reason you choose R. But now you are offered the chance to change your mind and switch to L, and the *switching argument* is pointed out to you.

Let x be the amount of money in R. Then there are two possibilities. Either R contains the smaller amount, in which case L contains $2x$, or R contains the larger amount, in which case L contains $x/2$. R is no more likely to contain the smaller amount than L, so the probability of each possibility is 1/2. So the expected monetary value of switching to L is $1/2.(2x) + 1/2.(x/2)$, which is $5x/4$, which is greater than x. So the expected monetary value of switching exceeds the monetary value of staying, no matter what x is, so it is rational to switch to L. But how can that be if there was nothing to choose between the two envelopes to start with, and you have gained no new information since? This is the *two envelope paradox*.[1]

1. That there is something wrong with the switching argument can be seen by a simple example. Suppose that you believe that the only possible values of x are 1 and 2. Then it is not true that whatever x is, the expected value of switching will exceed x: it is not true for $x=2$. Since what you believe is consistent with the description of the set-up, it follows that the switching argument is unsound.

Are there *any* sets of beliefs about what might be in the envelopes which are consistent with the description of the set-up which lead to the paradoxical conclusion? Let us say that such a set makes you *vulnerable* to the switching argument. If you think that there is some non-zero probability that x is a certain value m, then the switching argument won't tell you that whatever x is, the expected value of switching always exceeds x, unless you believe that there is some non-zero probability that x is $2m$. Otherwise, if x is m, the expected value of switching is $m/2$.

Suppose you attach a non-zero probability to x being 1. Then for the switching argument to work on you, you must attach non-zero probabilities p_0, p_1, …, p_n, … to the following unordered envelope possibilities: $(1,2)$, $(2,4)$, …, $(2^n,2^{n+1})$, … . So you will be vulnerable to the switching

[1] As far as we know, the paradox first appears in Nalebuff 1989.

Analysis 57.1, January 1997, pp. 42–50.

argument only if you believe there is no upper bound to the amount of money in the world.

All that just establishes a constraint on the probability distributions you must hold to be vulnerable to the switching argument. But it doesn't show there are any such probability distributions. Are there? Yes. We will establish a necessary and sufficient condition on p_0, p_1, p_2, ... for you to be vulnerable to the switching argument, and then show that there are probability distributions which satisfy the condition.

To make the argument more perspicuous, we are going to make two simplifying assumptions. First, we will assume that 1 is the smallest value of x to which you attach a non-zero probability. Second, we will assume that the only values of x to which you attach non-zero probabilities are of the form 2^n. The argument that follows can be made essentially unchanged if we drop these assumptions, but it's messier.

We are assuming that x is 2^n for some n. If $n=0$, then $x=1$ and the expected monetary value of switching is 2, so suppose $n > 0$. We want to find the expected monetary value of y given that x is 2^n. There are then two possibilities for y, 2^{n+1} and 2^{n-1}. So

$$(1)\quad \mathrm{E}(y \mid x=2^n) = \Pr(y=2^{n+1} \mid x=2^n).2^{n+1} + \{1 - \Pr(y=2^{n+1} \mid x=2^n)\}.2^{n-1}$$

Since the cheques were placed randomly in the envelopes, the probability of the ordered possibility $[2^n,2^{n+1}]$ is half that of the unordered possibility $(2^n,2^{n+1})$, i.e. $p_n/2$. So $\Pr(y=2^{n+1} \mid x=2^n) = \Pr(y=2^{n+1} \,\&\, x=2^n)/\Pr(x=2^n) = (p_n/2)/(p_{n-1}/2 + p_n/2)$. So we have

$$(2)\quad \Pr(y=2^{n+1} \mid x=2^n) = p_n/(p_{n-1} + p_n)$$

Substituting (2) into (1), with a little algebra, tells us

(3) $\mathrm{E}(y \mid x=2^n) > 2^n$ if and only if $p_n > p_{n-1}/2$.

Hence you will be vulnerable to the switching argument if and only if your probability distribution over envelope possibilities is such that $p_n > p_{n-1}/2$ for all n.

Are there any such probability distributions? Yes. Consider for example a distribution defined recursively by $p_0=1/4$ and $p_n=(3/4).p_{n-1}$ for $n>0$.[2] Let us call such a distribution *paradoxical*.

