

# Symmetry and Bias: The Exchange Paradox

Anubav Vasudevan

## 1 The Exchange Paradox

### 1.1 The Paradox

You are handed two envelopes, you hold one in your left hand and one in your right, and you are told that each contains a ticket on which is written some integer-valued number. You will be allowed to select one of the two envelopes and keep the ticket inside. If the envelope you choose contains the ticket with the greater of the two numbers written on it, you can exchange the ticket for a cash prize, the value of which *increases* with the number of the winning ticket.<sup>1</sup> The ticket with the lesser number written on it is worth nothing. The only information you are given concerning the contents of the two envelopes is that the difference between the two numbers is 1. Having no good reason to prefer one envelope to the other, you select on a whim (or perhaps on the basis of a coin flip) the envelope in your left hand. Before opening it, however, you are given the option to exchange it for the envelope in your right hand. Do you have any reason to accept the offer to switch?

Intuitively, the answer to this question is no, however, the following argument seems to suggest otherwise: Suppose that the number of the ticket in your envelope is  $n$ . Then the number of the ticket in the other envelope is either  $n + 1$  or  $n - 1$ . If  $f$  is the function which determines the cash value of the winning ticket, then in the first case you stand to gain  $f(n + 1)$  by switching and in the second case switching will cost you  $f(n)$ . Since you have no reason to judge either one of these possibilities more likely than the other, you assign to each an equal probability so that your expected gain from switching envelopes is:

$$\frac{1}{2}f(n + 1) - \frac{1}{2}f(n)$$

Since, by assumption,  $f$  is a strictly increasing function your expected gain is positive, and so, assuming that  $n$  is the number of the ticket in your envelope, you ought to switch. But since this is true *for any* value of  $n$ , you ought to switch envelopes.

Clearly, there is something wrong with this line of reasoning, for after selecting one of the two envelopes in an arbitrary fashion and without receiving any additional information you now believe that you have good reason to switch! This is the simple form of the Exchange Paradox.

### 1.2 Replies to the Paradox

Where does the argument for switching envelopes go wrong? We may note right away that the argument for switching depends crucially upon the assumption that the number of the winning ticket can take any value *no matter how large*, for if there were some limit to how large this number could be, there would exist the possibility that your chosen envelope contains the largest possible ticket. But, if you were to assume this to be the case, you would know with certainty that the ticket in your envelope is the winner, in which case you clearly ought to refuse the exchange.

---

<sup>1</sup>Thus, for example, it is worth more to win with ticket number '63' than it is to win with ticket number '14' than it is to win with ticket number '-234'.

The paradox, therefore, can only arise in *infinite* settings,<sup>2</sup> and as a result it has generally come to be regarded as yet another in the long line of puzzling and counterintuitive phenomena associated with infinity. In what follows, I will argue that this standard view of the paradox is mistaken. The fallacy in the argument for switching does not result from the application of some principle of reasoning which, while valid in finite settings, breaks down when applied to infinite domains, but rather from the agent's basing his decision to switch on a *biased* deliberative methodology.

In outline form, the argument for switching envelopes can be expressed as follows:

- (P1) For any  $n$ , assuming that  $n$  is the number of the ticket in the agent's envelope, the number of the ticket in the other envelope is as likely to be  $n + 1$  as it is to be  $n - 1$ ; hence,
  - (P2) For any  $n$ , assuming that  $n$  is the number of the ticket in the agent's envelope, the agent ought to switch envelopes; hence,
- 
- (C) The agent ought to switch envelopes.

The vast majority of discussions of the paradox accept that the agent's inference from (P1) to (P2) is valid. That is, it is generally taken for granted that if the agent were to suppose that his envelope contains ticket number  $n$ , and if he is right to conclude that the other envelope is as likely as not to contain the winner, then he ought to agree to the exchange. Having accepted this, one is forced to conclude that the fallacy in the agent's reasoning must be located either in his acceptance of (P1) or in his inferring on the basis of (P2) that he ought to switch envelopes.

In order to explain why either one or both of these assumptions are unwarranted, the standard analyses of the paradox all appeal, in one way or another, to the infinitary nature of the setup. In the following two sections, I argue that these standard analyses of the paradox are unconvincing since they all rely on the unnecessarily strong assumption that the agent is a full-fledged Bayesian decision maker whose preferences are the result of maximizing expected utility with respect to some well-defined probability function. While the authors of these proposals are correct to point out that complications arise when the Bayesian framework is applied in decision contexts in which it is assumed that the world may be in any one of an infinite number of states, and that as a result of such complications, standard expected value reasoning cannot be used to underwrite the agent's decision to switch, they fail to provide a satisfactory resolution to the paradox since they overlook the fact that the agent's rationale for switching can be expressed in terms of much more elementary principles of reasoning which do extend naturally to the infinite case.

On the basis of this discussion, I conclude that the mistake in the agent's reasoning does, in fact, consist in his inference from (P1) to (P2). This may seem like a startling result since (P2) seems to follow from (P1) by means of a straightforward assessment of the agent's conditional expected gain. In the remainder of the paper, I offer an explanation of why the agent is wrong to base his decision on the result of such an assessment. My analysis is based on a certain construal of the *symmetry* inherent in the problem. The notion of symmetry which figures in this analysis is not the obvious evidential symmetry which characterizes the agent's epistemic state at the time of his decision. Rather, it is a certain, 'higher-order' symmetry which requires that the agent's decision to switch be based on a deliberative methodology which does not depend on his initial choice of envelope.

From this symmetry-based analysis it follows that the agent cannot assess his conditional expected gain in the straightforward manner proposed in the argument for switching, and in the concluding section of the paper, I offer an alternative account of how these expectations ought to be assessed. The proposed methodology represents an initial attempt to formalize the intuition that, in deciding whether or not to

---

<sup>2</sup>More specifically, the paradox can only arise if it assumed that the number of the winning ticket has no *upper* bound. The argument for switching does not, however, require that the number of the winning ticket is bounded below since if the agent assumes that his envelope contains the smallest possible ticket, he obviously ought to switch. We assume throughout that the number of the ticket is unbounded in both directions (the difference between the bounded and the unbounded case is discussed briefly in section 6).

switch envelopes, the agent ought to ‘take seriously’ what he would have thought had he instead chosen the envelope in his right hand. While the model suffers from some important defects, it does, I believe, capture some of our general intuitions about how a rational agent ought to adjust his expectations to correct for the effects of bias.

### 1.3 Prior Probabilities and Proper Forms of the Paradox

The first of the two standard objections to the simple form of the paradox is addressed to premise (P1), i.e., the agent’s assumption that whatever may be the number of the ticket in his envelope, the number of the ticket in the other envelope is as likely as not to be the greater of the two. As has often been pointed out, nothing about the description of the setup implies that this should be the case. If, for example, the agent assumes that the probability of the winning ticket’s being numbered  $n$  decreases with the absolute value of  $n$ , then provided the number of the agent’s ticket is positive, he ought to judge it *more* likely than not that his ticket is the winner.

Of course, it is not enough to simply point out that the agent could consistently adopt such a hypothesis in order to provide him with a reason to reject (P1), for perhaps there is some other more reasonable assumption about the probabilistic mechanism at work in the setup that could justify the agent in accepting this claim. This, however, turns out not to be the case, for there is *no* assignment of probabilities consistent with the description of the setup that can support the uniform conditional probabilities expressed in (P1).

The argument is as follows: Suppose that the number of the ticket in the agent’s envelope is  $n$ . Then, (P1) asserts that this ticket is as likely as not to be the winner. Now, since the agent’s envelope was chosen arbitrarily, this conclusion should apply no matter which of the two envelopes is in the agent’s possession. Thus, no matter which of the two envelopes contains ticket number  $n$ , this ticket is as likely as not to be the winner. But this amounts to the claim that the number of the winning ticket is as likely to be  $n$  as it is to be  $n + 1$ , and this holds for all values of  $n$ . Thus, it follows from (P1) that every possible value of the number of the winning ticket is *equally* likely to occur. But this cannot be, since there is no such thing as a uniform probability distribution over an infinite set.

To express this argument in more precise terms, let  $L$  and  $R$  be the numbers of the tickets contained in the chosen (left-hand) and unchosen (right-hand) envelopes, respectively, and let  $W = \max\{L, R\}$  be the number of the winning ticket. The possible contents of the two envelopes can be represented by the set  $\Omega$  of all pairs of integers which satisfy the condition  $|L - R| = 1$ , i.e.,  $\Omega = \{(0, 1), (1, 0), (0, -1), (-1, 0), \dots\}$ , where the first term in each pair gives the value of  $L$  and the second term the value of  $R$ .

Now, suppose that the agent assigns probabilities to the states in  $\Omega$  and that these probabilities are given by the function  $Pr$ . Since the agent’s choice between the two envelopes was arbitrary, whatever the number of the winning ticket may be, it is as likely as not that the agent’s envelope contains the winning ticket. In other words, for all  $n$ :

$$Pr(L = n|W = n) = Pr(R = n|W = n)$$

This is equivalent to the condition that:

$$(1) \quad Pr(n, n - 1) = Pr(n - 1, n)$$

In what follows, we will refer to any probability function that satisfies this condition as an ‘exchangeable’ probability. The condition that the agent’s probabilities be exchangeable is clearly implied by the description of the setup.<sup>3</sup> This condition, however, is not what is expressed by (P1). Rather, (P1)

---

<sup>3</sup>If this is not immediately apparent, it is because we have taken certain liberties in our framing of the problem. Strictly speaking, we ought to have included in the space of possibilities not only the possible contents of the left and right-hand envelopes, but also the two possible outcomes of the agent’s choice of envelope. In other words, we ought to have let  $\Omega = \{(l, r, c) : l, r \in \mathbb{Z}, |l - r| = 1; c \in \{\text{left}, \text{right}\}\}$ , where the first term in each ordered triple gives the number of the

asserts that whatever the number of the ticket in the agent’s envelope may be, this ticket is as likely as not to be the winner. That is, for all  $n$ :

$$Pr(R = n + 1|L = n) = Pr(R = n - 1|L = n),$$

or, equivalently:

$$(2) \quad Pr(n, n + 1) = Pr(n, n - 1).$$

If we let  $Pr_W(n)$  be the probability that the number of the winning ticket is  $n$ , then it follows from (1) and (2) that, for all  $n$ :

$$\begin{aligned} Pr_W(n) &= Pr(n, n - 1) + Pr(n - 1, n) \\ &= 2Pr(n, n - 1) \\ &= 2Pr(n, n + 1) \\ &= Pr(n, n + 1) + Pr(n + 1, n) = Pr_W(n + 1) \end{aligned}$$

In other words, the function  $Pr$  assigns an equal weight to each of the possible values of the number of the winning ticket. But, since there are an infinite number of such possibilities, the sum of these weights must equal either zero or  $+\infty$ .<sup>4</sup> But this contradicts the original assumption that  $Pr$  is a probability function since, by definition, the sum of the probabilities assigned to a set of mutually exclusive and exhaustive events is 1.

