The Two-Envelope Paradox: A Complete Analysis?

David J. Chalmers

Department of Philosophy

Washington University

St. Louis, Missouri 63130

dave@twinearth.wustl.edu

March 22, 1994

A wealthy eccentric places two envelopes in front of you. She tells you that both envelopes contain money, and that one contains twice as much as the other, but she does not tell you which is which. You are allowed to choose one envelope, and to keep all the money you find inside.

This may seem innocuous, but it generates an apparent paradox. Say that you choose envelope 1, and it contains \$100. In evaluating your decision, you reason that there is a 50% chance that envelope 2 contains \$200, and a 50% chance that it contains \$50. In retrospect, you reason, you should have taken envelope 2, as its expected value is \$125. If your sponsor offered you the chance to change your decision now, it seems that you should do so. Now, this reasoning is independent of the actual amount in envelope 1, and in fact can be carried out in advance of opening the envelope; it follows that whatever envelope 1 contains, it would be better to choose envelope 2. But the situation with respect to the two envelopes is symmetrical, so the same reasoning tells you that whatever envelope 2 contains, you would do better to choose envelope 1. This seems contradictory. What has gone wrong?

The paradox can be expressed numerically. Let A and B be the amounts in envelope 1 and 2 respectively; their expected values are $E(A)$ and $E(B)$. For all n, it seems that $p(B > A|A = n) = 0.5$, so that $E(B|A = n) = 1.25n$. It follows that $E(B) = 1.25E(A)$, and therefore that $E(B) > E(A)$ if either expected value is greater than zero. The same reasoning shows that $E(A) > E(B)$, but the conjunction is impossible, and in any case $E(A) = E(B)$ by symmetry. Again, what has gone wrong?

This problem has been discussed in the pages of *Analysis* by Jackson, Menzies and Oppy [2], and by Castell and Batens [1], but for reasons that will become clear I think that their analyses are incomplete and mistaken respectively, although both contain insights that are important to the resolution of the problem. I will therefore present my own analysis of the "paradox" below.

Some distractions inessential to the problem arise from the facts that in the real world, money comes in discrete amounts (dollars and cents, pounds and pence) and that there are known limits on the world's money supply. We can remove these distractions by stipulating that for the purposes of the problem, the amounts in the envelopes can be any positive real number.

There are a number of steps in the resolution of the paradox. The first step is to note (as do the authors mentioned above) that the amounts in the envelopes do not fall out of the sky, but must be drawn from some probability distribution. Let the relevant probability density function be g, where the probability that the smaller amount falls between a and b is $\int_a^b g(x)dx$. We can think of this distribution as either representing the chooser's prior expectations, or as the distribution from which the actual values are drawn. I will generally write as if it is the second, but nothing much rests on this. To fix ideas, we can imagine that our sponsor chooses a random variable Z with probability density g, and then flips a coin. If the coin comes up heads, she sets $A = Z$ and $B = 2Z$; if it comes up tails, she sets $A = 2Z$ and $B = Z$.

Recognizing the existence of a distribution immediately shows us that the reasoning that leads to the paradox is not always valid, as Jackson *et al* note. For example, if the distribution is a uniform distribution over values between 0 and 1000, with amounts over 1000 being impossible, then if $A > 500$, it is always a bad idea to switch. It is therefore not true that for all distributions and all values of n, $p(B > A|A = n) = 0.5$. In general, $E(B|A = n)$ will not depend only on n; it will also depend on the underlying distribution.

In their analysis, Jackson *et al* are satisfied with this observation, combined with the observation that limitations on the worlds' money supply ensure that in practice the relevant distributions will always be bounded above and below. The paradox does not arise for bounded distributions, as we saw above. When A is a medium value, there may be equal chances that B is larger or smaller, but when A is large B is likely to be smaller, and when A is small B is likely to be larger, so the paradox does not get off the ground.

This practical observation is an insufficient response to the mathematical paradox, however, as Castell and Batens note. Unbounded distributions can exist in principle if not in practice, and in-principle existence is all that is needed for the paradox to have its bite. For example, it might seem that if the distribution were a uniform distribution over the real numbers, then $p(B > A|A = n) = 0.5$ for all n. This would seem to have paradoxical consequences for mathematics, if not for the world's money supply.

