Riches and Rationality

J. Dmitri Gallow [†]

Abstract

A one-boxer, Erica, and a two-boxer, Chloe, engage in a familiar debate. The debate begins with Erica asking Chloe: 'If you're so smart, then why ain'cha rich?'. As the debate progresses, each gets clearer about what connection they see between rational choice and long run riches. Erica says: long run riches give evidence about rationality, so long as the long run is one on which you face the same choice, and choose the same way, over and over again. Chloe objects that Erica unfairly compares the long run riches of people who were afforded different opportunities. As Erica pushes Chloe to get clearer about which comparison *is* fair, Chloe rehearses familiar formulations of causal decision theory. She is eventually driven to reject them all, and instead endorse a novel version of causal decision theory. This new theory allows Chloe to forge a connection between rational choice and long run riches. In brief: Chloe concludes that it is not long run wealth but rather long run wealth *creation* which is symptomatic of rationality.

Erica and Chloe appear on a game show, *Beat the Predictor!*. In the final round, contestants are presented with two boxes: one transparent, one opaque. Inside the transparent box, there is $10,000. Inside the opaque box, there is either $1,000,000 or nothing. Contestants must choose whether to take the contents of only the opaque box ('one-box') or to take the contents of both boxes ('two-box'). Whether you one-box or two-box makes no difference with respect to whether the million dollars awaits in the opaque box, but, before filming, an advanced AI bot, known as *Newcomb*, analyzes MRI scans of contestants' brains, the results of psychometric testing, and their social media accounts in order to make a prediction about whether they will one-box or two-box. If Newcomb predicts that a contestant will one-box, then the million dollars is placed in the opaque box before filming begins. If Newcomb predicts that they will two-box, then the opaque box is left empty. The show has been running for several seasons now, and, in the final round, Newcomb's predictions tend to be about 90% reliable.[1]

Draft of April 1, 2019 Word count: 9,404
Comments appreciated ✉: jdmitrigallow@pitt.edu

[†] Thanks to Daniel Drucker for many helpful conversations on this material.

[1] That is: the probability that Newcomb predicted you one-box, given that you do, is 90%. And the probability that Newcomb predicted that you two-box, given that you do, is 90%. This is how I'll continue to understand 'reliable' throughout.

Erica and Chloe are informed of all of this once filming begins (and not before).[2]

Erica decides to one-box, and she walks away with a million dollars. Chloe decides to two-box, and walks away with ten thousand. After filming, Chloe and Erica share a coffee and discuss their strategies.

Chloe asks Erica: 'What were you *thinking*, taking only the one box?'.

Erica explains: 'It was a simple calculation; the expected value of one-boxing was much higher than the expected value of two-boxing. Going for just the one box was clearly the better bet.'[3]

Chloe protests: 'But, no matter which prediction was made, if you'd taken both boxes, you'd be $10,000 richer!'

Erica agrees with Chloe about this, but remains skeptical of her claim that she ought to have taken both boxes. 'If you're so smart', she asks Chloe, 'why ain'cha rich?'

Chloe cries foul: 'It's no fair comparing my performance with yours—you had more money in front of you! A fair comparison would put you and I in the same circumstances. And, if we were in the same circumstances, then I would have ended up $10,000 richer, no matter which prediction was made.'[4]

Erica remains skeptical, so Chloe reminds her about one of the other contestants, Fred. During an earlier round, they were all presented with two transparent cups, and asked to take either only the cup on the left (one-cup) or to take both cups (two-cup). In this case, too, Newcomb had made a prediction about how they would behave. If it predicted that they would two-cup, then the left cup was empty. If, however, it predicted that they would one-cup, then there was $100 in the left cup. The right cup contained $10 no matter what. While Erica and Chloe's left cups were both empty, Fred's left cup had the $100 in it. Fred proceeded to take only the $100, leaving the $10 behind.[5]

Chloe: 'Surely you don't think that Fred did the right thing. Surely you agree that Fred should have taken the $10.' Erica agrees.

Chloe: 'But Fred ended up richer than you. So, if you're going to point to your

[2] Nozick (1969)

[3] Neither Erica nor Chloe are risk-adverse, and both of them value money *linearly*. No matter how rich they get, each additional dollar is just as valuable to them as the one before it.

[4] Lewis (1981b), Joyce (1999), Wells (forthcoming).

[5] Soares & Levinstein (ms) and Greene (2018).

riches as evidence of your rationality, then shouldn't you also point to *Fred*'s riches as evidence of *his*?'[6]

Erica cries foul: 'No fair comparing my performance with Fred's—we were making different choices! A fair comparison would have Fred and I making the *same* choice. In that case, I would have ended up $10 richer than Fred, no matter what prediction was made.'

Chloe: 'Wait...what does it take for us to be making the *same* choice?'

Erica: 'Well, for starters, we should have all the same options available to us. You shouldn't count as more rational than me just because you're able to do something I can't. And we should also have all the same evidence—you can't gloat if you only did better because you knew something I didn't. In particular, we should have the same evidence about what will result from each of our options, in each possible state of the world. Actually, let's say that, if we face the same choice, then you and I have exactly the same beliefs about the relevant probabilities—the same probabilities of states, of options, and of states conditional on options. And we should want the same things. If you really need that $10,000, then maybe it makes sense for you to not risk ending up with nothing. So let's say that we each value all the possible outcomes to the same degree. If all that's true, then I think that, if I end up richer than you, that's a reason to think that I'm more rational than you are.'

Chloe: 'I see. So, given that understanding of when two people are making the same choice, even though you and Fred were making *different* choices with the cups, you and I were making the *same* choice with the boxes.'

Erica: 'Yeah, and so I think that your ending up with only the $10,000 is evidence that you made an irrational choice. (Though Fred's ending up richer than I isn't evidence that I made an irrational choice.)'[7]

Chloe: 'Hmm...maybe it's evidence (I'm not sure yet), but I didn't *have* to end up with only $10,000. I got unlucky, but I *could* have ended up with $1,010,000—like that other two-boxer, Chris, remember? Chris ended up richer than you, so doesn't that give evidence that *you* are irrational?'