This argument shows that if we assume that the agent’s prior state of belief can be represented by a probability function on  $\Omega$ , then the agent cannot coherently accept (P1), and based on this fact, some authors have sought to dismiss the simple form of the paradox as resulting from a straightforward inconsistency on the part of the agent.<sup>5</sup> Thus, for example Scott and Scott (1997) claim that “[t]he simple form of the paradox can be rejected . . . since the argument itself makes an assumption which entails a contradiction.”<sup>6</sup>

To describe the situation this way, however, is misleading, for the contradiction only arises once we accept the assumption that the agent’s prior state of belief ought to be represented by a probability function, and, in this case, there is reason to doubt the legitimacy of this claim. In particular, since the agent is given no information concerning the number of the winning ticket, it *does* seem reasonable that he ought to regard each of the possible values of this quantity in a like manner,<sup>7</sup> and, as already noted, there is no probability function that can allow him to do so. Indeed, probabilities defined on infinite sets are not only non-uniform, they are all, in a sense, *radically* non-uniform, since it follows from the

---

ticket in the left hand envelope ( $L$ ) the second term the number of the ticket in the right-hand envelope ( $R$ ), and where the third term determines which of the two envelopes (“left” or “right”) was initially chosen by the agent. In this framework, rather than  $L$  and  $R$ , we would instead be concerned with the random variables  $M$  (“mine”) and  $N$  (“not mine”) defined by:

$$M(l, r, c) = \begin{cases} l & \text{if } c = \text{left} \\ r & \text{if } c = \text{right} \end{cases} \quad N(l, r, c) = \begin{cases} l & \text{if } c = \text{right} \\ r & \text{if } c = \text{left} \end{cases}$$

Thus, for instance, the condition of exchangeability would be expressed  $Pr(M = n|W = n) = Pr(N = n|W = n)$ , which follows immediately from the fact that the agent’s choice of envelope was arbitrary (i.e.,  $Pr(l, r, \text{left}) = Pr(l, r, \text{right})$ , for all  $l, r$ ). For ease of exposition, we have opted to characterize the setup *after* the agent has already chosen the left-hand envelope. Thus, in our model, we cannot represent the fact that the choice between the two envelopes was arbitrary. We simply assume that the agent (and the reader) “remembers” that this was the case.

<sup>4</sup>If  $Pr_W(n) = 0$ , then the sum obviously equals zero. If  $Pr_W(n) > 0$ , then, by the Archimedean property of the real numbers, the sum is infinite. If we wish to satisfy conditions (1) and (2), while, at the same time allowing the total weight assigned by  $Pr$  to be finite, we could assume that  $Pr$  takes values in a non-Archimedean field, like that of the (extended) hyperreals. This possibility is mentioned, though not discussed at any great length in Sobel (1994).

<sup>5</sup>To quote Ian Jewitt: “As I see it, the monstrous hypothesis is the uniform density on the set of . . . integers. Accept this and you accept anything.” (quoted in Nalebuff (1989), p. 176)

<sup>6</sup>Scott and Scott (1997), p. 35.

<sup>7</sup>This is not to suggest that the agent believes that the number of the winning ticket was selected ‘at random’, where this belief is meant to refer to some specific chance mechanism. The point is rather that a uniform distribution may provide the best modeling of the agent’s prior state of knowledge, or what is perhaps more accurate in this case, the agent’s prior state of ignorance.

fact that the probabilities must sum to 1 that for any arbitrarily small  $\varepsilon > 0$ , there will exist a *finite* set with probability greater than  $1 - \varepsilon$ . In other words, the agent can only assign probabilities to the possible contents of the two envelopes, if he is willing to posit a finite set from which it is all but certain that the number of the winning ticket has been drawn<sup>8</sup> (or, as de Finetti put it, if he is willing to treat an infinite set of possibilities as a ‘finite set, up to trifles’<sup>9</sup>), and this despite the fact that he has been given no information concerning the value of this quantity. When viewed in this light, such a demand seems rather unreasonable.

It is therefore misleading to suggest that the acceptance of (P1) involves the agent in a contradiction, for we may just as well describe the situation as one in which the agent’s uniform state of ignorance precludes him from assigning probabilities to the possible contents of the two envelopes.

Once we give up the requirement that the agent must adopt a prior probability function, we are free to seek out some alternative means of justifying the judgments of equiprobability expressed in (P1). If, for example, we wished to do so within a generalized Bayesian framework, we could model the agent’s prior state of belief by a uniform, but *improper* distribution, i.e., a distribution which sums to infinity.<sup>10</sup> Alternatively, we could simply treat the agent’s conditional probabilities as primitive, and assume that the agent reasons in accordance with the following heuristic: When asked to deliberate under the assumption that the number of the winning ticket belongs to some finite set, the agent concludes that it is equally likely that this quantity takes any one of the values in this set; otherwise, he refrains from assigning probabilities altogether. Given the agent’s limited information, this does not seem like an altogether unreasonable policy to adopt.<sup>11</sup>

Fortunately for us, however, there is no need to enter into any more detailed discussion of what justifies the agent in accepting (P1). For, as it turns out, the assumption that the agent is given no information about how the contents of the two envelopes were determined can be done away with in the formulation of the paradox. Indeed, as we shall see, with suitable modifications to the original scenario, the argument for switching can even be modified so as to apply to an agent with *precise knowledge* of the probabilistic mechanism at work in the setup.

The key point to note here is that in the argument for switching envelopes, the agent’s uniform conditional probabilities, as expressed in (P1), are only significant insofar as they entail that the agent’s *conditional expected gain* from switching envelopes is always positive. But the agent’s conditional expectations arise from a combination of his conditional probabilities and utilities, and so it stands to reason that this condition could be satisfied despite the agent’s conditional probabilities being non-uniform provided suitable adjustments are made to the payoff function  $f$ .

---

<sup>8</sup>The agent could perhaps justify the assumption that there must be a specific upper bound on the absolute value of the number of the winning ticket, on the grounds that there is a fundamental physical limit to how large a number can be written in the space contained by the envelope (the particular upper bound will, of course, depend upon the language in which the number is represented, but if, for example, it is assumed that this language has a finite alphabet consisting of  $N$  distinct symbols, then an (extreme) upper bound could be given by  $N^V$ , where  $V$  is the number of Planck volumes that fit within the envelope). However, even if such considerations could justify the agent in placing an upper bound on the magnitude of the number of the ticket, it is absurd to suppose that the agent’s error lies in his failure to take such considerations into account.

<sup>9</sup>de Finetti (1974) p. 122.

<sup>10</sup>The use of improper distributions as a means of representing non-informative priors has a long history in Bayesian statistics dating back, at least, to Laplace, who adopted the equivalent of what is today referred to as the Lebesgue measure as the prior distribution of a location parameter with values ranging over the whole real line. In more recent years, the use of improper priors has become a standard part of the Bayesian methodology, particularly since many formal rules for prior selection, such as Jeffrey’s rule, yield improper distributions when applied in common statistical settings. For a detailed discussion of this point, see Kass and Wasserman (1996). For a Bayesian argument against the use of improper priors, see Jaynes (2003), c. 15.

<sup>11</sup>Such a heuristic relies on a restricted version of the principle of indifference, i.e., the principle which asserts that when one has no reason to judge any one of a number of possible alternatives more likely than any other, one ought to assign to each an equal weight. The principle of indifference is, of course, a very controversial principle, and it is a well-known fact that, in its unqualified form, its application can issue in paradoxical results. For the classic discussion of the paradoxes of indifference, see Keynes (1920), ch. 4. For a more recent (and less optimistic) analysis of the paradoxes, see van Fraassen (1989), ch. 12. For a defense of the view that the simple form of the Exchange Paradox is yet another in the long line of paradoxes of indifference, see Priest and Restall (2008).

To see that this is indeed the case, let  $\rho(n)$  be the agent's conditional probability<sup>12</sup> that the right-hand envelope contains ticket  $n + 1$ , given that his own envelope contains ticket number  $n$ . Then, the agent's conditional expected gain from switching envelopes, given that the number of his ticket is  $n$ , is:

$$\rho(n)f(n + 1) - (1 - \rho(n))f(n)$$

This is positive whenever:

$$(3) \quad \rho(n) > \frac{f(n)}{f(n) + f(n + 1)}$$

If this condition is satisfied for all values of  $n$ , then the argument for switching envelopes will go through, since whatever the agent assumes to be the number of the ticket in his envelope, he will expect a positive return from the exchange. If this is the case, we can replace (P1) in the argument for switching with:

(P1)\* For all  $n$ , assuming that  $n$  is the number of the ticket in the agent's envelope, the probability that the other envelope contains ticket  $n + 1$  is  $\rho(n)$ .

If condition (3) is satisfied for all  $n$ , we will refer to the pair of functions  $(\rho, f)$  as 'paradoxical'.

Now, in the simple version of the paradox, it is assumed that  $\rho(n) = 1/2$  for all  $n$ , so that the pair  $(\rho, f)$  is paradoxical just in case  $f$  is a strictly increasing function. However, it is not difficult to see that so long as the function  $\rho$  has a positive lower bound (i.e., for some  $\varepsilon > 0$ ,  $\rho(n) \geq \varepsilon$  for all  $n$ ), there will exist *some* function  $f$ , such that the pair  $(\rho, f)$  is paradoxical (in particular, this will be true for the function  $f(n) = (1/\varepsilon)^n$ ). It follows that there exist many formulations of the paradox in which the number of the agent's ticket may, in fact, provide him with non-trivial information as to whether or not it is the winner. In such cases, though the agent will sometimes be led to conclude that it is more likely than not that his envelope contains the winning ticket, the increased rewards that the agent would acquire by winning with the other ticket are sufficiently great to make it worth his while to agree to the exchange.

We may now ask whether there exist any such alternative formulations of the paradox that are consistent with the agent's adopting an exchangeable prior probability (as we observed earlier, this is not possible in the simple version of the paradox, where the function  $\rho$  has the constant value  $1/2$ ). Let us refer to the function  $\rho$  as a 'proper' function<sup>13</sup> if for some exchangeable probability  $Pr$  on  $\Omega$  :

$$\rho(n) = Pr(R = n + 1 | L = n),$$

for all  $n$ . Since  $Pr$  is exchangeable, this is equivalent to the condition that:

$$(4) \quad \rho(n) = \frac{Pr_W(n + 1)}{Pr_W(n) + Pr_W(n + 1)},$$

for some probability function  $Pr_W$  on the set of integers. The above question can now be restated as follows: Can one identify functions  $\rho$  and  $f$ , such that  $\rho$  is proper and  $(\rho, f)$  is paradoxical? The following, simple example suffices to show that the answer to this question is yes.<sup>14</sup>

Suppose that the number of the winning ticket is determined by means of the following process: We first draw a ball at random from an urn containing three balls, marked '0', '+' and '-', respectively. The mark on the drawn ball is used to determine whether the number of the winning ticket is to be zero, positive or negative. In either of the latter two cases, the absolute value of the winning ticket is determined by tossing a fair coin repeatedly until the coin lands 'tails', and counting the number of

<sup>12</sup>From this point on, we treat the agent's conditional probabilities as primitive, and, in particular, we do not assume that they express the conditionalized form of a prior probability. Thus, we make no assumptions about the function  $\rho$ , except that it takes values between 0 and 1.

<sup>13</sup>See fn. 10.

<sup>14</sup>Examples of proper paradoxical pairs  $(p, f)$  can be found in Nalebuff (1989), Broome (1995), and Clark and Shackel (2000). In all of those works, it is assumed that the numbers of the tickets are bounded below (though not above). The example we provide shows that the boundedness assumption is not necessary for paradoxicality.

times the coin was tossed. Having thus determined the number of the winning ticket, we again toss a fair coin to determine in which of the two envelopes the winning ticket is to be placed (it is this last step which ensures that the probability is exchangeable).

The probability distribution that characterizes this process is given by

$$(5) \quad Pr_W(n) = \frac{1}{3} \left(\frac{1}{2}\right)^{|n|}$$

If, with respect to this distribution, we now define the function  $\rho$  as in equation (4), then  $\rho$  is, by definition, proper. All that remains to be shown, then, is that there exists a function  $f$ , such that  $(\rho, f)$  is paradoxical, and for this it suffices to show that  $\rho$  has a positive lower bound. This can be verified directly by observing that:

$$\rho(n) = \begin{cases} \frac{2}{3} & n < 0 \\ \frac{1}{3} & n \geq 0 \end{cases}$$

Thus, provided the payoffs for the game are suitably chosen (e.g., if  $f(n) = 3^n$ ), an agent who adopts the conditional probabilities given by  $\rho$ , can follow through with the argument for switching. That is, even if such an agent is made *fully aware* of the chance process by means of which the number of the winning ticket was determined, it will still be the case that for any given ticket that might be in the agent's envelope, if the agent assumes that this ticket *is* in his envelope, he will conclude that he is better off switching.