This leads to the second step in the resolution of the paradox, which is that taken by Castell and Batens. (We will see that this step is ultimately inessential to the paradox's resolution, but it is an important intermediate point of enlightenment.) There is in fact no such thing as a uniform probability distribution over the real numbers. To see this, let g be a uniform function over the real numbers. Then $\int_k^{k+1} g(x)dx$ is equal to some constant c for all k. If $c = 0$, then the area under the entire curve will be zero, and if $c > 0$, then the area under the entire curve will be infinite, both of which contradict the requirement that the integral of a probability distribution be 1. At one point Jackson *et al* raise the possibility of infinitesimal probabilities, but if this is interpreted as allowing c to be infinitesimal, the suggestion does not work any better. To see this, note that if the distribution is uniform:

$$\begin{aligned}
\int_0^\infty g(x)dx &= \int_0^1 g(x)dx + \int_1^2 g(x)dx + \int_2^3 g(x)dx + \cdots \\
&= \int_0^1 g(x)dx + \int_2^3 g(x)dx + \int_4^5 g(x)dx + \cdots \\
&= 1/2 \int_0^\infty g(x)dx,
\end{aligned}$$

so that the overall integral must be zero or infinite. A uniform distribution over the real numbers can only be an "improper" distribution, whose overall integral is not 1.

The impossibility of a uniform probability distribution over the real numbers is reflected in the fact that every proper distribution must eventually "taper off": for all $\epsilon > 0$, there must exist k such that $\int_k^\infty g(x)dx < \epsilon$. It is very tempting to suppose that this "tapering off" supplies the resolution to the paradox, as it seems to imply that if A is near the high end of the (proper) distribution, it will be more likely that B is smaller; perhaps sufficiently more likely to offset the paradoxical reasoning? This is the conclusion that Castell and Batens draw. They offer a "proof" that the distribution must be improper for the paradoxical reasoning to be possible.

Unfortunately Castell and Batens' proof is mistaken, and in fact there exist proper distributions for which the paradoxical reasoning is possible. The error lies in their assumption, early in the paper, that

$$p(B > A|A = n) = \frac{g(n)}{g(n) + g(n/2)}.$$

This seems intuitively reasonable, but in fact

$$p(B > A|A = n) = \frac{2g(n)}{2g(n) + g(n/2)},$$

which is significantly larger in general.

To see this, note that if A is in the range $n \pm dx$, then B is either in the range $2n \pm 2dx$ or in the range $n/2 \pm dx/2$. The probability of the first, relative to the initial distribution, is $g(n)dx$; the probability of the second is $g(n/2)dx/2$. The probabilities that B is greater or less than A therefore stand in the ratio $2g(n) : g(n/2)$, not $g(n) : g(n/2)$, as Castell and Batens suppose.

For example, given a uniform distribution between 0 and 1000, if A is around 100, it is in fact twice as likely that B is around 200 than that B is around 50. To dispel any lingering counterintuitiveness, note that something like this *has* to be the case to make up for the fact that when $A > 500$, B is always less than A. To find a distribution where the chances of a gain and a loss are truly equal for many n, we should turn not to a uniform distribution but to a decreasing distribution, where $g(n/2) = 2g(n)$ for many n. An example is the distribution $g(x) = 1/x$, where we cut off the distribution between arbitrary bounds L and U, and normalize so that it has an integral of 1. This distribution will have the property that for all n such that $2L < n < U/2$, $p(B > A|A = n) = 0.5$. To illustrate this intuitively, note that for such a decreasing distribution, the prior probability that the smaller value is between 4 and 8 is the same as the probability that it is between 8 and 16, and so on, if L and U are appropriate. Given the information that $8 < A < 16$, it is equally likely that B is in the range above or below.

This flaw in Castell and Batens' reasoning nullifies their proof that a distribution must be improper for the paradoxical reasoning to arise, but it does not yet show that the conclusion is false. It remains open whether there is a proper distribution for which the paradoxical reasoning is possible. The bounded distribution above will not work, as its bound will block the paradoxical reasoning in the usual fashion; and the *unbounded* distribution $g(x) = 1/x$ is improper, having an infinite integral. But this can easily be fixed, by allowing the distribution to taper off slightly faster. In particular, the distribution $g(x) = x^{-1.5}$, cut off below a lower bound L and normalized, allows the paradox to arise. The distribution has a finite integral, and even though for most n, $p(B > A|A = n) < 0.5$, it is still the case that for all relevant n, $E(B|A = n) > n$. To see this, note that if $n < 2L$, then $E(B|A = n) = 2n$; and if $n \geq 2L$, then

$$\begin{aligned} p(B > A|A = n) : p(B < A|A = n) &= 2g(n) : g(n/2) \\ &= 2n^{-1.5} : (n/2)^{-1.5} \\ &= 1 : \sqrt{2}. \end{aligned}$$

The expected value $E(B|A = n)$ is $(2n + \sqrt{2}n/2)/(1 + \sqrt{2})$, which is about $1.12n$. The paradox therefore still arises.