Erica: 'Well, I don't think we should be thinking about how much money an *individual* one-boxer made and comparing it to how much money an *individual* two-boxer made. We should instead be thinking about how much money a one-boxer would make *on average*, if they were to play the game over and over again a

[6] Gibbard & Harper (1978) and Arntzenius (2008, §7).

[7] Ahmed (2014, §7.3)

large number of times. Here—' Erica grabs a napkin and writes down:

$$\begin{aligned}\mathbf{ER}(O) &= \Pr(M \mid O) \cdot 1,000,000 + \Pr({\sim}M \mid O) \cdot 0 \\ &= 90\% \cdot 1,000,000 + 10\% \cdot 0 \\ &= 900,000\end{aligned}$$

'Imagine that you make this choice over and over. Each time, there will either be a million dollars in the opaque box or not—write 'M' for 'there's a million' and '${\sim}M$' for 'there's not'. The probability that the million is there, given that you one-box—that's the 'O'—is 90%. So, in this hypothetical long run, if you are a one-boxer, about 90% of the time, you'll get a million, and about 10% of the time, you'll get nothing. So, on average, you'll get $900,000. On the other hand, if you're a two-boxer, then in the hypothetical long run, about 90% of the time, you'll get ten thousand, and about 10% of the time, you'll get $1,010,000. So you'll only get $110,000, on average.' Erica writes:

$$\begin{aligned}\mathbf{ER}(T) &= \Pr(M \mid T) \cdot 1,010,000 + \Pr({\sim}M \mid T) \cdot 10,000 \\ &= 10\% \cdot 1,010,000 + 90\% \cdot 10,000 \\ &= 110,000\end{aligned}$$

'So you didn't just get unlucky—it was *predictable* that you'd get unlucky, given that you chose as you did. Chris may have gotten lucky, but, even so, both your and his *expected returns* were lower than mine. And *that's* what makes both you and Chris irrational. Both of you should be expected, in the long run, to end up poorer than me.'[8]

Chloe: 'Well, wait—Chris happened to get richer than you when we played the game once; couldn't it similarly turn out that, in the long run, Chris gets richer than you? Couldn't you and Chris both be predicted to act similarly in the long run? In that case, Chris would make $10,000 more than you, on average.'

Erica: 'Sure, I mean, it's *possible* that I get unlucky in the long run where we play the game over and over again. But the long run still helps us to clarify our thinking about what it's rational to choose, because, as we play the game more and more times, the probability that my average return matches my expected return approaches 100%.[9] So, by imagining ourselves playing the game more and more

[8] Of course, Erica is supposing that the long-run probability of being in the state K, given that you choose A, is equal to your subjective probability that you're in state K, conditional on your choosing A. Erica won't blame someone who gets unlucky in the long run by having subjective probabilities which don't line up with the objective probabilities (nor will Chloe).

[9] Erica is here appealing to the weak law of large numbers; if she were more careful, she'd have said: 'for any ϵ, the probability that my average return and my expected return differ by no

times, we can transform a choice that we make in conditions of *uncertainty* into a choice that we make under conditions of *certainty* (or, as near to certainty as we like). Whatever is rational to do once is rational to do repeatedly, if you're facing the same choice over and over again. And it's irrational to choose to be poorer in the long run. So if a choice leads you to be poorer over the long run, that choice can't be the rational one to choose even once.'

Chloe: 'Okay, I guess I see how it's useful to think about how much you win, on average, in this hypothetical long run. But won't *Fred* end up making more money than you, in the hypothetical long run where we play the cup game over and over again?'

Erica: 'Well, in a sense, he will. I mean: in a hypothetical long run where Fred gets presented with $100 in his left cup over and over, and I get presented with nothing in my left cup, over and over, I'll end up poorer. But *that* long run isn't an appropriate long run to be using to think about how I ought to choose. In those long runs, Fred and I are repeatedly asked to make *different choices*. But someone who gets offered a better choice isn't rational for that reason. When you're thinking about how to choose, you should imagine a long run in which you're asked to make the *same choice* over and over again.'[10]

Chloe: 'I see. Okay, I think I agree with you that it's not fair to compare long runs in which you're facing different choices. Just because Fred does better in *his* long run than you do in *yours*, that doesn't mean that Fred is more rational than you are, since Fred got better choices than you did.'

Erica: 'Okay, good—but, notice that, when you and I play the game with the boxes, in the long run, we're facing the *same* choice over and over again. And I'll end up richer in that long run. So that shows us that you're not playing rationally.'

Chloe: 'Maybe that's what I want to disagree with. It seems to me that, just like comparing your and Fred's long runs is making an unfair comparison, comparing your and my long runs is similarly making an unfair comparison. You'll make more money, on average, in your hypothetical long run, but that's just because you're more likely to be rewarded by Newcomb in your long run than I am in mine. In your long run, you end up with a million dollars in front of you more often than I do in mine. But that just shows that you will be given more opportunities than me, in the long run. It doesn't show that you'll end up making the most of the opportunities you're provided. In fact, you're squandering those opportunities, ending up ten thousand dollars poorer, on average, than you could

more than ϵ approaches 100% as we play the game more and more times'.

[10] Ahmed (2014, §7.3.3)

have been, if only you had two-boxed.'

Erica: 'Well, that doesn't seem right to me. I'm *making* those opportunities for myself by deciding to one-box. If you get to decide whether to one-box or two-box, then you get to decide which long run to face, right?'