2. Let us state the switching argument more precisely. Assume you hold a paradoxical probability distribution. Then the total outcome space of

[2] This is a trivial variation of an example Broome gives (see his 1995: 7). In fact, any probability distribution defined recursively by $p_0 = q$ and $p_n = r.p_n - 1$ for $n > 0$ where $q + r = 1$ and $1 > r > 1/2$ will make you vulnerable to the switching argument, as well as a host of less well behaved distributions.

envelope possibilities is {[1,2], [2,1], [2,4], [4,2], ...}, where [1,2] means \$1 in L, \$2 in R. Then there is a partition Π_R of this outcome space according to the amount of money that is in R: {{[2,1]}, {[1,2], [4,2]}, {[2,4], [8,4]}, ...}. For any element P of this partition, if you hold a paradoxical probability distribution, then the expected value of L conditional on P obtaining is greater than the expected value of R conditional on P obtaining. Broome and others have claimed that this appears to entail that it would be irrational for you to stay with R (Broome 1995: 8).

But there had better be something wrong with this reasoning, for it generates a clear inconsistency. There is another partition Π_L of this outcome space according to the amount of money that is in L: {{[1,2]}, {[2,1], [2,4]}, {[4,2], [4,8]}, ...}. For any element Q of this partition, if you hold a paradoxical probability distribution, then the expected value of R conditional on Q obtaining is greater than the expected value of L conditional on Q obtaining. Exactly the same reasoning would then tell you that it would be irrational for you to switch to L. So if this reasoning is valid, there are probability distributions which entail that it is both irrational to stay, and irrational to switch!

Standard decision theory says to go by expected value: given a choice of two actions, it is rational to choose one whose expected value is at least as great as the other. But the reasoning we have just seen talks about conditional expectations, so to apply standard decision theory, we need to look at the unconditional expectations of staying and switching.

One might claim that the expected value of switching exceeds the expected value of staying by claiming the following:

(4) If there is a partition Π of a total outcome space S such that for every element P_j of Π, the expectation of X conditional upon P_j obtaining is finite and is greater than the expectation of Y conditional upon P_j obtaining, then the (unconditional) expectation of X is greater than the (unconditional) expectation of Y.

At this point someone might object that standard decision theory tells you to go by expected utility value, not expected monetary value, so strictly speaking, (4) does not entail that it is irrational to stay or that it is irrational to swtich. But this is of no real help. It follows from (4) and facts about the partitions that if you hold a paradoxical distribution, the expected monetary value of switching is both greater and less than the expected monetary value of staying. Not a problem for standard decision theory, perhaps, but still an inconsistency.

Let us look more closely at (4). What exactly is the connection between conditional and unconditional expectations? The following looks plausible.

(5) Let Q be some random variable over an outcome space S. Then for any partition Π of S, the unconditional expectation value E(Q) equals the sum $\Sigma_j E(Q|P_j)Pr(P_j)$, where one sums over all the elements P_j of Π.

But even if (5) is true, (4) does not follow. It would only follow if it was in general true that if one has two sequences (A_j) and (B_j) of finite terms such that for each j, $A_j > B_j$, then $\Sigma_j A_j > \Sigma_j B_j$. But this is not in general true for infinite sequences: consider the sequences 2, 2, 2, … and 1, 1, 1, … . And in fact, as we will prove later, the unconditional expectation of switching and the unconditional expectation of staying are each infinite, so neither is greater than the other, so (4) yields the wrong conclusion about the unconditional expectations of staying and switching. Moreover, since these expectations are infinite, standard decision theory neither says that it is irrational to switch, nor that it is irrational to stay. One might baulk at talk of infinite expectations, but we will argue later that there is no need to do so.

There is no doubt that (4) is false, and that pointing out that the expectation of staying and the expectation of switching are infinite blocks the switching argument as we described it. But the switching argument can be stated in another way without appealing to (4), and the most general analysis of the two envelope paradox must be able to show what is wrong with this revised version. Any clarity we have on this point is due to John Norton, who has provided such an analysis (Norton 1996), on which we rely in the following.

Consider the difference between what is in L and what is in R, i.e. consider the random variable D = L – R. If the expectation E(D) is positive, it would appear to be rational to switch, even in the case in which E(D) is (positively) infinite. But if (5) is true, then E(D) is positive. For consider the partition Π_R. Within each member of the partition, E(D) is positive. So by (5), the unconditional expectation E(D) is positive (in fact, positive infinity). The revised version of the switching argument therefore says that for this reason, it is rational to switch.