Most of the recent literature on the subject has focused on these 'proper' forms of the paradox, since in these cases, the agent's conditional probabilities can coherently be thought of as expressing the agent's knowledge of the chance mechanism at work in the setup. In what follows, we adopt a slightly different tack. Based on the existence of proper forms of the paradox, we conclude that the agent's conditional probabilities are not where the fault lies with the agent's reasoning, and we extend this conclusion even to *improper* forms of the paradox. In doing so, we do not mean to imply that, in the simple form of the paradox, the agent's conditional probabilities are immune to criticism, only that such criticism should play no role in a satisfactory account of what is wrong with the argument for switching. From this point on, then, we will take it for granted that the agent is justified in accepting (P1)\* (whether or not the function  $\rho$  is proper).

## 1.4 Unconditional Expectations and the Sure-Thing Principle

Let us now turn to the second of the two standard responses to the paradox. This response locates the fallacy in the final step of the argument, in which the agent concludes on the basis of (P2) that he ought to agree to the exchange. What justifies the agent in making this inference?

Recall that, in the context of the argument for switching, (P2) follows from the fact that whatever the agent may assume to be the number of the ticket in his envelope, he will expect a positive return from the exchange. But, according to standard Bayesian decision theory, the agent is only justified in accepting the exchange provided he expects a positive return from switching *without assuming anything* about the number of his ticket. These two claims are not equivalent. While the first asserts that the agent's *conditional* expected gain from switching envelopes (given the number of his ticket) is always positive, the second asserts that the agent's *unconditional* expected gain from switching envelopes is positive. According to the standard Bayesian account, then, if the final step in the argument is to be justified, the latter claim must follow from the former, but does it?

The first and most obvious point to note is that if the agent's prior state of belief cannot be represented by a probability function, then the agent's unconditional expectations are not well-defined. In the previous section, we observed that in the simple version of the paradox, the agent can only avoid the charge of incoherency by refusing to assign probabilities to the possible contents of the two envelopes. Without

such probabilities, however, the agent cannot meaningfully assess the gains he expects to acquire from the exchange.<sup>15</sup> Thus, in the simple form of the paradox, standard expected value reasoning cannot be used to underwrite the agent’s decision to switch.

But what if we restrict our attention to *proper* forms of the paradox, in which the agent *can* coherently adopt a prior probability function? In these cases, does it follow from (P2) that the agent’s unconditional expected gain from switching envelopes is positive? Again, the answer is no, but this time for reasons of a slightly more technical nature.<sup>16</sup>

Suppose that  $Pr$  is the agent’s prior probability function. Then, the agent’s unconditional expected gain from switching envelopes is:

$$(6) \quad \sum_{\omega \in \Omega} Pr(\omega)G(\omega),$$

where  $G$  is the net gain or loss that the agent would incur by switching from the left to the right-hand envelope, i.e.:

$$G(\omega) = \begin{cases} f(R(\omega)) & \text{if } R(\omega) > L(\omega) \\ -f(L(\omega)) & \text{if } R(\omega) < L(\omega) \end{cases}$$

In order to assess an infinite sum such as this, we must first arrange the terms in an ordered sequence, then compute the sum of the first  $m$  terms (as they appear in this sequence) and take the limit of these ‘partial’ sums as  $m$  goes to infinity. It follows that when we express an infinite sum as above, without making reference to any particular ordering of its terms, we are tacitly assuming that the value of this sum does *not* depend upon the order in which its terms enter into the assessment. In other words, if  $\omega_1, \omega_2, \dots$  and  $\omega'_1, \omega'_2, \dots$  are any two orderings of  $\Omega$ , then in order for the above sum to be well-defined it must be the case that:

$$(7) \quad \sum_{k=1}^{\infty} Pr(\omega_k)G(\omega_k) = \sum_{k=1}^{\infty} Pr(\omega'_k)G(\omega'_k),$$

Now, if the sum were finite, this would, of course, follow immediately from the fact that addition is a commutative operation, but in the case of an infinite series, it turns out that this condition does not always hold. To see why, we first note that, roughly speaking, there are two distinct ways for the partial sums of an infinite series to converge: On the one hand, it could be that the magnitudes of the terms eventually grow to be very small, and that their approach to zero proceeds sufficiently quickly to limit the growth of the series’ partial sums.<sup>17</sup> When a series converges in this way, it is said to converge ‘absolutely’,<sup>18</sup> and it can be shown that the sum, in this case, will not depend upon the order of the terms.<sup>19</sup>

On the other hand, it is possible for an infinite series to converge not because its terms (eventually) grow very small very quickly, but rather because negative terms occur periodically in the series and act to ‘cancel out’ the previously accumulated sums. Consider, as an analogy, the alternating series:

$$1 + (-1) + 1 + (-1) + \dots$$

<sup>15</sup>In fn. 10, we noted that in the simple form of the paradox, the agent’s uniform conditional probabilities can be represented within a generalized Bayesian framework by attributing to the agent an improper prior distribution. It would be wrong, however, to use this distribution to assess the agent’s unconditional expectations, as the use of improper distributions for the purpose of computing averages leads to a great many paradoxical results, most notably the marginalization paradoxes first discussed in Dawid et al. (1973).

<sup>16</sup>If the reader wishes to circumvent this technical excursion, he or she may skip ahead to the paragraph which begins “The upshot of all this...”

<sup>17</sup>It is not enough for the terms of the series to approach zero in the limit, to ensure that the series converges. For example, while the terms of the  $1 + \frac{1}{2} + \frac{1}{3} + \dots$ , go to zero, they do not do so sufficiently quickly to prevent the series from diverging to  $+\infty$ .

<sup>18</sup>Formally, the series  $\sum_{k=1}^{\infty} a_k$  converges absolutely, just in case the series  $\sum_{k=1}^{\infty} |a_k|$  converges.

<sup>19</sup>Roughly put, this is because the convergence behavior of an absolutely convergent series is determined by properties of its ‘tail’, and, in general, tail-properties of a sequence are invariant under permutations.

What constrains the growth of the partial sums of this series is not any decrease in the magnitude of its terms, but rather the fact that every other term in the series cancels out the term that immediately precedes it. In this case, if we wish to allow the series to grow more quickly, we can simply rearrange its terms so that these cancellations take place less frequently:<sup>20</sup>

$$1 + 1 + (-1) + 1 + 1 + (-1) + \dots$$

While neither of these two series converges, only the latter diverges to  $+\infty$ .

If an infinite series converges, but not absolutely, then its convergence is owing to this sort of effect, where the negative terms in the series cancel out the positive terms. As a result, the sum of such a series will depend upon the order of its terms, since by rearranging these terms one can alter the relative frequency with which negative terms appear in the series. In fact, it is a well known theorem, owing to Riemann, that if an infinite series converges but not absolutely, then by rearranging its terms, its sum can be made to take *any* value, either finite or infinite.<sup>21</sup>

Now, in the context of the paradox, it is easy to see that the infinite sum representing the agent's expected gain can be ordered so as to yield an alternating series, and it turns out that if an agent's prior probabilities, in conjunction with the payoff function  $f$ , give rise to a proper form of the paradox, then this series does *not* converge absolutely.<sup>22</sup>

Thus, in any instance in which a proper paradox arises, equation (7) does not hold. This means that either the agent's expected gain does not converge regardless of how we order the states in  $\Omega$ ,<sup>23</sup> or else for any number – either positive or negative, finite or infinite – we can order the states so as to ensure that the agent's expected gain is equal to this number. In the latter case, there is no reason to accept that the sum assessed with respect to any particular ordering of  $\Omega$  corresponds to the agent's 'true' unconditional expected gain, since the framework of decision theory attributes no significance at all to any ordering of the space of relevant contingencies. Thus, even in the context of a proper paradox, it is wrong to assert that the agent's unconditional expected gain from switching envelopes is positive – rather, it is undefined.<sup>24</sup>

The upshot of all this is that with respect to both improper *and* proper forms of the paradox, it cannot strictly be said of the agent that he expects a positive return from the exchange, and so, according

<sup>20</sup>The possibility of dispersing the negative terms in this way is what differentiates the infinite from the finite case. In the finite case, we can only disperse the negative terms in one part of the series by clustering them together somewhere else, so that the overall effect of the dispersion is eventually cancelled out.

<sup>21</sup>The last two paragraphs are merely meant to give the reader an intuitive sense of how it can be that the sum of a convergent series might depend on the order of its terms. For a mathematically rigorous discussion of the subject, the reader is referred to any standard textbook in analysis, e.g., Rudin (1976, ch 3.).

<sup>22</sup>Suppose, for contradiction, that the series  $\sum_{k=1}^{\infty} Pr(\omega_k)G(\omega_k)$  converges absolutely. Then:

$$\begin{aligned} \sum_{k=1}^{\infty} |Pr(\omega_k)G(\omega_k)| &= \sum_{n \in \mathbb{Z}} \{Pr(n, n-1)|G(n, n-1)| + Pr(n-1, n)|G(n-1, n)|\} \\ &= \sum_{n \in \mathbb{Z}} \left\{ \left( \frac{1}{2} Pr_W(n) f(n) \right) + \left( \frac{1}{2} Pr_W(n) f(n) \right) \right\} \\ &= \sum_{n \in \mathbb{Z}} Pr_W(n) f(n) \end{aligned}$$

is a convergent series. This implies that the sequence of terms in the series must converge to zero, but from (3) and (4) we have that  $Pr_W(n+1)f(n+1) > Pr_W(n)f(n)$ , for all  $n$ . Thus, regardless of how the terms in this series are ordered, one can find a subsequence of this sequence which is strictly increasing. But since all the terms in the sequence are positive, this subsequence (and thus the sequence as a whole) cannot converge to zero. Contradiction.

<sup>23</sup>Moreover, the agent's expected gain cannot always diverge to either  $\pm\infty$ , for it follows from the fact that  $Pr$  is exchangeable, that we can choose  $\omega_1, \omega_2, \dots$  such that every even term in the series  $\sum_{k=1}^{\infty} Pr(\omega_k)G(\omega_k)$  is the negative of the term which precedes it. But this means that the series cannot approach either  $\pm\infty$  since there is a subsequence of the sequence of its partial sums which converges to 0.

<sup>24</sup>Clark and Shackel (2000) show that there exist proper paradoxes for which the agent's average conditional expected gain has a finite positive value. The authors label such cases as 'best' paradoxes, implying that they are somehow more problematic than those paradoxical cases in which this average diverges to infinity. I agree with the point made in Meacham and Weisberg (2003) that no new difficulties are raised by these 'best' versions of the paradox, since the proposed error in the agent's reasoning is not with the assumption that the weighted average of the agent's conditional expected gains is positive, but rather with the assumption that the agent's unconditional expected gain can be identified with this average.

to standard Bayesian decision theory, despite the fact that the agent's conditional expected gains are always positive, the agent has no reason to switch. This point is made in Wagner (1999) as follows:

The lesson to be learned here is to keep steadfastly in mind that it is an inequality between unconditional expectations that is our criterion for preferring one act to another. While a certain family of inequalities involving conditional expectations can often serve as a surrogate for that criterion, this is not universally the case.<sup>25</sup>

And, again, in Norton (1998):

[The agent] holds his envelope in his hand and thinks: "For any definite amount that may be in this envelope, I have an expectation of gain if I swap. Therefore, no matter what is in the envelope, I have an expectation of gain if I swap." The fallacy is in the final step . . . There is no expectation.<sup>26</sup>

I believe that these responses to the paradox misrepresent the agent's rationale for switching envelopes. In particular, the agent's decision to switch need not rely upon any numerical estimate of his unconditional expected gain, but can rather be based upon the much more primitive maxim referred to in the decision-theoretic literature as the 'sure-thing' principle:

**Sure-Thing Principle:** Let  $E_1, E_2, \dots$  be mutually exclusive and exhaustive events, and let  $A$  and  $B$  be two acts which are available to the agent. If, for all  $k$ , the agent knows that he ought to prefer  $A$  to  $B$ , assuming that  $E_k$  has occurred, then the agent ought to prefer  $A$  to  $B$ .