Chloe: 'Not at all—let's think about this hypothetical long run. We're playing the final round of *Beat the Predictor!* over and over again. I assume that how I choose in the first round won't causally affect what Newcomb is likely to predict in round two. In particular, one-boxing in round one doesn't cause Newcomb to predict that I'll one-box in round two—else, I'd agree with you that I should one-box in round one. Also, if that's the case—if my decision in round one causes Newcomb to predict differently in round two—then I don't think that what happens in this hypothetical long run tells me anything interesting about what to do in the one-off case, where I know I won't be on the game show ever again. And you agree with that—right? In that hypothetical long run, I'm making a different choice than the one I'm actually making, since, in that long-run, the consequences of my one-boxing are different.'

Erica: 'Yeah, okay, that's right. We should say that, in the long run, what you do in round one doesn't causally affect what happens in the other rounds. But *even so*, one-boxing *makes it more likely* that you'll face a good long run.'

Chloe: 'I agree—if, by 'makes it more likely that you'll have a good long run', you mean 'gives evidence that you'll have a good long run'. But I don't think that kind of evidence is relevant to how I ought to choose. Think about it like this: if how I choose in the earlier rounds doesn't causally affect what Newcomb predicts in the later rounds, then Newcomb could make all the predictions about how I'll choose in every round right at the beginning—right?'

Erica: 'I suppose it could, sure.'

Chloe: 'But then, in round one, it's already determined which long run I'm going to face. Nothing I do there will change that. So these aren't really opportunities that I'm making for myself. These are opportunities which Newcomb has either provided for me or not, before I get to make my choice.'

Erica: 'Okay, so I guess I agree that you're not literally *making* those opportunities for yourself, but I still think that you should want to give yourself the evidence that you're going to face the best long run possible. I mean, making money is the whole *point* of the game. Surely you agree that there's some connection between playing the game rationally and the money you get, right?'

Chloe: 'Of course there's *some* connection; I just don't think that the connection is what you say it is.'

Erica: 'Okay, so what is it, then? I've got my answer: you're choosing rationally if, and only if, someone who chooses like you would be expected to be at least as well off, in a hypothetical long run in which you make the same choice over and over, as anyone choosing in any other way would be expected to be, in a hypothetical long run in which *they* make the same choice over and over. In general, that's how I decide to what to do. When I have to make a decision, I list off all of the potential acts, $A_1, A_2, \ldots, A_M$, and I list off all the ways things might be, for all I have any control over, $K_1, K_2, \ldots, K_N$.[11] For each action A, I ask myself, firstly, how good it would be to choose A if it turns out that each K is true, and also how likely each K is, if I choose A. I multiply these together, add them up, and that gives me the *expected return* of choosing A' —she scribbles on the napkin:

$$\mathbf{ER}(A) = \sum_K \Pr(K \mid A) \cdot V(AK)$$

"$\Pr(K \mid A)$' is the probability that K obtains, given that I choose act A. And '$V(AK)$' is how much I value choosing A in the state K. I choose whichever act, A, has the highest expected return. But, so long as you and I are making the same choice—so long as we have the same acts available to us, we have the same beliefs about how likely each state K is, conditional on each act, and we value everything to the same degree, then your expected return for each act will be the same as mine. And so, if you and I are making the same choice, then I'll always end up making at least as much money as you, on average, in the long run.'

Chloe: 'I see.'

Erica: 'But what about you? You say that there's some connection between your rationality and your riches, or your expected riches—but what could that connection *be*, if it's not mine?'

Chloe: 'I guess I'll have to think about that for a bit.' She thinks, and then says: 'Well, I think that, when we're comparing your and my performance in the long run, we need to be considering a long run in which you and I are afforded the same opportunities.[12] So we need to *equalize* those opportunities. Maybe we

[11] Erica forgets to note: these K's are what Lewis (1981a) calls 'dependency hypotheses', and they are mutually exclusive and jointly exhaustive. (Erica, by the way, doesn't think that she *has* to use these kinds of states when making up her mind about what to do; any other partition of possibilities would do just as well. Even so, this is how she does it.)

[12] Wells (forthcoming).

should put it like this—' Chloe writes down:

$$\mathbf{CER}(A) = \sum_K \Pr(K) \cdot V(AK)$$

'There. ('*C*' for '*Chloe*'.) So, here's the suggestion: if you're rational, then you'll make the most money possible in a hypothetical long run in which you face the same choice over and over again—*and*, for each state in which you could face that choice, you are in that state, over the long run, a proportion of the time which is equal to your *unconditional* probability in that state (and not, like you would have it, the probability *conditional on* your action). Then, when you compare our long run riches, we're both facing the state K the same proportion of the time. So your long run is my long run, and we're afforded the same opportunities.'

Erica: 'Hold on. The probability that you're in state K could depend upon how likely you are to choose each act. Like, during the final round of *Beat the Predictor!*—the probability that there's a million dollars changes, depending upon how likely you are to one-box or two-box. If t is how likely you are to two-box, then...wait, hold on...' Erica walks off and comes back with a handful of napkins. She then writes out:

$$\begin{aligned}\Pr(M) &= \Pr(M \mid T) \cdot t + \Pr(M \mid O) \cdot (1-t)\\ &= t \cdot 10\% + (1-t) \cdot 90\%\\ &= 90\% - t \cdot 80\%\end{aligned}$$

Erica: 'Here. So, as the probability of your two-boxing rises—as t goes up towards 100%—it gets less likely that the million dollars is there. And as the probability of your two-boxing falls—as t goes down to 0%—it gets more likely that the million dollars is there. So which unconditional probability do you want to use?'

Chloe: 'I see. Okay—let's try this: say it's the probability *once I've chosen*. If you want to show me that I'm irrational, then you should show me that you could do better than I, in the long run which I would expect to face, after choosing, were I to make that choice over and over again.'

Erica: 'But isn't that just what I said? You're now telling me to calculate your 'Chloe' expected return, conditional on your having chosen A—let's write that '$\mathbf{CER}(A \mid A)$'—and that's this expectation here:' Erica writes:

$$\mathbf{CER}(A \mid A) = \sum \Pr(K \mid A) \cdot V(AK)$$

'But isn't that expectation just the same as mine?'