But something has gone seriously wrong, since exactly the same argument applied to the partition Π_L will yield the conclusion that the unconditional expectation E(D) is negative.

We went wrong in accepting (5). Very roughly speaking, the problem is as follows. The various possible differences D multiplied by their respective probabilities Pr(D), are like 1, –1, 2, –2, 4, –4, … . The argument that the expectation value is positive infinity relies on ordering these terms as follows: 1 + (2 – 1) + (4 – 2) + (8 – 4) + … . The argument that the expectation value is negative infinity relies on ordering these terms as follows: –1 + (–2 + 1) + (–4 + 2) + (–8 + 4) + … . But for the sum of a countable set

of terms to be well-defined it must not depend on the order in which the terms are summed. So E(D) is undefined. This shows why (5) is false and why the revised version of the switching argument is unsound.

3. Broome and others have suggested that there are strong similarities between the two envelope paradox and the St. Petersburg paradox.[3] Indeed, one way of proving as we claimed earlier that the expected (dollar) value of switching is infinite if you are vulnerable to the switching argument is by showing how the set up of the two envelope paradox can be changed fairly simply to get a St. Petersburg paradox.

Suppose you are vulnerable to the switching argument, and suppose that instead of initially being offered your choice of envelopes for nothing, you were offered the chance to pay to pick an envelope. How much should you pay?

To be conservative, assume you will pick the envelope with the smaller of the two cheques. Then you will win exactly \$1 with probability p_0, \$2 with probability p_1, …, $\$2^n$ with probability p_n, and so on, for all n. But since you are vulnerable to the switching argument, for all $n > 0$, $p_n > p_0/2^n$. So this gamble is better than a gamble in which you win exactly \$1 with probability p_0, \$2 with probability $p_0/2$, …, $\$2^n$ with probability $p_0/2^n$, and so on, for all n. But this is just a St. Petersburg gamble, a gamble with infinite expectation, and the fact that, if you are like most people, you would not be prepared to pay very much, is what makes the St. Petersburg paradox.

Broome briefly makes two claims about the relation between the two envelope paradox and the St. Petersburg paradox (1995: 9).The first is this.

(6) The St. Petersburg paradox relies on infinite expectation values, but the two envelope paradox does not. Since infinite expectation values are not strictly speaking well-defined, the two envelope paradox is in one respect more powerful.

We say:

(6*) It is sometimes useful to allow infinite expectation values, and doing so makes standard decision theory more powerful. Moreover the St. Petersburg paradox can be stated without relying on infinite expectation values.

To ban all talk of infinite expectations (in the limiting sense) is unnecessarily strong medicine that precludes us from saying things that are perfectly sensible and useful – and hence should not always be eschewed – for it enables us to represent certain facts about rational preferences.

[3] Broome 1995: 8–9. Broome attributes noting the connection to Doug MacLean.

Suppose that there is some quantity Q (dollars, utilities, or otherwise), of which all possible values are finite and positive, such that for any gambles with finite Q-expectations you always prefer the higher Q-expectation, and you are indifferent between any gambles of equal Q-expectations. Now let F be a Q-gamble with finite Q-expectation, and I a Q-gamble with an infinite Q-expectation. Then it is very plausible that you prefer I to F. For I can be decomposed into the sum of two gambles, F* and I*, where the expected value of F* is finite and identical to the expected value of F, and the expected value of I* is infinite. Since the expected values of F* and F are identical you will be indifferent between F and F*. Moreover if you accept gamble I, for any possible circumstance the payoff of I is the sum of the payoffs of F* and I* in that circumstance (that is what we meant when we said that I can be decomposed into F* and I*). So if you accept I, then for any possible outcome you will get at least what you would get were you to accept F*. Clearly you should prefer I to F* and hence you should prefer I to F. It follows that you prefer I to any gamble over Q with finite expectation values. This can be naturally represented if we allow talk of infinite expectations, for we could then represent your preferences by adding that you prefer any Q-gamble with infinite expectation to any Q-gamble with finite expectation.

This is not to say that infinite expectations aren't strange in any way. For instance consider a St. Petersburg gamble G and another St. Petersburg gamble G′ arrived at by modifying G by doubling each of its possible payoffs, while keeping the probabilities the same. Clearly one would prefer G′ to G, and yet, both expectations are the same, in the sense that both have infinite expectations. Thus allowing claims of infinite expectation in the coarse way that we have suggested does not allow us to represent all our intuitive preferences about such gambles in terms of expectations.