In the scenario under consideration, the sure-thing principle has application since, presumably, the agent is aware of the fact that no matter what he assumes to be the number of the ticket in his envelope, his expected gain from switching envelopes is positive. Hence, according to the sure-thing principle, the agent ought to agree to the exchange.

Now, it is important to stress that the sure-thing principle is not a *consequence* of the principle of maximum expected utility. Indeed, on the contrary, in the standard axiomatic formulations of Bayesian decision theory, the sure-thing principle itself figures among the theory's most elementary axioms. Thus, from a foundational point of view, its justification is independent of the more substantial assumptions needed for the representation of the decision-maker as a 'full-fledged' Bayesian agent with precise probabilities and utilities.<sup>27</sup> Consequently, the sure-thing principle could provide the agent with a reason to switch despite the fact that, in paradoxical settings, the agent cannot meaningfully assess what he expects to acquire from the exchange.

If we understand the sure-thing principle in this way – as a self-standing principle which imposes direct constraints on an agent's rational preferences and not as a corollary of the principle of maximum expected utility – is there any reason to think that the principle cannot be applied in the context of the Exchange Paradox?

In order to address this question, it will be helpful to offer a more concrete interpretation of the principle, and, more specifically, of the agent's conditional preferences. To this end, let us modify the original scenario slightly by supposing that the agent is told prior to choosing his envelope that one minute after making his initial selection, he will be allowed to *open* his envelope and observe the number of the ticket

---

<sup>25</sup>Wagner (1999), p. 239.

<sup>26</sup>Norton (1998), p. 44.

<sup>27</sup>In Savage's framework, for example, the finite form of the sure-thing principle follows directly from the two most elementary axioms of the theory, *viz.*, axiom *P1*, requiring that the agent's preference relation be a total ordering on the set of available acts, and axiom *P2*, requiring that the agent's conditional preferences be well-defined. From a foundational point of view, then, the sure-thing principle can be justified without assuming that the agent assigns precise probabilities to the relevant states, or even that the agent (partially) orders the states in terms of their relative likelihood (see Savage, 1954, ch. 1-5).

inside. After selecting on a whim the envelope in his left-hand, the agent is given the opportunity to exchange envelopes before a minute has passed, and the paradox consists in the fact that he seems to have reason to accept the exchange immediately upon its being offered.

We must take care, right away, to ensure that by modifying the scenario in this way, we have not altered the problem in some crucial respect. After all, it can sometimes be the case that one's merely being informed that at some later point in time one will come to know the value of a certain unknown quantity, can *itself* provide one with a reason to act.<sup>28</sup>

To take a simple example, suppose that you are deciding whether or not to purchase a certain mystery novel when you are informed that your talkative and rather indiscreet friend, whom you know to have just finished reading the novel, will be paying you a visit this evening. Anticipating that by the end of the visit, you will know how the mystery is resolved, you decide not to purchase the book.

While this may be a perfectly reasonable thing to do, this is only because the enjoyment that you will derive from reading the novel depends not just on *how* the book ends, but also on whether or not you *know* how the book ends prior to reading it. In general, if an agent is merely informed of the fact that he will soon come to know the value of an unknown quantity, this can only provide him with a reason to act if his actions have consequences whose worth to him depends not only on the value of this quantity, but also on whether or not he *knows* this value.<sup>29</sup> In the context of the Exchange Paradox, this is clearly not the case, since, by assumption, all the agent cares about is money, and his epistemic state has no direct monetary implications. Thus, if the agent has good reason to switch having been told that he will shortly be allowed to open his envelope, then he must have good reason to switch even *without* being given this information.

From this point on then we will focus on the modified version of the scenario, in which the agent is informed that he will soon be allowed to open his envelope and observe the number of the ticket inside. In this setting, we will interpret the claim that 'the agent ought to switch envelopes, assuming that the number of the ticket in his envelope is  $n$ ', to mean that the agent ought to switch envelopes, if, after one minute, he opens his envelope and observes that  $n$  is the number of his ticket. In other words, we will replace (P2) in the argument for switching with:

(P2)\* For all  $n$ , if the agent opens his envelope one minute from now and observes ticket number  $n$ , he ought to switch envelopes.

On this reading, the sure-thing principle asserts that if the agent knows that one minute from now, after opening his envelope, he will have good reason to agree to the exchange, then he has good reason to agree to the exchange *now*.

Interpreted in this way, the sure-thing principle seems undeniable, for given that neither the actions available to the agent nor the consequences of his actions are of a time sensitive nature, what possible reason could the agent have to refuse the exchange knowing full well that one minute from now he will have good reason to accept it? Such a refusal on the part of the agent cannot be justified on the grounds that he knows that he will soon be receiving new information. For if, on the one hand, the agent believes that this information will somehow be misleading, then he should deny the antecedent assumption that

---

<sup>28</sup>The 'merely' here is intended to rule out those cases in which being told that you will come to know the value of a certain quantity provides you with non-trivial information as to that quantity's value. For instance, if you are the defendant in a murder trial in which the jury deliberates for only a very short time before submitting its verdict to the judge, the fact that you will be informed of the jury's decision one minute after the closing arguments have been delivered, may provide you with information as to what the verdict will be (which in turn may give you reason to flee the courtroom).

<sup>29</sup>It was pointed out to me by Sidney Felder that the disagreement that arises in the context of Newcomb's Problem can be viewed as a disagreement over whether or not the amount of money in the opaque box *depends upon* how much money the agent believes to be in the box. The two-boxer contends that these two quantities are independent, and thus is free to base his decision upon the sort of sure-thing reasoning outlined above. The one-boxer, on the other hand, argues that the amount of money in the opaque box depends upon the agent's beliefs (in so far as these beliefs have a role to play in the predictor's methodology for assessing which of the two options the agent will choose) and thus rejects the two-boxer's argument from dominance.

after opening his envelope, he ought to agree to the exchange.<sup>30</sup> If, on the other hand, he believes that this information will *not* be misleading, then he ought to accept his future decision to switch since, by his own lights, he will, at that time, occupy an improved epistemic state.

Thus, it seems, if the agent refuses to agree to the exchange now despite knowing that he will agree to the exchange a minute from now, it can only be because the agent believes that his rational commitments are subject to change with the mere passage of time, and this is surely an unacceptable consequence for any theory of rationality.<sup>31</sup> Indeed, if the agent were to persist in this conceit, it is tempting to accuse him of engaging in some sort of odd *ritualism*: he is unwilling to exchange until he has opened his envelope and looked at the number inside, but he knows that the particular number he sees will make no difference, since he knows that *regardless* of the number he sees, he will be willing to agree to the exchange. Thus, it seems, the agent is insisting that rationality demands that he be allowed to take part in what he knows to be the purely ritual act of his opening his envelope before he agrees to the exchange. But this is absurd! Moreover, the absurdity of this result does not depend at all on whether the number of the ticket in the agent's envelope can take only a finite or an infinite number of possible values, for in either case, provided the agent possesses a minimal deductive capacity, he is in a position to judge that his expected gain (if calculated in the manner suggested by the argument for switching) will be positive no matter what the number of his ticket may be (if, for some reason, he is unaware of this fact, we can always show him the proof).

At this point, one could simply insist that the agent's rejection of sure-thing reasoning ought to be accepted as yet another strange consequence of decision-making in infinite settings. But if we assume that reasoning in infinite settings opens up the possibility of such radical deviations from normality, then we may as well call into question the claim that, prior to opening his envelope, the agent has no reason to switch. I admit that if one wishes to insist that reasoning in infinite settings is so strange, so far-removed from our ordinary intuitions, that paradoxes cannot even arise, there is nothing to prevent him from doing so. But, otherwise, if one regards the Exchange Paradox as a genuine paradox, then one cannot simply disregard so elementary a principle of reasoning as the sure-thing principle on the grounds that we cannot expect intuitive results where infinity is concerned.

We will henceforth take it for granted, then, that the sure-thing principle does apply in the context of the Exchange Paradox, and thus that the agent is justified in inferring from (P2)\* that he ought to switch envelopes. While a complete defense of the infinitary version of this principle would require a more thorough analysis, hopefully, the above remarks suffice to shift the burden of proof back to those who would claim that it is a mistake for the agent to appeal to the principle in this context.

Let us pause and take stock of what we have said thus far. Over the course of the last two sections, we have recast the argument for switching envelopes in the following terms:

- (P1)\* For all  $n$ , assuming that  $n$  is the number of the ticket in the agent's envelope, the probability that the other envelope contains ticket  $n + 1$  is  $\rho(n)$ ; hence,
- (P2)\* For all  $n$ , if the agent opens his envelope one minute from now and observes ticket number  $n$ , he ought to switch envelopes; hence,

---

The agent ought to switch envelopes.

---

<sup>30</sup>Throughout this discussion, it is important to keep in mind that when we speak of the agent's rational preferences, we are not concerned to address the psychological, or otherwise causal determinants of the agent's intentional actions. Instead, we are interested in the agent's 'reasons' for acting, understood in the *normative* sense of the word. We could make this normative interpretation explicit by always speaking of what the agent ought to prefer, instead of what the agent does prefer, but to do so would make our language more cumbersome than it needs to be. Instead, we will trust that the present remark will suffice to prevent any confusion on this front.

<sup>31</sup>There are, of course, well-known philosophers who have expressed skepticism about the possibility of diachronic constraints on an agent's preferences (see, e.g., Levi (1987)). To provide a satisfactory reply to their objections, however, would take us too far afield. Instead we refer the reader to Gaifman and Vasudevan (2010), sec. 2, where the issue is discussed in greater detail.

We have argued that the fallacy in this argument neither consists in the agent’s acceptance of (P1)\*, nor in the agent’s concluding on the basis of (P2)\*, that he ought to agree to the exchange. The only alternative left, then, is to conclude that the fallacy consists in the inference from (P1)\* to (P2)\*. To reject the validity of this inference is to admit that *even after opening his envelope and observing ticket number  $n$ , and despite having correctly assessed how likely it is that the other envelope contains ticket  $n + 1$ , it is not always the case that the agent ought to agree to the exchange.*

This may seem like a very strange result, for once the agent has opened his envelope, his decision to switch is based on what appears to be a straightforward assessment of his expected gain. In the following sections, we will try to explain why the agent cannot reason in this way. In doing so, we will appeal to the *symmetry* inherent in the problem.

## 2 Methodological Symmetry

### 2.1 Methodological Symmetry and Bias

When a paradoxical argument is first encountered, in spite of one’s having a very clear intuition that the argument must be somehow mistaken, it is often the case that the reasons which underlie this intuition are vague and imprecise. In such cases, in addition to identifying the fallacy in the paradoxical argument, it is often worthwhile to articulate, in detail, the counterparadoxical intuition to show why it implies that the argument must be fallacious.

In the case of the Exchange Paradox, it is clearly the *symmetry* of the situation that underlies the initial intuition that the agent has no reason to switch. Intuitively, the agent ought to recognize that the decision facing him would have been essentially the same had he initially opted for the envelope in his right hand, and, as a result, whatever reasons he might have for agreeing to the exchange are, in effect, ‘cancelled out’ by the corresponding reasons that he would have had for switching had he initially chosen the right-hand envelope.