Chloe: 'Hmm...well, yes and no. I think we may agree about how to understand expected returns. What we're disagreeing about is how to *compare* people in terms of their expected returns. Let's try putting it like this: you say that it's fair to compare two people's expected returns in the long run, so long as it's a long run in which they face the same choice over and over again. So, since I chose to two-box, T, and you chose to one-box, O, you think it's fair to compare us by comparing $\mathbf{CER}(T \mid T)$ and $\mathbf{CER}(O \mid O)$. Since $\mathbf{CER}(T \mid T)$ is less than $\mathbf{CER}(O \mid O)$, I have a lower expected return than you do, and you think I'm irrational for that reason.'

Erica: 'Right.'

Chloe: 'But *I* think that it's not always fair to compare people's expected returns in long runs where they face the same choice, since it could be that one person's long run is better than the other's. Like, if I choose to two-box, and you choose to one-box, then your long run will look better than mine. It will look better not because you're being more rational in that long run, but just because there will be more money for the taking on your long run. So, before you call me irrational, I think we need to equalize the opportunities, and ask how you would have done in *my* long run. That is, I think we need to compare $\mathbf{CER}(T \mid T)$ with $\mathbf{CER}(O \mid T)$. Once we've equalized the opportunities in this way, I'll always end up $10,000 richer than you. So you don't have a higher expected return than I do, once we've equalized the opportunities we face.'

Erica: 'Wait—what's $\mathbf{CER}(O \mid T)$? That's the Chloe-expected return of one-boxing, in the long run you expect to have if you've two-boxed?'

Chloe: 'Yeah. Here, I just mean this: for any actions—call them '*A*' and '*B*'...' Chloe writes:

$$\mathbf{CER}(A \mid B) = \sum_K \Pr(K \mid B) \cdot V(AK)$$

'By conditioning the probability function on B, we look at the long run you'd expect to have if you selected B, and then we ask about how good it would be, in *that* long run, to have done A instead.'

Erica: 'Oh, interesting. So, maybe we can think about it like this:' Erica writes down the following.

Erica's Proposal: Choosing A is irrational iff the alternative, $\sim A$, is such that

$$\mathbf{CER}(A \mid A) < \mathbf{CER}(\sim A \mid \sim A)$$

Chloe's Proposal: Choosing A is irrational iff the alternative, $\sim A$, is such that

$$\mathbf{CER}(A \mid A) < \mathbf{CER}(\sim A \mid A)$$

Chloe: 'Alright, good. So you'll say that I'm irrational for two-boxing, since you do better on your long run than I do on mine. And I'll say that you're irrational for one-boxing, since I will do better in your long run than you will. Of course, these proposals only say something about decisions where you have two options: A and $\sim A$. Presumably, we'll want to generalize these proposals to handle cases where there's more than two options.'

Erica: 'I agree, but in the interests of simplicity, I'd prefer to just focus on the two-option case for now—is that okay?'

Chloe: 'Sure, that's fine with me. Okay...so *that's* my connection between riches and rationality. If you're rational, then you'll make the most possible of the long run you would face, were you to make the same choice over and over again. What's wrong with that?'

Erica: 'Let me think about it for a bit.' She thinks, and then says: 'Do you remember that game we played with the two doors?'

Chloe: 'Yeah, that one was strange.'

In the door game, contestants were asked to open one of two closed doors: one door black, the other white. Again, Newcomb made a prediction about how they would act. If it predicted that they would open the black door, then $100 was placed behind the white door and nothing was placed behind the black door. And if it predicted that they would open the white door, then $100 was left behind the black door, and nothing was left behind the white door. Newcomb's predictions are 80% reliable in this game. Contestants were told all of this in advance.[13]

Erica: 'So it looks to me like what you're proposing says that both choosing white and choosing black is irrational in the door game. Because the Chloe expected return of choosing black, in the long run in which you choose black, is $20, which is less than the Chloe expected return of choosing white in that same long run, which is $80.' She writes down:

$$\begin{aligned}\mathbf{CER}(B \mid B) &= \Pr(K_B \mid B) \cdot 0 + \Pr(K_W \mid B) \cdot 100 \\ &= 80\% \cdot 0 + 20\% \cdot 100 \\ &= 20 \\ \mathbf{CER}(W \mid B) &= \Pr(K_B \mid B) \cdot 100 + \Pr(K_W \mid B) \cdot 0 \\ &= 80\% \cdot 100 + 20\% \cdot 0 \\ &= 80\end{aligned}$$

[13] Gibbard & Harper (1978), Richter (1984), Weirich (1985), Egan (2007), Joyce (2012).

'(Here, I'm using 'K_W' to mean that Newcomb predicted you'd choose white, and 'K_B' to mean that it predicted you'd choose black.) So your proposal says that choosing black is irrational. And it says the same thing about choosing white, since the Chloe expected return of choosing white, in the long run in which you choose white, is $20, while the Chloe expected return of choosing black in that same long run is $80. (Black and white are perfectly symmetric in this game.)'

Chloe: 'Shoot. I suppose I could say that there's no rational choice here[14]...but I'd rather not. I'm not really sure what to say about the door case, honestly, but surely there's *some* rational choice to be made.'

Erica: 'Okay, so then, do you concede that my theory about the connection between rationality and riches makes better sense?'

Chloe: 'Well, hold on. I still think it's unfair to compare your performance in *your* long run with my performance in *mine*—if you're getting more money offered to you in your long run, that shouldn't speak in your favor. And I still think that, if we want to get a *fair* comparison, we need to equalize the opportunities afforded to us in each of our hypothetical long runs. What I'm thinking you've shown me here is that I had the wrong way of equalizing those opportunities.'

Erica: 'What's the alternative?'

