Book review:
Paradoxes in Probability Theory, by William Eckhardt

Olle Häggström

William Eckhardt's recent book *Paradoxes in Probability Theory* has an appetizing format with just xi+79 pages. Some might say it's a booklet rather than a book. Call it what you will, this thought-provoking text treats the following seven delightful problems or paradoxes.

1. **The Doomsday Argument.** Let a given human being's *birth rank* be defined as the number of human beings born up to and including his or her birth. Consider the scenario that humanity has a bright future, with a population of billions thriving for tens of thousands of years or more. In such a scenario, our own birth ranks will be very small, as compared to the "typical" human. Can we thus conclude that Doomsday is near?

2. **The Betting Crowd.** You are in a casino, together with a number of other people. All of you bet on the roll of a pair of fair dice *not* giving double sixes; either all of you win, or all of you lose, depending on the outcome of this single roll. You seem to have probability 35/36 of winning, but there's a catch. The casino first invites one person to play this game. If the casino wins, then the game is played no more; otherwise another 10 people are invited to play. If the casino wins this time, play closes, whereas otherwise 100 new players are invited. And so on, with a tenfold increase of the number of players on each round. This ensures that when the whole thing is over, more than 90% of all players have lost. Now what is your probability of winning?

3. **The Simulation Argument.** Assume that computer technology continues to develop to the extent that eventually we are able to run, at low cost, detailed simulations of our entire planet, down to the level of atoms or whatever is needed. Then future historians will likely run plenty of simulations of world history during those interesting transient times of the early 21st century. Hence, the number of people living in the real physical world in 2012 will be vastly outnumbered by the number of people who believe themselves to do so but actually live in computer simulations run in the year 2350 or so. Can we thus conclude that we probably live in a computer simulation?

4. **Newcomb's Paradox.** An incredibly intelligent donor, perhaps from outer space, has prepared two boxes for you: a big one, and a small one. The small one (which might as well be transparent) contains \$1,000. The big one contains either \$1,000,000 or nothing. You have a choice between accepting both boxes, or just the big box. It seems obvious that you should accept both boxes (because that gives you an extra \$1,000 irrespective of the content of the big box), but here's the catch: The donor has tried to predict whether you will pick one box or two boxes. If the prediction is that you pick just the big box, then it contains \$1,000,000, whereas if the prediction is that you pick both boxes, then the big box is empty. The donor has exposed a large number of people before you to the same experiment, and predicted correctly 90% of the time, regardless of whether subjects choose one box or two.[1] What should you do?

5. **The Open Box Problem.** This is the same as Newcomb's Paradox, except that you get to see the contents of the big box before deciding.

6. **The Hadron Collider Card Game.** Phycisits at CERN are looking for an hitherto undetected particle X.[2] A radical physical theory Y has been put forth,[3] that particle X can in principle be produced, except that "something in the future is trying by any means available to prevent the production of [particle X]". A way to test theory Y is as follows. Prepare and shuffle a deck with a million cards, including one ace of spades. Pick one card at random from the deck, after having made international agreements to abandon the search for particle X provided the card picked turns out to be the ace of spades. If the ace of spades is picked, this can be seen as evidence in favor of theory Y. Does this make sense?

7. **The Two-Envelopes Problem.** Two envelopes are prepared, one with a positive amount of money, and the other with twice that amount. The envelopes are shuffled, and you get to pick one, and open. You may then decide whether you wish to keep that amount, or to switch to the other envelope. If the amount you observe is X, then the amount in the other envelope is either $X/2$ or $2X$ with probability $1/2$ each,

[1] Eckhardt forgets to mention this last condition ("regardless of..."), but it is clear that he intends it.

[2] In Eckhardt's account X is the Higgs boson, which unfortunately (for his problem formulation) has been detected since the time of writing.

[3] It really has [NN].

> for an expectation of $\frac{X/2+2X}{2} = \frac{5X}{4}$ which is greater than X, so it seems that you should switch envelope. But this is true regardless of the value of X, so it seems that you have incentive to switch even before you open the envelope. This, however, seems to clash with an obvious symmetry between the two well-shuffled envelopes. What is going on here?