Let us first try to formulate, in more abstract terms, the general structure of this sort of reasoning. The problem that the agent is faced with is whether or not to switch from the left to the right-hand envelope. In attempting to solve the problem, the agent initially observes that there is a *second* problem that is in all relevant respects identical to that which he is concerned to address (namely, the problem of whether or not he ought to have switched from the right to the left-hand envelope, had he chosen the right-hand envelope instead). Based on this fact, he concludes, on the one hand, that both problems must have the same solution. On the other hand, these solutions, in so far as they are answers to different questions, correspond to different claims, and these claims stand in a certain objective relationship to one another. In particular, since only one of the two envelopes contains the winning ticket, if the agent ought to switch from the left to the right hand envelope, then he ought not to have switched from the right to the left-hand envelope, and vice-versa. These two facts together suffice to determine the solution to the problem.

Note that this sort of reasoning requires that we conceive of the solution to a problem in *two* distinct ways. On the one hand, the solution is to be viewed abstractly, as the formal output of a certain process of reasoning. It is in this sense that we can meaningfully speak of two distinct problems as having the ‘same’ solution. On the other hand, the solution is to be viewed as the answer to a specific question, and, as such, as making a contentful claim about what is or what ought to be the case. It is in this sense that the solutions to two problems can stand in an objective relationship to one another. This dual conception of the solution to a problem (in terms of its structural form and its particular content) is characteristic of all symmetry-based reasoning.

In order to express this reasoning in slightly more formal terms, let us first suppose that the possible forms of the solution to a problem comprise the set  $\Theta$ . We may think of a ‘problem’ as a function mapping each  $\theta \in \Theta$  to a proposition expressing what would be the case were  $\theta$  to be the problem’s

solution. To judge that two problems are *symmetrical* is to judge that these problems must have the same formal solution. In other words, two problems  $\varphi_1$  and  $\varphi_2$  are symmetrical just in case:<sup>32</sup>

$$(8) \quad \exists \theta \in \Theta(\varphi_1(\theta) \wedge \varphi_2(\theta))$$

Of course, the mere recognition that two problems have the same formal solution does not, by itself, impose any constraints on what this solution might be. For this, we need independent knowledge which relates the two problems to one another. Such knowledge will typically be of the form:

$$(9) \quad \forall \theta_1, \theta_2 \in \Theta((\varphi_1(\theta_1) \wedge \varphi_2(\theta_2)) \rightarrow \psi(\theta_1, \theta_2)),$$

where  $\psi$  is some relation on  $\Theta$ . Knowledge of this claim allows us to conclude on the basis of the symmetry of  $\varphi_1$  and  $\varphi_2$ , that their shared solution  $\theta$  must satisfy the condition  $\psi(\theta, \theta)$ , and depending on the relation  $\psi$ , this is sometimes sufficient to determine  $\theta$  uniquely.

While formalizing the reasoning in this way is, itself, a rather trivial exercise, it does serve to highlight the three essential components of any symmetry-based argument, namely:

$\Theta$  : A description of what counts as a formal solution to the problems under consideration;

$\varphi_1, \varphi_2$ : Two problems, each viewed as a means of interpreting these formal solutions; and

$\psi$  : A constraint clause, describing how the solutions to these two problems are related.

Provided these three components are chosen in such a way that claims (10) and (9) are satisfied, the symmetry of the situation will impose constraints on the problem's solution.

Let us now return to the Exchange Paradox, and consider how the intuitive argument described at the beginning of this section can be formalized within this framework. The most obvious sense in which the decision facing the agent 'would have been essentially the same' had he chosen the right-hand envelope, is that he would have had available to him *exactly the same information* concerning the contents of the two envelopes as that which is available to him now. But, since this is the only information relevant to the agent's decision as to whether or not to switch, it follows that if, after choosing the left-hand envelope, the agent concludes that he has good reason to switch, he ought likewise to conclude that he *would* have had good reason to switch had he initially opted for the envelope in his right hand. This, however, is absurd, since the agent knows that only one of the two envelopes contains the winning ticket, and so any reason he might have for switching from the left to the right-hand envelope, should count as an equally compelling reason *not* to switch from the right to the left.

We can formalize this line of reasoning in the above framework as follows: The agent's decision problem can be represented as a choice between the following three options: (1) accept the exchange, (2) refuse the exchange, or (3) be indifferent to the exchange (we will represent these three options by the values 1, -1 and 0, respectively, so that  $\Theta = \{1, 0, -1\}$ ).<sup>33</sup> The fact that the agent's evidence is symmetrical entails that the solution to this problem should not depend on the agent's initial choice of envelope, that is, it should not depend on whether the exchange in question consists of a switch from the left to the right-hand envelope, or vice-versa. Hence, the two problems:

<sup>32</sup>By expressing the condition of symmetry as an existential rather than a universal claim, we are building into the judgment of symmetry the tacit assumption that the problems under consideration both have solutions in  $\Theta$ . In other words, we are assuming that the problems are "well-formed". In the case of the Airplane Seat Problem this is obvious since the chance mechanism is fully described in the problem statement. However, when symmetries are appealed to for the purposes of assigning *inductive* probabilities to events, such judgments of well-formedness constitute empirical assumptions. It is for this reason that prior probabilities that are chosen on the basis of symmetry considerations are subject to revision in the light of new evidence. A failure to recognize this fact has led, in the past, to misplaced criticisms that symmetry-based reasoning involves the agent in an illicit brand of a priorism. In the first chapter of my dissertation, I argue that the real power of such reasoning derives not from its alleged a priority, but rather from the fact that our empirical knowledge is organized in such a way that, in practice, we are often able to judge of a given problem that it has a well-defined solution, even without being able to offer any account of what that solution might be, or even how it can be obtained.

<sup>33</sup>Expressed this way, these three options may not appear to be mutually exclusive, since the agent may be indifferent to the exchange and still accept the offer to switch. Obviously, by "accept the exchange" what we have in mind is something more like "conclude that he ought to accept the exchange." But to describe things this way would make our language more cumbersome than it needs to be.

$LR(\theta) =$  “If given the chance to switch from the left to the right-hand envelope, the agent ought to  $\theta$ .”

$RL(\theta) =$  “If given the chance to switch from the right to the left-hand envelope, the agent ought to  $\theta$ .”

must have the same formal solution.

At the same time, since only one of the two envelopes contains the winning ticket, if the agent has good reason to switch from the left to the right-hand envelope, then he has good reason *not* to switch from the right to the left (and vice-versa), i.e.:

$$(LR(\theta_1) \wedge RL(\theta_2)) \rightarrow \theta_1 + \theta_2 = 0$$

Together with the symmetry of the situation, this entails that  $LR(0)$ , i.e., if given the chance to switch from the left to the right-hand envelope, the agent ought to be indifferent to the exchange.

This simple symmetry-based argument is what underlies our initial intuition that the agent has no reason to switch. Unfortunately, while this line of reasoning clearly entails that the argument for switching envelopes must be fallacious, it cannot help to explain where the argument goes wrong. For, with the exception of the sure-thing principle, the argument for switching envelopes only relies upon claims about what the agent ought to do *once he has opened his envelope and observed the number of the ticket inside*. But once he has done this, his epistemic situation is no longer symmetrical in the above sense since he now possesses information which would *not* have been available to him had he chosen the other envelope (e.g., he knows that his envelope contains ticket number  $n$ ).

This points to the real tension at the heart of the Exchange Paradox. On the one hand, the intuition that the agent has no reason to switch is based upon a certain ‘evidential’ symmetry which characterizes his epistemic state at the time of his decision. On the other hand, the agent is aware of the fact that after a minute has passed this symmetry will be *broken*, and, once this has occurred, it is no longer clear that the agent should be indifferent to the exchange. Thus, it seems, the symmetry of the agent’s situation cannot prevent him from concluding that one minute from now, he ought to switch envelopes, and more generally, that he ought to do so no matter what the number of his ticket may be. In the remainder of this section, I will argue that what, in fact, prevents the agent from accepting this last claim is a certain *higher-order* symmetry inherent in the problem, which persists even after the agent has opened his envelope.

Suppose then that the agent has already opened his envelope and observed the number of the ticket inside. As we already noted, the agent now possesses information which would not have been available to him had he chosen the other envelope. Nevertheless, there is still a sense in which the agent’s decision problem is symmetrical, for while it is true that the agent’s decision whether or not to switch, insofar as it depends on the number of the agent’s ticket, may depend on the agent’s initial choice of envelope, clearly, the sort of *reasoning* by means of which the agent arrives at this decision should not. Thus, for example, if after opening the left-hand envelope the agent sees ticket number  $n$ , and on the basis of this information decides that he ought to switch, he should likewise conclude that he ought to have switched had he instead opened the right-hand envelope and observed the same ticket. In other words, the agent’s *general methodology* for deciding whether or not to switch, should not depend on his initial choice of envelope.

What is important to note is that this sort of ‘methodological’ symmetry can be acknowledged by the agent despite his knowing that the two envelopes contain different information. In this respect, the agent’s situation is analogous to that in which one judges two individuals to be ‘equally trustworthy’ with respect to their judgments in a given domain. Even if one is then informed that the opinions of these two individuals differ on some particular point, their disagreement does not necessarily provide a sufficient reason to withdraw one’s original judgment of equal expertise.<sup>34</sup> Similarly, in the agent’s

<sup>34</sup>Although it may give one reason not to base his opinion solely on the testimony of one of the two experts without trying to take into account the opinion of the other (see section 6, below).

case, despite knowing that the two envelopes carry different information, he may still judge them to be ‘equally valid’ sources of information, in the sense that information obtained from the one should enter in the same way in the agent’s subsequent deliberations as information obtained from the other.

In order to represent this symmetry in the framework described above, we must construe the decision problem confronting the agent not as a simple choice between the three options of accepting, rejecting or being indifferent to the exchange, but rather as a *higher-order* decision problem in which the agent is forced to commit to a general deliberative methodology which determines how he is to act once he has opened his envelope and observed the number of the ticket inside (note that to represent the agent’s decision problem this way is not an ad-hoc maneuver since this is exactly the decision problem that the agent must solve if he is to carry through with the argument for switching). In formal terms, such a methodology can be represented as a function which maps the possible numbers of the ticket in the agent’s envelope to values in the set  $\{1, 0, -1\}$ . The higher-order decision problem facing the agent is the problem of selecting such a function (we let  $\Theta$  be the set of all such functions), and, just as above, the symmetry of the situation entails that the solution to this problem should not depend upon the agent’s initial choice of envelope.

Now, in the above argument from symmetry the agent’s choice of envelope was only significant insofar as it determined whether the proposed exchange involved a switch from the left to the right-hand envelope or vice-versa. In the context of this higher-order decision problem, however, the agent’s choice of envelope also determines what *type* of information the agent is to receive (that is, whether he is to be informed of the number of the ticket in the left or the right-hand envelope).

We first define the propositional form:

$LR(E, x) =$  “If the agent knows only that the event  $E$  has occurred, then, if given the chance to switch from the left to the right-hand envelope, the agent ought to  $x$ .”

where  $E$  is any event and  $x \in \{1, 0, -1\}$  ( $RL(E, x)$  is defined similarly).

Now, when the agent opens his envelope and observes ticket number  $n$ , all that he knows is that the event  $L^{-1}(n)$  has occurred. Thus, the claim that, “if the agent chooses the left-hand envelope, he ought to adopt the methodology  $\theta$ ”, can be expressed as follows:

$$Left(\theta) = \forall n \in Z [LR(L^{-1}(n), \theta(n))]$$

Similarly, the claim that, “if the agent chooses the right-hand envelope, he ought to adopt the methodology  $\theta$ ”, can be written:

$$Right(\theta) = \forall n \in Z [RL(R^{-1}(n), \theta(n))],$$

The higher-order symmetry of the situation entails that these two problems must have the same solution.