The distribution here may be unintuitive, but it is easy to illustrate a similar distribution intuitively. Take a distribution in which the probability of a value between 1 and 2 is c, the probability of a value between 2 and 4 is just slightly less, say $0.9c$, the probability of a value between 4 and 8 is $0.81c$, and so on. This distribution has a finite integral, as the integral is the sum of a decreasing geometric series; and it is sufficiently close to the case in which the probability of a value between 2^k and 2^{k+1} is constant that the paradoxical reasoning still arises. Even though $p(B < A|A = n)$ is now slightly less than 0.5, due to the incorporated factor of 0.9, it has decreased by a sufficiently small amount that $E(B|A = n)$ remains greater than n. The case $g(x) = x^{-1.5}$ is just like this, except that the factor of 0.9 is replaced by a factor of $1/\sqrt{2}$, which is around 0.7.

The paradox has therefore not yet been vanquished; there are perfectly proper distributions for which the paradoxical reasoning still applies. This leads us to the third and final step in the resolution of the paradox. Note that although the distributions above have finite integrals, as a probability distribution should, they have infinite *expected value*. The expected value of a distribution is $\int_0^\infty xg(x)dx$. When $g(x) = x^{-1.5}$ (cut off below L), the expected value is $\int_L^\infty x^{-0.5}dx$, which is infinite. But if the expected value of the distribution is infinite, there is no paradox! There is no contradiction between the facts that $E(B) = 1.12E(A)$ and $E(A) = 1.12E(B)$ if both $E(A)$ and $E(B)$ are infinite. Rather, we have just another example of a familiar phenomenon, the strange behavior of infinity.[1]

To fully resolve the paradox, we need only demonstrate that for distributions with finite expected value, the paradoxical situation does not arise. To do this, we need to precisely state the conditions expressing the paradoxical situation. In its strongest form, the paradoxical situation arises when $E(B|A = n) > n$ for all n. However, it arises more generally whenever reasoning from B's dependence on A leads us to the conclusion that there is expected gain *on average* (rather than all the time) by switching A for B. This will hold whenever $E(K-A) > 0$, where K is the random variable derived from A by the transformation $x \mapsto E(B|A = x)$. We therefore need to show that when $E(A)$ is finite, $E(K - A) = 0$.

Let h be the density function of A. Then $h(x) = 1/2.(g(x) + g(x/2)/2) = 1/4.(2g(x) +$

[1]Castell and Batens note some similar consequences of infinite expected values in another context, in which the distribution is over a countable set. They say that infinite expected values are "absurd", but I do not see any mathematical absurdity.

$g(x/2))$. (Note that $h \neq g$, as g is the density function of the *smaller* value.) Then

$$
\begin{aligned}
E(K-A) &= \int_0^\infty h(x)(E(B|A=x)-x)dx \\
&= \int_0^\infty \frac{2g(x)+g(x/2)}{4}.(\frac{2x.2g(x)+x/2.g(x/2)}{2g(x)+g(x/2)}-x)dx \\
&= \int_0^\infty \frac{2xg(x)-x/2.g(x/2)}{4}dx \\
&= 1/4.(\int_0^\infty 2xg(x)dx - \int_0^\infty 2yg(y)dy) \\
&= 0.
\end{aligned}
$$

Note that the fourth and fifth steps above are valid only if $\int_0^\infty xg(x)dx$ is finite, which holds iff $E(A)$ is finite. (If $\int_0^\infty xg(x)dx$ is infinite, it is possible that $\int_0^\infty 2xg(x) - x/2.g(x/2)dx \neq 0$, even though $\int_0^\infty 2xg(x)dx = \int_0^\infty x/2.g(x/2)dx$.)

It follows that when $E(A)$ is finite, consideration of the dependence of B on A will not lead one to the conclusion that one should switch A for B. A colollary of the result is that when $E(A)$ is finite, it is impossible that $E(B|A=n) > n$ for all n, so that the strong form of the paradox certainly cannot arise.

If $E(A)$ is infinite, this result does not hold. In such a case, it is possible that $E(A) = E(K)$ (both are infinite) but that $E(K-A) > 0$. Here, the "paradoxical" reasoning will indeed arise. But now the result is no longer paradoxical; it is merely counterintuitive. It is a consequence of the fact that given infinite expectations, *any* given finite value will be disappointing. The situation here is somewhat reminiscent of the classical St. Petersburg paradox: both "paradoxes" exploit random variables whose values are always finite, but whose expected values are infinite. The combination of finite values with infinite expected values leads to counterintuitive consequences, but we cannot expect intuitive results where infinity is concerned.[2]

[2] I first learned of this paradox from Doug Hofstadter in 1988. I reached the conclusions contained in this paper in discussion with Hofstadter and a number of others in 1988 and 1989. The appearance of the problem in a philosophical journal prompts me to publish the analysis.

References

[1] P. Castell and D. Batens, 'The Two-Envelope Paradox: The Infinite Case'. *Analysis* 54:46-49.

[2] F. Jackson, P. Menzies, and G. Oppy, 'The Two Envelope "Paradox"', *Analysis* 54:43-45.