Of these seven paradoxes, two are new, whereas the other five are known from the literature – such as the Simulation Argument, which was put forth by Bostrom [B] in 2003 and has been the topic of intense discussion in the philosophy literature ever since.[4] The two new ones – the Betting Crowd and the Open Box Problem – were invented by Eckhardt, mainly as pedagogical vehicles to help thinking more clearly about the others.

Eckhardt has no ambition of providing complete coverage of the literature on the five previously known paradoxes. Rather, the task he sets himself is to resolve them, once and for all, and the review of previous studies that he does provide is mostly to set the stage for his own solutions. He claims to be successful in his task, but realizes that not everyone will agree about that: in the introductory paragraph of his chapter on Newcomb's Paradox, he writes that "there exist a variety of arguments both for and against one-boxing, but in keeping with the design of this book, I search for an incontrovertible argument. (Of course it will be controverted.)" The book is a pleasure to read, not so much for Eckhardt's solutions (which, indeed, I find mostly controvertible), but for the stimulus it provides for thinking about the problems.

Eckhardt is a financial trader with a background in mathematical logic and a keen interest in philosophy with a few academic publications in the subject prior to the present one. Hence it is not surprising that he may have a different view of what is meant by probability theory, as compared to an academic mathematician and probabilist such as myself, who considers the title *Paradoxes in Probability Theory* to be a bit of a misnomer. To me, probability theory is the study of internal properties of given probability models (or classes of probability models) satisfying Kolmogorov's famous axioms from 1933 [Ko], the focus being on calculating or estimating probabilities or expectations of various events or quantities in such models. In contrast, issues about how to choose a probability model suitable for a particular real-world situation (or for a particular philosophical thought experiment) are part of what we may call applied probability, but not of probability theory proper. This is not to say that we probabilists shouldn't engage in such

[4] Even I have found reason to discuss it, in an earlier book review in the *Notices* [Hä].

modelling issues (we should!), only that when we do so we step outside of the realms of probability theory.

In this strict sense of probability theory, all seven problems treated by Eckhardt fall outside of it. Take for instance the Two-Envelopes Problem, which to the untrained eye may seem to qualify as a probability problem. But it doesn't, for the following reason. Write Y and $2Y$ for the two amounts put in the envelopes. No probability distribution for Y is specified in the problem, whereas in order to determine whether you increase your expected reward by changing envelopes when you observe $X = \$100$ (say), you need to know the distribution of Y, or at least the ratio $P(Y = 50)/P(Y = 100)$. So a bit of modelling is needed. The first thing to note (as Eckhardt does) is that the problem formulation implicitly assumes that the symmetry $P(X = Y) = P(X = 2Y) = \frac{1}{2}$ remains valid if we condition on X (i.e., looking in the envelope gives no clue on whether we have picked the larger or the smaller amount). This assumption leads to a contradiction: if $P(Y = y) = q$ for some $y > 0$ and $q > 0$, then the assumption implies that $P(Y = 2^k y) = q$ for all integer k, leading to an improper probability distribution whose total mass sums to ∞ (and it is easy to see that giving Y a continuous distribution doesn't help). Hence the uninformativeness assumption must be abandoned. Some of the paradoxicality can be retained in the following way. Fix $r \in (0, 1)$ and give Y the following distribution:

$$P(Y = 2^k) = (1 - r)r^k \quad \text{for } k = 0, 1, 2, \ldots\ .$$

For $r > \frac{1}{2}$, it will always make sense (in terms of expected amount received) to switch envelope upon looking in it. Eckhardt moves quickly from the general problem formulation into analyzing this particular model. I agree with him that the behavior of the model is a bit surprising. Note, however, that $r > \frac{1}{2}$ implies $E[Y] = \infty$, so the paradox can be seen as just another instance of the familiar phenomenon that if I am about to receive a positive reward with infinite expected value, I will be disappointed no matter how much I get.