Chloe thinks for a while, and then says: 'When I was thinking about what to do in the door game, I found myself vacillating back and forth between black and white. As soon as I found myself leaning towards black, white seemed like the better choice—because, after all, if I'm leaning towards black, that means that Newcomb probably predicted that I'd choose black, and so the money is more likely to be behind the white door. But, then, when I found myself leaning towards white, black seemed like the better choice—because, if I'm leaning towards white, then Newcomb likely predicted that I'd take white, and so the money is more likely to be behind the black door. After a while, I ended up stuck in the middle—and I didn't feel like I had anything much else to think about, so I just...*picked*. I ended up choosing black, but I could have gone for white instead.'[15]

Erica: 'You're a mystery to me, Chloe. What was the point of all that hemming and hawing? By the symmetry of the case, there's nothing to tell between black and white, so either choice should be permissible.'

Chloe: 'Maybe you're right about that...I'm not sure, really—I mean, if I'm more likely to choose black, then Newcomb was more likely to predict black, and that

[14] Harper (1986, p. 33)

[15] Skyrms (1990), Joyce (2012, 2018), Arntzenius (2008).

may be an important consideration...or, in any case, it breaks the symmetry. But my vacillation gives me an idea about how to decide upon the right hypothetical long run to be using in order to equalize the opportunities. As I was deliberating about what to do, my probability that I would choose black or white was changing. And there were some middling probabilities for choosing black and choosing white at which both options looked equally good to me —let's call these my *equilibrium* probabilities. I guess that, in this case, it would be a 50% probability for choosing black and a 50% probability for choosing white. So here's a new idea: let's use *these* probabilities when we think about my hypothetical long run. If the equilibrium probability for the final game with the boxes is 100% two-boxing and 0% one-boxing, this will agree with what I said earlier about two-boxing, but maybe it will allow me to avoid this trouble you're pointing out with the door game. I guess I'd have to think about how to find these equilibrium probabilities in general—and I hope that there always are some equilibrium probabilities like these—but that's the start of an answer.'[16]

Erica: 'Alright, so is this your proposal now? Choosing A is irrational if the alternative has a higher average return, in the hypothetical long run you'd face if you choose...well, the long run determined by your equilibrium probabilities?'

Chloe: 'Maybe we shouldn't be talking about an individual choice being rational. I'm not sure I should choose black every single time I play the door game, in the long run. Maybe, instead, I should choose black half of the time, and white half of the time. So maybe we need to re-think which things we're evaluating. Let's not say that selecting *an act* is rational or irrational. Let's say instead that it's a *method for selecting* acts which is rational or irrational—we can call it a 'strategy'. And a strategy needn't tell you to select one act—it could tell you instead to choose an act with a certain probability, and to choose the alternative with a certain probability.'

Erica: 'Okay, good.'

Chloe: 'So, then, in those terms, here's my proposal:' She writes:

Chloe's 2nd Proposal: A strategy, S, is irrational iff there's some alternative strategy, S^*, which has a higher expected return than it, in the long run faced by the strategy S:

$$\mathbf{CER}(S^* \mid S) > \mathbf{CER}(S \mid S)$$

$$\sum_K \Pr(K \mid S) \cdot V(S^*K) > \sum_K \Pr(K \mid S) \cdot V(SK)$$

[16] Chloe needn't worry—there is a way of working out these equilibrium probabilities in general, and there will always be some equilibrium, given this way of working them out. See Skyrms (1990) and Arntzenius (2008). (However, there could be more than one equilibrium. Chloe should probably think about that, but, unfortunately, it doesn't occur to her.)

Erica: 'That's really helpful, Chloe. So let's think about your strategy of picking black half of the time and white half of the time. If that's your strategy, then, in the long run, you should expect Newcomb to predict that you pick black half of the time, and you should expect Newcomb to predict that you pick white half of the time.'

Chloe: 'Right. And then, when I'm evaluating a strategy which tells you to choose black with some probability, call it 'b', I'll calculate its expected return, in the long run I expect to face, when I'm picking black 50% of the time, like this...' Chloe writes:

$$\begin{aligned}
\mathbf{CER}(S_b \mid S_e) &= \Pr(K_B \mid S_e) \cdot V(S_b K_B) + \Pr(K_W \mid S_e) \cdot V(S_b K_W) \\
&= 50\% \cdot (0b + 100(1-b)) + 50\% \cdot (100b + 0(1-b)) \\
&= 50(1-b) + 50b \\
&= 50
\end{aligned}$$

Erica: 'What do 'S_b' and 'S_e' stand for?'

Chloe: 'Oh, sorry—'S_b' says that I've adopted the strategy of choosing black with a probability of b, and 'S_e' says I've adopted the *equilibrium* strategy of choosing black with 50% probability.'

Erica: 'I see. So your expected return doesn't depend upon your probability of choosing black at all! In this long run, any strategy is as good as any other. So I guess your proposal doesn't say that picking black half of the time is irrational.'

Chloe: 'No, I don't think so. But I think that it *does* say that choosing black *every* time is irrational.' Because, if you choose black every time, then, in the long run, you should expect Newcomb to predict that you choose black 80% of the time, and you should expect Newcomb to predict that you choose white 20% of the time. And in that long run, choosing black every time has an average return of $20, whereas choosing white every time has a higher average return of $80.'

Erica: 'And I guess the same will be true for any strategy *other* than your 50/50 split?'

Chloe: 'Right—if your probability of taking black is greater than 50%, then taking white every time will do better than your strategy, in your long run. And if your probability of taking black is less than 50%, then taking black every time will do better than your strategy in your long run.'

Erica: 'I see...that's really cool, Chloe.' Erica scribbles on one of the napkins for a bit. Then she says: 'Well, wait. Actually, I'm a little bit confused. You say that, in the long run you expect to face, your strategy (like every *other* strategy) will have an average return of $50. But, as I calculate things, your average return in your long run will be $20.'