In order to impose constraints on this solution, we must offer some account of how these two problems are related to each other. Here, we again appeal to the fact that only one of the two envelopes contains the winning ticket, so that, in any given epistemic state, if the agent has good reason to switch from the left to the right-hand envelope, then he has good reason *not* to switch from the right to the left, and vice-versa. In other words, for any event  $E$ :

$$(10) \quad (LR(E, x_1) \wedge RL(E, x_2)) \rightarrow x_1 + x_2 = 0$$

In addition, we assume that the agent’s conditional preferences ought to be preserved under arbitrary disjoint unions:

$$(11) \quad \text{If } E_1, E_2, \dots \text{ are mutually exclusive events, and if } LR(E_k, x) (k = 1, 2, \dots), \text{ then } LR(\bigcup E_k, x).$$

(this, of course, is also assumed to be true for claims of the forms  $RL(E, x)$ ).

From these two constraints and the symmetry of the situation, it follows that:<sup>35</sup>

<sup>35</sup>The same argument can be used to show that if  $Left(\theta)$ , then there exists some  $n$ , such that  $\theta(n) \neq -1$ .

If  $Left(\theta)$ , then there exists some  $n$  such that  $\theta(n) \neq 1$ .

The argument is very simple. Assume, for contradiction that  $Left(\theta)$ , where  $\theta(n) = 1$ , for all  $n$ . By the definition of the problem  $Left$  we have  $LR(L^{-1}(n), 1)$  for all  $n$ , and so by (11),  $LR(\Omega, 1)$ . By symmetry, however, the same reasoning can be applied to the problem  $Right$  to show that  $RL(\Omega, 1)$ . But this, in conjunction with  $LR(\Omega, 1)$ , violates the zero-sum condition.

The most important point to take from this discussion is that there are, in fact, *two* distinct symmetries exhibited by the problem. The first is an obvious evidential symmetry, which reflects the fact that at the time of the agent's initial decision, all the relevant information that the agent possesses concerning the contents of the left-hand envelope, likewise applies to the right. The second is a higher-order, methodological symmetry which is based on the fact that the two envelopes viewed as potential sources of information are equally valid, in the sense that information obtained from the one ought to be handled in the same way as information obtained from the other. What seems so compelling about the argument for switching envelopes is that it appears to provide the agent with a reason to switch which respects the first of these two symmetries. However, as we have seen, it can only do so by violating the second.

We can now explain why, even after opening his envelope, the agent cannot coherently base his decision to switch on the straightforward assessment of his expected gain appealed to in the argument for switching. Since the agent's decision to switch can be conceptualized as a move in a zero-sum game, if, after opening the envelope, he decides to switch, it must be that he believes that the number of the ticket in his envelope provides him with special information. Now, if this information is special, this can either be owing to the fact that the particular number of the ticket in his envelope is somehow significant, or else it can be owing to the fact that the information concerns the contents of the left rather than the right-hand envelope. But since the agent knows that his methodology recommends switching no matter what the number of his ticket may be, he must treat his methodology as distinguishing somehow between the left and the right-hand envelopes, but this violates the (higher-order) symmetry of the situation.

This argument shows that the symmetry of the situation precludes the agent from adopting a general methodology which instructs him to switch no matter what the number of his ticket may be. To put the point contrapositively, the agent cannot coherently appeal to such a methodology as a means of handling information from the left-hand envelope without denying that the same methodology can be applied to information from the right-hand envelope.

The fallacy in the argument for switching envelopes thus consists in the agent's inference from (P1)\* to (P2)\*. The reason why this inference is invalid is that it requires the agent to commit to a methodology which treats the two envelopes in an asymmetrical fashion. One natural term to use to describe such a methodology is 'biased' since the methodology discriminates between information obtained from two equally valid sources.<sup>36</sup> In a certain sense, the conclusion that the argument for switching commits the agent to reasoning in a biased fashion is obvious, for, intuitively, what is wrong with the argument is that the agent could have applied exactly the same sort of reasoning to the *other* envelope to justify the conclusion that he ought to refuse the exchange. Thus, in reasoning as he does, the agent is treating his own envelope as "special". What is not so obvious is that what prevents the agent from doing so is *not* the fact that his evidence is symmetrical, but is rather the higher-order judgment that the two envelopes represent equally valid sources of information.

In spite of all this, it may still seem odd that the deliberative methodology suggested by the argument for switching turns out to be biased, for it appears to be based on a straightforward assessment of the agent's conditional expected gain. One way of emphasizing the bias inherent in this sort of decision-making procedure is to construe the agent's conditional expected gain in frequentist terms, as a measure of the long-run rate of return in a sequence of repeated trials.<sup>37</sup> To this end, let us suppose that the

<sup>36</sup>Thus, for example, if a person is willing to accept the conclusion of a given argument when it is presented to him by one person, but is unwilling to accept the conclusion of the same argument when presented to him by someone else, we would say that the agent's reasoning is biased, since we assume that all that should matter in deciding whether or not to accept the conclusion, is the strength of the argument offered in its support, and not who rendered it.

<sup>37</sup>This way of viewing the problem is discussed in Clark and Shackel (2000). I do not wish to commit myself to the view that probabilities must be interpreted in frequentist terms. I will assert, however, that doing so can often help to sharpen our intuitions, particularly in cases in which standard methodologies are being called into question.

same game were to be repeated a very large number of times. Since the agent's conditional expected gain from switching envelopes (given that  $n$  is the number of his ticket) is positive, if the agent were to agree to switch whenever he observed ticket number  $n$  in his envelope, then the agent's per game winnings in these particular cases would eventually converge to some positive value.

Does this fact suffice to provide the agent with a reason to switch once he has opened his envelope and observed ticket number  $n$ ? Not necessarily. For it is not obvious why the agent ought to assess his strategy according to its long-run benefits in just those repeated trials in which the number of the ticket in his *own* envelope is held fixed. Why not, for example, assess such a strategy according to its long-run benefits in repeated trials in which the number of the ticket in the *other* envelope is held fixed? Even without knowing the actual number of the ticket in the other envelope, the agent knows that in trials in which this number is held fixed, the most effective long-run strategy is to refuse the exchange. Thus, in basing his decision to switch on a standard calculation of expected gain, the agent *is*, in a certain sense, treating his own envelope as special, since he is assessing his expected gain relative to a sequence of repeated trials in which the number of the ticket in his envelope is held fixed.

One important point to note (and one which is often overlooked) is that this same sort of reasoning applies equally well to *finite* versions of the scenario in which the number of the winning ticket has a known upper bound. In such cases, if after opening his envelope and observing the number of the ticket inside, the agent bases his decision to switch on a standard calculation of his expected gain, then he can likewise be criticised for reasoning in a biased fashion. Unlike the infinite case, however, such a criticism cannot be used to underwrite a direct charge of incoherency on the part of the agent since the methodology, in this case, does not recommend switching no matter what the number of the agent's ticket may be.

## 2.2 A Brief Note on Common Knowledge

The symmetry argument presented in the preceding section has a natural formulation in terms of *common knowledge*. To see this, we adopt the model for representing common knowledge first introduced in Aumann (1976). In the context of this model, it is assumed that each agent comes to acquire information about the world by learning the value of certain fixed random variable defined on the set of possible states of the world,  $\Omega$ . If an agent's information takes the form of the value of the variable  $X$ , and if the true state of the world is  $\omega$ , then the agent *knows* that the world is in one of the states in the set  $E_X(\omega) = X^{-1}(X(\omega))$ .<sup>38</sup> We may thus associate with every variable  $X$ , a corresponding knowledge operator  $K_X$  defined by

$$K_X(E) = \{\omega \in \Omega : E_X(\omega) \subset E\}.$$

$K_X(E)$  is the event that an agent who learns the value of  $X$ , knows that the event  $E$  has occurred.<sup>39</sup>

Now consider *two* agents, the first of whom learns the value of  $X$  and the second of whom learns the value of  $Y$ . We say that an event  $E$  is common knowledge between these two agents if its status as knowledge is preserved through any finite alternating sequence of iterations of the knowledge operators  $K_X$  and  $K_Y$  (i.e.,  $K_X(E)$ ,  $K_Y(E)$ ,  $K_X(K_Y(E))$ ,  $K_Y(K_X(E))$ , etc.).

In order to represent the paradox in this framework, we first have to recast the scenario as one in which the agent is involved in a two-player game against an opponent who is in possession of the other envelope. After opening their respective envelopes, the first (left-hand) agent is informed of the the value of  $L$ , and the second (right-hand) agent is informed of the value of  $R$ , and they are then asked whether or not they wish to switch envelopes.

As before, we will assume that both players adopt comprehensive methodologies for playing the game. In other words, for any event  $E$ , it is pre-determined what each of the players will do (switch, stay or

<sup>38</sup>The set of events  $\{E_X(\omega)\}_{\omega \in \Omega}$  is referred to by Aumann as the agent's *information partition*.

<sup>39</sup>It is easy to verify that for any random variable  $X$ ,  $K_X$  is an S5 modality. That is, in addition to the basic axioms of modal logic,  $K_X$  satisfies both positive introspection (if an agent knows  $E$ , then he knows that he knows  $E$ ) and negative introspection (if an agent does not know  $E$ , then he knows that he does not know  $E$ ).

be indifferent) if they were to be informed of the fact that  $E$  has occurred. Moreover, we will assume that rationality demands that their methodologies satisfy conditions (10) and (11) described above.

Note that an alternative way of expressing condition (10) is to say that there must exist some function  $w : 2^\Omega \rightarrow \{1, 0, -1\}$ , such that, for any event  $E$ ,  $LR(E, w(E))$  and  $RL(E, -w(E))$ . The zero-sum condition thus asserts that, given the same information, both agents must *agree* as to which of the two envelopes is more desirable (or if both envelopes are equally desirable, they must agree that this is the case). If  $w(E) = 1$  then, given  $E$ , the right-hand envelope is judged by both players to be more desirable (i.e., the left hand player will accept the offer to exchange, while the right-hand player will refuse); if  $w(E) = -1$  then the left-hand envelope is judged by both players to be more desirable; and if  $w(E) = 0$ , then both players agree that neither envelope is to be preferred to the other (i.e., they will both be indifferent to the exchange). The assumption that both agents ought to subscribe to the infinitary form of the sure-thing principle requires that the value of the function  $w$  is preserved under arbitrary disjoint unions.

Now, suppose that both players, after opening their envelopes, share their opinions as to whose envelope is to be preferred, so that their opinions are made common knowledge. Then it follows from what we have said so far that, if they are to be rational, their opinions *must* coincide. We may state this condition formally by defining two random variables  $w_L$  and  $w_R$  corresponding to the left and right-hand players' assessments of what they ought to do after opening their respective envelopes, i.e.,  $w_L(\omega) = w(E_L(\omega))$  and  $w_R(\omega) = w(E_R(\omega))$ . It follows that if  $w_L = \alpha$  and  $w_R = \beta$  are common knowledge, then  $\alpha = \beta$ .<sup>40</sup> In other words, if the two agents are rational, they cannot 'agree to disagree' as to whose envelope is to be preferred. Thus, if the actual exchange between the players is conducted in such a way that it is a necessary condition for the trade to go through that it is common knowledge that both players agree to the exchange (e.g., if the trade is only made official by a "handshake" between the two players) then, if both agents are rational, the exchange will not take place.<sup>41</sup>

In order to see what constraints this condition imposes on the deliberative methodologies of the two players, we should say something more about how it is that their respective decisions might come to be common knowledge in the first place. One way for such common knowledge to be generated is through a repeated sequence of exchanges of information. Such a process would begin with both players stating their initial opinions as to the value of  $w$ . Each player would then revise his opinion in the light of what the other says, and they would both report their revised opinions to each other.<sup>42</sup> They then revise their opinions again and report again, and the process is continued *ad infinitum*. It can be shown that, in the limit, this process suffices to generate common knowledge between the two players. It thus follows that, if such a process is carried out *ad infinitum*, the opinions of the two agents must, in the limit, converge.<sup>43</sup>

<sup>40</sup>The proof of this claim is exactly the same as that given in Aumann (1976) if the conditional probability function is replaced by the function  $w$ . We leave it to the reader to verify this fact.