Consider next Newcomb's Paradox. Here, Eckhardt advocates one-boxing, i.e., selecting the big box only. The usual argument against one-boxing is that since the choice has no causal influence on the content of the big box, one-boxing is sure to lose \$1,000 compared to two-boxing no matter what the big box happens to contain. On the other hand, one-boxers and two-boxers alike seem to agree that if the observed correlation between choice and content of the big box reflected a causal effect of the choice on the box content, then one-boxing would be the right choice. Eckhardt's argument for one-boxing, in the absence of such causality, is an appeal to what he calls the Coherence Principle, which says that decision problems that can be put in *outcome alignment* should be played in the same way. Here outcome alignment is a particular case of what probabilists call a *coupling* [L]. Two decision problems with the same set of options to choose from are said to be outcome alignable if they can be constructed on the same probability space in such a way that any given choice yields the same outcome for the two problems. Eckhardt tweaks the original problem formulation NP to produce a causal variant NPc, where the choice does influence the box content causally, in such a way that a one-boxer gets the $1,000,000 with probability 0.9, and a two-boxer gets it with probability 0.1. He also stipulates the same probabilities for NP, and notes that NP and NPc can be coupled into outcome alignment. Since everyone agrees that one-boxing is the right choice in NPc, we get from the Coherence Principle that one-boxing is the right choice also in NP.

The trouble with Eckhardt's solution, in my opinion, is that his stipulation of the probabilities in NP for getting the million, given one-boxing or two-boxing, glosses over the central difficulty of Newcomb's Paradox. Suppose I find myself facing the situation given in the problem formulation. If I accept that I have a 0.9 probability of getting the million in case of one-boxing, and a 0.1 probability in case of two-boxing, then the decision to one-box is a no-brainer. But why should I accept that those conditional probabilities apply to me, just because they arise as observed frequencies in a large population of other people? It seems that most people would resist such a conclusion, and that there are (at least) two psychological reasons for this. One is our intuitive urge to believe in something called free will, which prevents even the most superior being from reliably predicting whether we will one-box or two-box.[5] The other is our notorious inability to take into account base rates (population frequencies) in judging uncertain features of ourselves, including success rates of future tasks [Ka]. The core issue in Newcomb's Paradox is whether we should simply overrule these cognitive biases and judge, based on past frequencies, our conditional probabilities of getting the $1,000,000 given one-boxing or two-boxing to be as stipulated by Eckhardt, or if there are other more compelling rational arguments to think differently.

The story is mostly the same with the other problems and paradoxes

[5] This intuitive urge is so strong that, in a situation like this, it tends to overrule a more intellectual insight into the problem of free will, such as my own understanding (following, e.g, Hofstadter [Ho] and Harris [Har]) of the intuitively desirable notion of free will as being simply incoherent.

treated in this book. The real difficulty lies in translating the problem into a fully specified probability model. Once that is done, the analysis becomes more or less straightforward. My main criticism of Eckhardt's book is that he tends to put too little emphasis on the first step (model specification) and too much on the second (model analysis). The few hours needed to read the book are nevertheless a worthwile investment.

References

[B] Bostrom, N. (2003) Are we living in a computer simulation? *Philosophical Quarterly* **53**, 243–255.

[Hä] Häggström, O. (2008) Book review: Irreligion, *Notices of the American Mathematical Society* **55**, 789–791.

[Har] Harris, S. (2012) *Free Will*, Free Press, New York.

[Ho] Hofstadter, D.R. (2007) *I Am a Strange Loop*, Basic Books, New York.

[Ka] Kahneman, D. (2011) *Thinking, Fast and Slow*, Farrar Straus Giroux, New York.

[Ko] Kolmogorov, A.N. (1933) *Grundbegriffe der Wahrscheinlichkeitsrechnung*, Springer, Berlin.

[L] Lindvall, T. (1992) *Lectures on the Coupling Method*, Wiley, New York.

[NN] Nielsen, H.B. and Ninomiya, M. (2008) Search of effect of influence from future in Large Hadron Collider, *International Journal of Modern Physics A* **23**, 919–932.