Chloe: 'No, I don't think so. Here, you calculate my expected return like this:' Chloe writes:

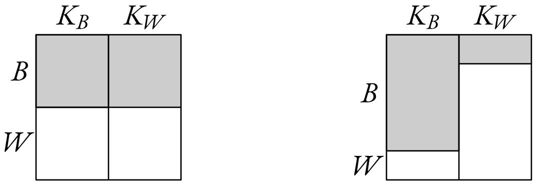
$$\begin{aligned}\mathbf{ER}(S_e) &= \Pr(K_B \mid S_e) \cdot V(K_B S_e) + \Pr(K_W \mid S_e) \cdot V(K_W S_e) \\ &= 50\% \cdot V(K_B S_e) + 50\% \cdot V(K_W S_e)\end{aligned}$$

And, if Newcomb predicted that you would choose black, then the value of the equilibrium strategy is just 50% times zero dollars plus 50% times 100 dollars,' She writes:

$$\begin{aligned}V(K_B S_e) &= \Pr(B) \cdot V(BK_B) + \Pr(W) \cdot V(WK_B) \\ &= 50\% \cdot 0 + 50\% \cdot 100\end{aligned}$$

'which is $50. And it's exactly the same if Newcomb predicted that you would choose white. So my average return in the long run will be $50.'

Erica: 'No, I think you're mis-calculating the value of using your strategy. You're ignoring important correlations between what your strategy tells you to do and what Newcomb has predicted. We agree that, in your long run, Newcomb predicts that you'll take black 50% of the time, and Newcomb predicts that you'll take white 50% of the time, but when *you* calculate the value of using your strategy, you're assuming that what your strategy recommends is independent of what Newcomb predicted. Here—' Erica draws out two squares like the ones shown below.



'You're assuming the distribution on the left—you're assuming that your strategy will tell you to pick black 50% of the time, independent of whether Newcomb predicted you'd pick black. But if that were so, then Newcomb wouldn't be a reliable predictor. You *should* be assuming the distribution on the right. When Newcomb predicts that you'll choose black, your strategy will tell you to pick white only 20% of the time. And, when Newcomb predicts that you'll choose white, your strategy will tell you to pick black only 20% of the time. So you should calculate the value of using your strategy, when Newcomb has predicted that you'll pick black, like this:' Erica writes:

$$\begin{aligned} V(K_B S_e) &= \Pr(B \mid K_B S_e) \cdot V(BK_B) + \Pr(W \mid K_B S_e) \cdot V(WK_B) \\ &= 80\% \cdot 0 + 20\% \cdot 100 \\ &= 20 \end{aligned}$$

'And, similarly, the value of using your strategy, when Newcomb has predicted that you'll pick *white*, will be $20. So the expected return of your strategy, in this long run, is going to be $20.'

Chloe: 'Oh...well, couldn't Newcomb just be good at predicting the probability that I'll choose black or white—that is, couldn't it just be good at predicting which *strategy* I end up adopting? After that, whether I pick black or white is just a roll of the dice, and surely Newcomb wouldn't be able to predict how those dice land.'

Erica: 'Well, I don't have any idea how Newcomb works, but I do know this: given that you take black, it's 80% likely to have predicted that you take black. That's what we were told, and that's all I'm assuming. And, you know, we're not talking about literal dice here (remember, the producers were very insistent that we couldn't flip coins or roll dice or anything like that). So whatever corresponds to the 'rolling of the dice', it's something going on up in your brain. And, remember, Newcomb saw scans of our brains. So maybe it *is* able to predict how the dice will land. Anyhow, let's suppose that it is, since that's the interesting case.'

Chloe: 'Hmm...okay, I guess that's right. So, if Newcomb can predict what I'll end up choosing when I play the equilibrium strategy, then the equilibrium strategy will have an average return of $20 in the long run. So what's going on here?' She thinks for a while, and then says: 'Okay, I think what's going on is this: when you calculated my expected return, in the long run, you took for granted that Newcomb could predict how I choose—so, you held fixed how likely Newcomb was to have predicted I'd take black, given that I do, $\Pr(K_B \mid B)$. But, when I was asking about someone *else's* expected return on *my* long run, I didn't do that...in fact, I *couldn't* have done that. Since I'm asking about my long run, I'm holding fixed the probability that Newcomb predicts black. But this, together with Newcomb's reliability, determines how likely I am to pick black.[17] Considering how any *other* strategy fares in that long run requires us to suppose that Newcomb isn't 80% reliable at predicting *that* strategy, in my long run. So, when you were calculating my expected return in my long run, you were supposing that I'm predictable; but, when I was thinking about someone *else's* expected return on my long run, I wasn't supposing that *they* were predictable (because that's

[17] $\Pr(B) = [\Pr(K_B) - \Pr(K_B \mid W)] \div [\Pr(K_B \mid B) - \Pr(K_B \mid W)]$

impossible).'

Erica: 'Yeah, that seems right—so doesn't this show us that the whole exercise of trying to 'equalize the opportunities', and ask about how someone else would fare in your long run, is confused from the get-go? In the door game, you shouldn't be comparing your performance in your long run with someone else's performance in your long run, since that *necessarily* involves an unfair comparison. It necessarily involves comparing someone who's predictable with someone who's not predictable. If you're going to make the comparison fair, both parties should be predictable if either is. And that's what my proposal accomplishes.'

Chloe: 'Well, maybe it just shows that, if I'm going to compare my performance in my long run with yours, I need to suppose that I'm *not* predictable. Then, the comparison will be fair.'

Erica: 'Is that a new proposal?'