<sup>41</sup>It is important to stress that this result only holds if it is assumed that both players subscribe to the *infinitary* version of the sure-thing principle. If rationality only requires that the agent's conditional preferences be preserved under finite disjoint unions, then it does not follow that two rational players cannot shake hands and agree to an exchange. Thus, the exchange paradox does not provide a counterexample to the well-known 'no-trade' theorem in Economics (Milgrom and Stokey (1980)), which denies the possibility of trade between two expected utility maximizers once market conditions have reached a state of pareto optimality. The proof of the no-trade theorem turns on the fact that if an agent's conditional preferences are the result of calculations of his expected gain, then his preferences will be preserved under finite disjoint unions (the role of this assumption in the proof is highlighted in Rubinstein and Wolinsky (1990)), but as we have already seen this is not always true in infinite settings (in particular, it does not hold when the expected gains assessed over the whole space do not converge absolutely). Thus, the no-trade theorem does not hold in general in the infinite case. Aumann's original result, however, which asserts that two agents with a common prior cannot agree to disagree about the posterior probabilities of an event, does apply in the infinite case.

<sup>42</sup>In addition to knowledge of the function  $w$ , this process requires that it be common knowledge between the players what type of information each player receives. In this simple setting the assumption is a natural one, but in more complicated settings it is totally unrealistic. It is this fact (that it is only in highly contrived settings that the information partition for each agent is made public in advance) which severely limits the practical significance of Aumann's theorem and other related results.

<sup>43</sup>In Geanakoplos and Polemarchakis (1982) it is shown that for finite information partitions, common knowledge of the players' opinions must occur after a number of steps not greater than the sum of the number of the cells in the two partitions. In Parikh and Krasucki (1990), this result is generalized so as to apply to the case of more than two agents who make sequential pairwise reports to each other in accordance with a 'fair' communication protocol. The fact that common knowledge must occur in the limit even if the information partitions are countably infinite was pointed out to me by Haim

This requirement that the two players must ultimately reach some consensus as to who the ‘winner’ of the game is, clearly rules out the possibility that both players decide whether to switch on the basis of a straightforward assessment of their expected gains, for, if this were the case, the repeated process described above would provide neither player with any information as to the number of the ticket in their opponent’s envelope. Both of the players would still agree to the exchange, even after it had become common knowledge that both players would agree to do so. But this is incoherent.

On the other hand, the requirement of convergence does not rule out the possibility that *one* of the two agents ought always to agree to the exchange, provided the same is not true of the other agent, but it is here that the symmetry of the situation enters the story. The symmetry of the situation entails that if one of the two players has reason to switch, after finding ticket number  $n$  in his envelope, then so too does the other. Thus if one of the two players ought always to switch, then so too should the other. Hence, rationality requires that neither of the two players adopt a methodology based on the straightforward calculation of expected gain relied upon in the argument for switching.

### 2.3 Symmetrical Expectations and the Iterated Subtraction Model

We have thus far argued that the symmetry of the situation precludes the agent from coherently adopting a deliberative methodology according to which he ought to switch no matter what the number of his ticket may be. This, however, is a very weak constraint on the agent’s choice of method. It does not, for example, rule out the possibility of the agent’s adopting a methodology which, for certain values of  $n$ , would instruct him to switch despite knowing that he would have switched had he chosen the other envelope. Nor does it preclude further iterations of this counterfactual. That is, it does not rule out the possibility that the agent ought to switch despite knowing that he ought to have switched had he chosen the other envelope, and despite knowing that he would have known *this* (viz., that he ought to have switched had he chosen the other envelope) had he chosen the other envelope, and so forth.

In order to assess more fine-grained possibilities of this sort, we must offer some positive account of what deliberative methodology the agent *should*, in fact, employ in deciding whether or not to switch once he has opened his envelope and observed the number of the ticket inside.

In this section, we offer one such account, which we term the “Iterated Subtraction” methodology, or ISM. In the context of this model, the agent’s expected gain, calculated in the standard way, is treated as a first-order estimate of what the agent ought to expect to acquire from the exchange. This initial estimate is subsequently revised through an iterated process of counterfactual self-reflection, in which the agent tries to take seriously what he would have thought had he, in fact, chosen the other envelope. The reason that the process repeats itself is that the general instruction to “adjust your expectations by what you would have expected had you chosen the other envelope,” has a recursive form. Thus every time the agent adjusts his expectations to account for what he would have thought has he chosen the other envelope, he ought to realize that he would have made similar adjustments had he chosen the other envelope, so that he ought to make further adjustments to account for this fact, and so forth.

As in the last section, it will again be helpful to imagine that the agent is involved in a two-player game in which he is playing an opponent who has the same information and the same utilities as him, and who is in possession of the other envelope. In this context, the crucial assumption of ISM is that it is common knowledge between the two agents that their expectations, though perhaps different, are, in a certain sense (to be made more precise) *equally legitimate*.

To see how the method works, let us suppose that the agent opens his envelope and finds ticket number  $n$  inside. Initially, without taking any account of his opponent’s expectations, the agent assesses his expected gains in the straightforward manner suggested in the argument for switching:

$$U_0(n) = \rho(n)f(n+1) - (1 - \rho(n))f(n)$$

Since the agent thinks that his opponent’s epistemic position is no less legitimate than his own, he determines that he ought to take into account his opponent’s expectations in deciding whether or not

---

Gaifman in a workshop on common knowledge held at Columbia University in the Summer of 2011.

to switch, but, unfortunately, he does not know exactly what his opponent's expectations are since he does not know whether his opponent finds ticket  $n + 1$  or ticket  $n - 1$  in his envelope. However, since he knows the probabilities of these two events, he can assess what he *expects* his opponent's expected gain from switching will be. This is just:

$$V_0(n) = \rho(n)U_0(n + 1) + (1 - \rho(n))U_0(n - 1)$$

Now, since the agents are engaged in a zero-sum game, if the agent is to take his opponent's expectations seriously, he ought to reckon his opponent's expected gains as his own expected losses. Thus, the agent ought to revise his initial assessment by subtracting from it his own expectation of his opponent's expected gains. The agent's revised expected gain is then:

$$U_1(n) = U_0(n) - V_0(n)$$

But now that the agent has revised his initial estimate, he should expect that his opponent will have done the same, and his expectation of his opponent's revised expected gain is:

$$V_1(n) = \rho(n)U_1(n + 1) + (1 - \rho(n))U_1(n - 1)$$

To take his opponents revised expectations into account, the agent adjusts his own expectations by again altering his expected gain from  $U_1(n)$  to

$$U_2(n) = U_1(n) - V_1(n)$$

Clearly this sort of reasoning can be iterated indefinitely. After  $m$  iterations, the agent's revised expected gain from switching envelopes is:

$$U_{m+1}(n) = U_m(n) - V_m(n),$$

where:

$$V_m(n) = \rho(n)U_m(n + 1) + (1 - \rho(n))U_m(n - 1)$$

If we let the function  $U$  be the point-wise limit of the sequence of functions  $\{U_m\}$ , i.e.:

$$U(n) = \lim_{m \rightarrow \infty} U_m(n)$$

then ISM is the methodology which instructs the agent to base his decision as to whether or not to switch (after opening his envelope and observing ticket numebr  $n$ ) on the sign of  $U(n)$ .<sup>44</sup>

In the rest of this section we will state a few very elementary results about the behavior of ISM. To begin with, let us consider the simple version of the paradox in which  $\rho = 1/2$ . We then have the following general result:

*Claim 1.* If  $\rho = 1/2$  and  $f$  is a polynomial function of order  $k$ , then  $U_m(n) = 0$ , for all  $n$  and for all  $m > k$ .

Thus, in the simple version of the paradox, provided  $f$  is a polynomial function, ISM will instruct the agent to be indifferent to the exchange, no matter what the number of his ticket may be (see Appendix A).

Now, let us consider the specific case in which  $f(n) = 2^n$ , since this is the version of the simple paradox most often discussed in the literature. In this case, it turns out that ISM also recommends that after opening his envelope, no matter what ticket he sees, the agent ought to be indifferent to the exchange.<sup>45</sup>

<sup>44</sup>We will allow  $U$  to take values in the extended real line, since it seems natural to suppose that if  $U_m(n) \rightarrow +\infty$ , then the agent ought to switch, and similarly, that the agent ought to refuse to switch when this sequence goes to  $-\infty$ . In all other cases, however, we will assume that the methodology leaves it undetermined what the agent ought to do.

<sup>45</sup>More generally, if  $\rho = 1/2$  and  $f(n) = c^n$  ( $c > 1$ ), then  $U_m(n) = k^m U_0(n)$ , where  $k = 1 - \left(\frac{c^2+1}{2c}\right)$ . Thus, if  $c < 2 + \sqrt{3}$ ,  $U(n) = 0$ , otherwise  $U$  is everywhere undefined.

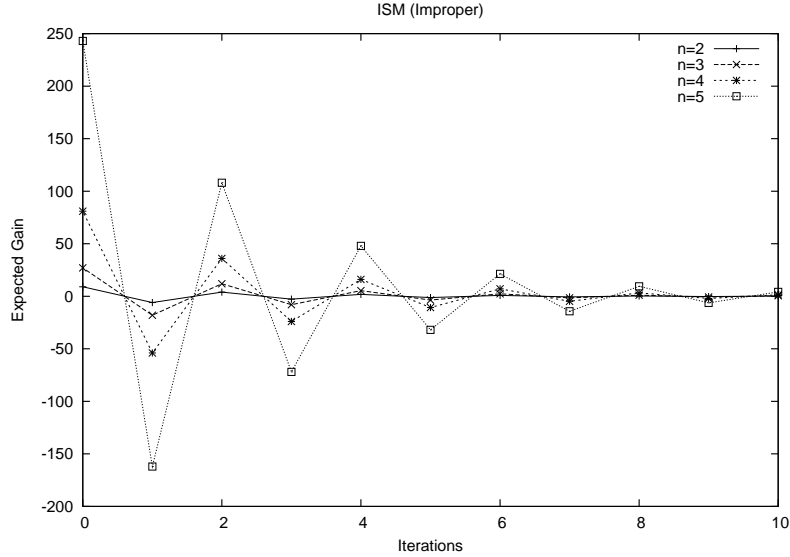


Figure 1: ISM Reasoning in the case where  $\rho(n) = 1/2$  and  $f(n) = 2^n$ .

*Claim 2.* If  $\rho = 1/2$  and  $f(n) = 2^n$ , then  $U(n) = 0$ , for all  $n$ .

In this case, the agent's expectations follow the pattern of a damped oscillation: at every stage of the iteration, the magnitude of what the agent expects to net from switching envelopes decreases, and, at each stage, the agent alternates between expecting to gain and expecting to lose (see Figure 1). In this context, it is natural to view ISM as the requirement that the agent's opinions be allowed to 'equilibrate' with those of his opponent (or in the single agent case, with his own counterfactual expectations) before they can be acted upon. When the agent first learns of the number of the ticket in his envelope, this information causes an initial perturbation in the agent's expectations, which, in this case, gradually settles down as the agent and his opponent take each other's points of view into account.

With respect to simple forms of the paradox in which the agent's conditional probabilities are always uniform, it is natural to suppose that the agent ought to remain indifferent to the exchange even after he has opened his envelope and observed the number of the ticket inside. The more interesting and difficult cases, however, concern *proper* forms of the paradox. For, here, there are no obvious grounds upon which to base a judgment as to what the agent should do after opening his envelope. It is in these cases, therefore, that ISM (and other proposed methodologies) may be employed to interesting effect.