For the second part of our claim, notice that for any finite amount of money, there is always an initial segment of the standard St. Petersburg gamble whose expected monetary value is finite but exceeds that amount. Suppose you were offered the chance to pay for the initial segment of a St. Petersburg gamble where you get to choose how long the initial segment will be. Then no matter how much you pay, you can always choose an initial segment whose expectation is finite but far exceeds that amount. But if you are like most people, you won't be prepared to pay very much, and this brings out the essence of the original St. Petersburg paradox without talking of infinite expectation values.

Broome's second claim is this.

(7) The St. Petersburg paradox and the two envelope paradox are very similar. Solutions to the St. Petersburg paradox will also dissolve the two envelope paradox. Nevertheless, these solutions

> are unsatisfactory in that there are possible worlds in which both the two envelope paradox and the St. Petersburg paradox remain.

We say:

> (7*) There is nothing intrinsically paradoxical about the St. Petersburg paradox, so solutions to it are not needed. But the two envelope paradox is worse than a paradox: the reasoning that leads you to it leads to an inconsistency. But a solution to this exists even if the standard solutions to the St. Petersburg are unavailable.

The St. Petersburg phenomenon can be explained by saying that you distinguish between value and money, and that the value you attach to dollar gambles is bounded. Another version of the St. Petersburg paradox makes two assumptions. First, there is something – happiness, say – of which you prefer more to less and of which there is no limit to the amount God can dish out. Second, you are indifferent between happiness gambles which have the same finite expected payoff. Now suppose God offers you a St. Petersburg paradox where the payoffs are in units of happiness. Then it is very plausible that rationality requires you to be willing to pay an arbitrarily large amount to play this gamble. But there is nothing intrinsically paradoxical about this – it just reflects your preferences.

Now if the value you attach to all dollar gambles is bounded, you will not be vulnerable to the switching argument. But if there are possible worlds in which the two assumptions about happiness in the previous paragraph are true, then there will be paradoxical probability distributions which would make you vulnerable to the switching argument, if it were a valid argument. So it is still important to show that the argument rests on a fallacy, and that is what we claim to do in this paper.

Finally, let us note that both the two envelope paradox and the St. Petersburg paradox share a feature which can be used to generate a different sort of paradox. Assume that the two assumptions about happiness are true, and that the two envelope paradox and St. Petersburg paradox both have payoffs in units of happiness. Suppose we were told of the contents of our envelope and not told of the contents of the other envelope. If we hold a paradoxical probability distribution, we will immediately prefer to switch to the other envelope, since the expected value of switching is higher than the expected value of staying. And if instead we were told only of the contents of the other envelope, we would prefer to stay where we are. Bizarre, but not inconsistent. Similarly, suppose we accept a St. Petersburg gamble. As soon as we are told of the outcome we will be disappointed in the following sense. We would prefer to abandon our winnings for another shot at the gamble, no matter how much we won. Bizarre, but not incon-

sistent. The bizarreness of this case, we think, derives from the fact that we know in advance that no matter what we get, it will be less than the expectation value of the gamble. This is strange indeed, but not inconsistent, and should not prevent us from talking sensibly of infinite expectations.

Still, there lurks a paradox which may be a genuine version of what the standard St. Petersburg paradox gestures towards, the *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. You are indifferent between any two gambles with the same expected value, and you prefer a gamble with a greater expected value to a gamble with lesser expected value. Furthermore, it is natural to extend this to infinite expectations in the way indicated earlier.

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 finite 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, he'll give you another shot at the St. Petersburg gamble. But if you decline, after your finite stay in Heaven you'll 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 finite 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.[4] What's gone wrong?

University of Southern California
arntzeni@mizar.usc.edu

Johns Hopkins University
dmccarth@sph.jhu.edu

References

Broome, J. 1995. The two-envelope paradox. *Analysis*, 55: 6–11.
Nalebuff, B. 1989. The other person's envelope is always greener. *Journal of Economic Perspectives* 3: 171–81.
Norton, J. 1996. When the sum of our expectations fails us: the exchange paradox. (Manuscript.)

[4] Note that God didn't have to offer you a gamble with infinite expectations to get this effect. He could have started out by offering you one day in Heaven, then the next day three in heaven in return for a day in Hell, then the next five in Heaven in return for a day in Hell, and so on. What is important is that there is no limit to how many days in Heaven he can offer you.