Chloe: 'Yeah, I think so. How should I put it?' Chloe scribbles on the napkin for a bit, and then says: 'Here, let's not ask about the Chloe-expected return of a strategy—that's tantamount to assuming that I'm predictable. Instead, let's ask about the *unpredictable* expected return of the strategy—we can call that '**UER**." Chloe points at two equations from her napkin:[18]

$$
\begin{aligned}
\mathbf{CER}(S^* \mid S) &= \sum_K \Pr(K \mid S) \cdot V(S^* K) \\
&= \sum_K \Pr(K \mid S) \cdot \sum_A \Pr(A \mid S^* K) \cdot V(AK) \\
\mathbf{UER}(S^* \mid S) &= \sum_K \Pr(K \mid S) \cdot \sum_A \Pr(A \mid S^*) \cdot V(AK) \\
&= \sum_A \Pr(A \mid S^*) \cdot U(A \mid S)
\end{aligned}
$$

'(I'm using '$U(A \mid S)$' for the *unpredictable* value of the act A, in the long run faced by S—it's just $\sum_K \Pr(K \mid S) \cdot V(AK)$.) So, when we calculate the *unpredictable* expected return of a strategy, S^*, in the long run faced by the strategy S, we set each term $\Pr(A \mid S^* K)$ equal to $\Pr(A \mid S^*)$, and thereby ignore any correlations between the act you end up selecting and Newcomb's predictions.'[19] She then writes:

[18] Chloe and Erica both accept and take for granted that, for any proposition ϕ, and any partition of propositions $\{\psi_1, \dots, \psi_N\}$, $V(\phi) = \sum_i \Pr(\psi_i \mid \phi) \cdot V(\phi\psi_i)$. This is why Chloe allows herself to exchange $V(S^* K)$ with $\sum_A \Pr(A \mid S^* K) \cdot V(AK)$. (Chloe also doesn't intrinsically value her strategies, which is why she writes just '$V(AK)$' instead of '$V(AS^* K)$'.)

[19] Chloe is now evaluating strategies in the same way as Skyrms (1990), Arntzenius (2008), and Joyce (2012).

Chloe's 3rd Proposal A strategy, S, is irrational iff there's some alternative strategy, S^*, which has a higher *unpredictable* expected return than it, in the long run faced by the strategy S:

$$\mathbf{UER}(S^* \mid S) > \mathbf{UER}(S \mid S)$$

Erica thinks about this for a bit, and then says: 'I guess I'm a bit confused by this proposal, Chloe. I was asking you what connection you saw between rational choice and the money you win in the long run. Now, you're telling me that you think the rational choice will make more money than the alternatives, *not* in the hypothetical long run in which you make that choice over and over again, but instead in a *different* hypothetical long run—one in which you're not predictable?'

Chloe: 'Well...I'm just trying to make the comparison fair. The other strategies aren't predictable in my long run, so if we want a fair comparison, then *I* shouldn't be predictable, either.'

Erica: 'Yeah, I see why you're going for this proposal, but it still just seems wrong to me. You should care about the riches you expect to earn in your *actual* long run, and not some *other* long run which you know you definitely won't face.'

Chloe: 'This *is* my actual long run...it's just that the distribution of my choices throughout the long run has changed a bit from what you would expect. They've changed so as to make my strategy as predictable as any other.'

Erica: 'But aren't you just ignoring important information when you evaluate strategies in that way? You're pretending that you're not predictable, when you *know* that you are. I don't know. Maybe it would help if we could we think about what this proposal says about that game with the envelopes.'

In the envelope game, contestants were presented with two envelopes, labeled 'X' and 'Y'. Newcomb made a prediction about how the contestants would choose. There was a guaranteed \$20 placed in envelope Y, no matter what Newcomb predicted. If it predicted that they would take envelope X, then X was left empty. If it predicted that they would take envelope Y, then \$100 was placed in envelope X. With this game, Newcomb is 90% reliable.[20]

Chloe: 'Okay, sure. I guess we first need to figure out what the equilibrium strategy is in this game. If I'm definitely going to take envelope X, then I'll think that it's only 10% likely to contain the \$100. That's worse than a guaranteed \$20, so, if I'm definitely going to take X, Y looks like a more attractive option. And, if I'm definitely going to take Y, then I'll think that it's 90% likely that X contains

[20] *Cf.* Egan (2007).

$100, which is better than a guaranteed $20. The two envelopes will look equally appealing to me when...' Chloe scribbles on the napkin for a bit,[21] and concludes: '...I've got a 7/8ths—or a 87.5% probability for taking X.'

Erica: 'Hmm, I think you're right about how to calculate the equilibrium probabilities here—but, doesn't that seem wrong to you? Don't you think that you should just go ahead and take the guaranteed $20 in Y? If you try to go for X, Newcomb will likely have predicted your choice, and you'll likely end up with nothing!'

Chloe: 'I'm not sure. I mean, if I go for Y, I should think that there's likely $100 waiting for me in envelope X—how could it be rational for me to ignore that information and go ahead and take envelope Y?'

Erica: 'Okay, so, if *that* is your strategy, then you should expect Newcomb to have predicted you to take X...I guess, 80% of the time?' She jots down some equations to check.[22] 'Yeah, 80% of the time. And so Newcomb will predict that you take Y 20% of the time.'

Chloe: 'Yeah...actually, I think that this is just like the game with the doors. Any strategy is going to be as good as any other, in the long run I face when taking X 87.5% of the time, and taking Y 12.5% of the time.'

Erica: 'Hold on, I'm not so sure about that. Could we work it out?'

Chloe: 'Sure. If we calculate the *unpredictable* value of taking X, in the long run faced by my equilibrium strategy, we'll get $20, since...' Chloe writes:

$$\begin{aligned} U(X \mid S_e) &= \Pr(K_X \mid S_e) \cdot V(X K_X) + \Pr(K_Y \mid S_e) \cdot V(X K_Y) \\ &= 80\% \cdot 0 + 20\% \cdot 100 \\ &= 20 \end{aligned}$$

'(I'm using 'K_X' to stand for 'Newcomb predicts I take X', and similarly for 'K_Y'.) And since there's always $20 in envelope Y, no matter what, the unpredictable value of taking Y is also $20.'