Let us consider again the proper paradox described at the end of section 2. What does ISM instruct the agent to do in this case after he has opened his envelope and observed the number of the ticket inside? Interestingly enough, the answer depends on whether the number of the observed ticket is *even* or *odd*:

*Claim 3.* If  $\rho(n) = \begin{cases} \frac{2}{3} & n < 0 \\ \frac{1}{3} & n \geq 0 \end{cases}$ , and  $f(n) = 3^n$ , then  $U(n) = \begin{cases} +\infty & \text{if } n \text{ is odd.} \\ -\infty & \text{if } n \text{ is even.} \end{cases}$

In other words, if the agent opens his envelope and finds an odd-numbered ticket inside, he ought to agree to the exchange; otherwise, he ought to refuse it.<sup>46</sup> It is not entirely clear why the agent's decision to switch should depend, in this case, upon the parity of the number of the ticket in his envelope. The phenomenon clearly has something to do with the fact that the agent's prior probability distribution has a unimodal form (in this case, it is symmetrical about zero). Indeed, the typical pattern of damped oscillation proceeds as normal until the point at which one of the two agents is forced to entertain some

<sup>46</sup>Since this specific example is merely meant to be illustrative, and since the proof of claim 3 is in itself both cumbersome and unilluminating, we omit the proof of this claim.

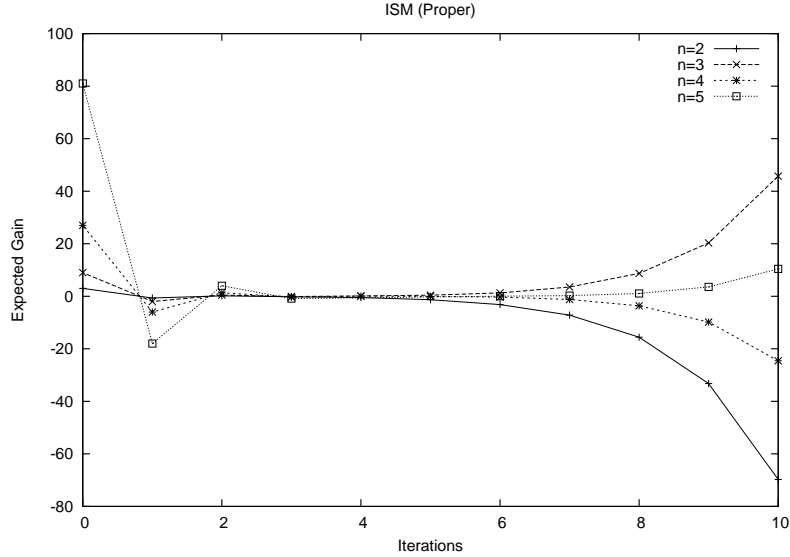


Figure 2: ISM Reasoning in the case where  $\rho(n) = \begin{cases} \frac{2}{3} & n < 0 \\ \frac{1}{3} & n \geq 0 \end{cases}$  and  $f(n) = 3^n$

hypothesis to the effect that either the other agent sees ticket number zero in his envelope, or else the other agent thinks that he sees zero in his envelope, or else the other agent thinks that he thinks that the other agent sees zero in his envelope, etc., and at this point the expectations of the two agent's diverge (see Figure 2).

The ISM method simply represents an initial attempt to make sense of how an agent ought to revise his expectations when he knows that they are based on biased information. As such, one should not invest too much stock in the results stated above. Indeed, the model itself suffers from a number of important defects, the most noteworthy of which is that it can only be meaningfully applied if we assume that the number of the winning ticket is *unbounded* in either direction. This is because, by introducing a bound on this value, we undermine the common knowledge assumption upon which the method is predicated.

Suppose, for example, that the number of the winning ticket has a lower bound of zero. If an agent opens his envelope and sees ticket number 1, then while he may adjust his initial expectations in accordance with ISM, he *cannot* assume that his opponent will do the same, since it may be that his opponent sees ticket number zero, in which case his opponent knows with certainty that switching will earn him  $f(1)$ . At this point, of course, it would be absurd for the agent's opponent to revise his expectations in accordance with ISM, and so, by admitting the possibility that his opponent may no longer be revising his expectations, the agent's own revisions come to a halt. In general, if the number of the winning ticket has a lower bound of zero, then if the agent sees ticket number  $n$  in his envelope, he will only be able to revise his expectations in accordance with ISM  $n$  times before the assumption that both agent's have equally legitimate expectations can no longer be sustained.<sup>47</sup>

### 3 Conclusion

In this essay, I have argued that the philosophical significance of the Exchange Paradox does not consist in what it reveals to us about the limits of finitary reasoning. Rather, the importance of the paradox

<sup>47</sup>This limitation of the ISM model, in fact, serves to highlight a rather interesting point, namely, that while it is often the case that two agents have common knowledge that their respective information was obtained from equally valid sources, it is almost *never* the case that they have common knowledge that the conclusions that they form on the basis of this information are equally legitimate.

consists in the new insight it provides us into the way in which judgments of *symmetry* can serve to inform an agent's decision-making. Traditionally, epistemologists have focused their attention on evidential symmetries which impose constraints on the doxastic attitudes that an agent can adopt towards a space of relevant contingencies. But an agent's deliberations are just as often impacted by the higher order symmetries inherent in a situation. Such symmetries serve to fix the scope of the agent's deliberative methodologies, and in this way, determine when an agent's decision is to count as biased. The Exchange Paradox is of interest precisely because it provides an example of a case in which the general prescription to act in an unbiased way imposes direct constraints on an agent's rational preferences. At its most general level, then, the real question raised by the paradox is the question of how a rational agent ought to cope with bias. What *should* the agent do after he has opened his envelope and observed the number of the ticket inside? This is a concrete question that can be posed in either the infinite or the finite setting, and in either case its answer is unclear.

## Appendix A:

In this appendix, we prove the following claim: If  $\rho = 1/2$  and  $f$  is a polynomial function of order  $k$ , then  $U_m(n) = 0$ , for all  $n$  and for all  $m > k$ .

**Lemma.** *If  $\rho = 1/2$  and  $f$  is a polynomial function of order  $k$ , then*

$$U_m(n) = \sum_{k=0}^{2m+1} \binom{2m+1}{k} \frac{(-1)^{k+m+1} f(n+k-m)}{2^{m+1}}$$

*Proof.* The proof proceeds by induction on  $m$ . If  $m = 0$ , the theorem is easily verified. Suppose the theorem holds for all numbers less than  $m$  ( $m \geq 1$ ). Then:

$$\begin{aligned} U_m(n) &= U_{m-1}(n) - \frac{1}{2}(U_{m-1}(n-1) + U_{m-1}(n+1)) \\ &= \frac{1}{2^m} \left( \sum_{k=0}^{2m-1} \binom{2m-1}{k} (-1)^{k+m} f(n+k-m+1) \right) \\ &\quad - \frac{1}{2^{m+1}} \left( \sum_{k=0}^{2m-1} \binom{2m-1}{k} (-1)^{k+m} f(n+k-m) \right) \\ &\quad - \frac{1}{2^{m+1}} \left( \sum_{k=0}^{2m-1} \binom{2m-1}{k} (-1)^{k+m} f(n+k-m+2) \right) \\ &= \frac{1}{2^{m+1}} \left\{ 2 \left( \sum_{k=1}^{2m} \binom{2m-1}{k-1} (-1)^{k+m+1} f(n+k-m) \right) \right\} \\ &\quad + \frac{1}{2^{m+1}} \left( \sum_{k=0}^{2m-1} \binom{2m-1}{k} (-1)^{k+m+1} f(n+k-m) \right) \\ &\quad + \frac{1}{2^{m+1}} \left( \sum_{k=2}^{2m+1} \binom{2m-1}{k-2} (-1)^{k+m+1} f(n+k-m) \right) \end{aligned}$$

Regrouping the terms according to the argument of  $f$ , we have:

$$\begin{aligned} U_m(n) &= \frac{1}{2^{m+1}} \{ (-1)^{m+1} f(n-m) + (-1)^m (2m+1) f(n-m+1) \} \\ &\quad + \frac{1}{2^{m+1}} \{ (-1)^{m+1} (2m+1) f(n+m) + (-1)^m f(n+m+1) \} \\ &\quad + \frac{1}{2^{m+1}} \left\{ \sum_{k=2}^{2m-1} (-1)^{k+m+1} f(n+k-m) \left[ 2 \binom{2m-1}{k-1} + \binom{2m-1}{k} + \binom{2m-1}{k-2} \right] \right\} \end{aligned}$$

From the recursive definition of the binomial coefficients:

$$\begin{aligned} 2 \binom{2m-1}{k-1} + \binom{2m-1}{k} + \binom{2m-1}{k-2} &= \left[ \binom{2m-1}{k} + \binom{2m-1}{k-1} \right] + \left[ \binom{2m-1}{k-1} + \binom{2m-1}{k-2} \right] \\ &= \binom{2m}{k} + \binom{2m}{k-1} \\ &= \binom{2m+1}{k} \end{aligned}$$

Substituting this back into the above, we obtain the desired result.  $\square$

The above claim follows from this lemma and the fact that for any polynomial function  $P(x)$  of order  $< n$ :

$$\sum_{j=0}^n (-1)^j \binom{n}{j} P(j) = 0.$$

## References

- Aumann, R. J. (1976). Agreeing to disagree. *The Annals of Statistics*, 4(6):1236–1239.
- Broome, J. (1995). The two-envelope paradox. *Analysis*, 55(1):6–11.
- Clark, M. and Shackel, N. (2000). The two-envelope paradox. *Mind*, 109(435):415–442.
- Dawid, A. P., Stone, M., and Zidek, J. V. (1973). Marginalization paradoxes in bayesian and structural inference. *Journal of the Royal Statistical Society. Series B (Methodological)*, 35(2):189–233.
- de Finetti, B. (1974). *Theory of Probability, Volume 1*. John Wiley and Sons.
- Gaifman, H. and Vasudevan, A. (2010). Deceptive updating and minimal information methods. *Synthese* (forthcoming).
- Geanakoplos, J. and Polemarchakis, H. (1982). We Can’t Disagree Forever. *Journal of Economic Theory*, 28:192–200.
- Jaynes, E. T. (2003). *Probability Theory: The Logic of Science*. Cambridge University Press.
- Kass, R. E. and Wasserman, L. (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, 91(435):1343–1370.
- Keynes, J. M. (1920). *A Treatise on Probability*. Cosimo, Inc., New York, NY, 2006 edition.
- Levi, I. (1987). The demons of decision. *The Monist*, 70:193–211.
- Meacham, C. and Weisberg, J. (2003). Clark and Shackel on the two-envelope paradox. *Mind*, 112(448):685–689.
- Milgrom, P. and Stokey, N. (1980). Information, trade and common knowledge. *Journal of Economic Theory*, 26(1):17–27.
- Nalebuff, B. (1989). Puzzles: The other person’s envelope is always greener. *The Journal of Economic Perspectives*, 3(1):171–181.
- Norton, J. (1998). When the sum of our expectation fails us: The exchange paradox. *Pacific Philosophical Quarterly*, 78:34–58.
- Parikh, R. and Krasucki, P. (1990). Communication, Consensus, and Knowledge. *Journal of Economic Theory*, 52:178–189.
- Priest, G. and Restall, G. (2008). Envelopes and indifference. In Cédric Dégrémont, L. K. and Rückert, H., editors, *Dialogues, Logics and Other Strange Things: Essays in Honor of Shahis Rahman*. College Publications.
- Rubinstein, A. and Wolinsky, A. (1990). On the logic of “agreeing to disagree” type results. *Journal of Economic Theory*, 51:184–193.
- Rudin, W. (1976). *Principles of Mathematical Analysis*. McGraw Hill, third edition edition.
- Savage, L. (1954). *The Foundations of Statistics*. John Wiley and Sons.
- Scott, A. D. and Scott, M. (1997). What’s in the two envelope paradox? *Analysis*, 57(1):34–41.

Sobel, J. H. (1994). Two envelopes. *Theory and Decision*, 36:69–96.

van Fraassen, B. C. (1989). *Laws and Symmetry*. Clarendon: Oxford University Press.

Wagner, C. G. (1999). Misadventures in conditional expectation: The two-envelope problem. *Erkenntnis*, 51(2/3):233–241.