Chloe: 'Then, the *unpredictable* expected return of a strategy which takes X with probability x—let's write that 'S_x'—in the long run I face using my equilibrium

[21] She sets $\mathbf{CER}(X)$ equal to $\mathbf{CER}(Y)$ and solves for $\Pr(X)$, getting $\Pr(X) = 7/8$.

[22] She writes: '$\Pr(K_X \mid S_e) = \Pr(K_X \mid X S_e) \cdot \Pr(X \mid S_e) + \Pr(K_X \mid Y S_e) \cdot \Pr(Y \mid S_e) = (9/10) \cdot (7/8) + (1/10) \cdot (1/8) = 4/5$'.

strategy, S_e, will be...' Chloe writes out:

$$\begin{aligned}\mathbf{UER}(S_x \mid S_e) &= \Pr(X \mid S_x) \cdot U(X \mid S_e) + \Pr(Y \mid S_x) \cdot U(Y \mid S_e) \\ &= 20x + 20(1-x) \\ &= 20\end{aligned}$$

'So, in the *unpredictable* long run, it doesn't matter how often you take X and how often you take Y. Your unpredictable expected return will be $20.'

Erica: 'Okay, good. This clarifies things for me—so here's what I'm confused about. Suppose you and I *actually play* this game a large number of times. You take X 7/8ths of the time and Y 1/8ths of the time, and I take Y every time. I get $20 each time, and—as you would expect—you walk away, on average, with about $11.25.[23] At the end of the game, I ask you: 'Why ain'cha rich?' What are you going to say to me?'

Chloe: 'Oh...hmm...well, I'm going to say: 'I'm not rich because...I faced a different long run than you did; but, on my long run, I did as good as any strategy could have done."

Erica: 'But that's just not true, Chloe. You clearly could have done better, *even on your own long run*, if only you'd taken envelope Y every time. What's true is that you did as well as anybody would have done, if Newcomb hadn't been any good at predicting how you'd choose. But, since it *is* good at predicting how you'll choose, and since you know this, I don't understand why your riches in that unpredictable long run, that you know you *definitely won't* face, should tell you anything at all about how to choose in the long run you *definitely will*.'

Chloe: 'Yeah...you're right. I knew that was wrong as I was saying it. Damn. I guess I really shouldn't be thinking about my riches in the unpredictable long run. So...what to say?' Chloe thinks for a bit. 'Honestly, right now I'm feeling tempted by the thought that there is no connection between rational choice and your riches...but that feels drastic.'

Erica: 'You know, you could always just accept *my* proposal.'

Chloe: 'I could...but I'm still pretty convinced that you're making unfair comparisons. I still think that, in the game with the boxes, you squandered the opportunities Newcomb provided you; and that I made the most of those opportunities. And I still think that we should be equalizing opporuntities in some way before we start comparing how much money our strategies make. It's just that the games

[23] This is what you would expect because $\mathbf{ER}(S_e) = \Pr(Y \mid S_e) \cdot 20 + \Pr(X \mid K_Y S_e) \cdot \Pr(K_Y \mid S_e) \cdot 100 = (1/8) \cdot (20) + (7/16) \cdot (1/5) \cdot 100 = 45/4 = 11.25$.

with the doors and the envelopes have left me confused about how to equalize those opportunities in general.'

Chloe gets quiet and thinks. After a while, she says: 'Well...actually...maybe that's wrong. I guess that, throughout this conversation, I've really been saying two different things. The first thing I've been saying is: in the game with the boxes, you squandered the riches Newcomb provided you. The second thing I've been saying is: had I been in your long run, I would have done better than you. All my proposals have been attempting to develop that second idea. I've been trying to find some way to equalize our opportunities by finding a long run in which your and my performance can be fairly compared. But I think that what I've learned is that that was the wrong way to go. What I've learned is that, in the game with the doors, there's no fair way to compare your and my performance in my long run, because we can't both be predictable in my long run. I could try to make things fair by making us *both* unpredictable, but then I'm not comparing *my* average riches and yours, but rather the average riches that would be acquired in my long by someone who chose like me but wasn't predictable. And that doesn't teach me anything about how good *my* strategy is, since, unlike that hypothetical person, I *am* predictable.'

Erica: 'Right.'

Chloe: 'But let me go back to that first idea: in the game with the boxes, you squandered your opportunities, and I didn't. Maybe I should agree with you that we *can* compare your performance in your long run with my performance in mine. But I should disagree that we should be making that comparison on the basis of the total amount of money we end up with. That comparison confuses the riches Newcomb provided us with the riches we ourselves earned with our choices. Instead, we should be comparing our performances in each of our respective long runs by asking about whether we did the best we could in those long runs. In the long run I expect to face two-boxing, I do the best I can; but in the long run you expect to face one-boxing, you do worse than you could have. *That's* why you're irrational. You got more money than I did; but, even so, you got less than you could have. I got less money than you; but, even so, I got as much as I possibly could have.'

Erica: 'Yeah, but, in the game with the black and white doors, you definitely *won't* get as much money as you possibly could in your long run, no matter what you do. So if you say this, won't you be forced to say that neither door is a rational choice?'

Chloe: 'Yeah, right, I don't want to say that, but...maybe I don't have to. What if we think about it like this: in the game with the boxes, one-boxing squanders riches, and two-boxing does not. In the game with the doors, *every* choice will, in the long run, end up squandering some riches—in the sense that, no matter what you choose, you'll face a hypothetical long run in which some other choice would earn you more money. But we can still ask: *how much money* do those choices squander? And if you squander more in your hypothetical long run than I squander in mine, then I'll say that you've acted irrationally.'

Erica: 'Oh, that's really interesting, Chloe. So, in the game with the doors, I guess you'd say that both choosing black and choosing white squander the same amount of riches and so either is permissible. Is that right? How are you thinking about measuring the degree to which riches are squandered?'

Chloe: 'Hmm....I don't know. Let me think about it for a bit.' She scribbles on the napkin for a while, and eventually says: 'Alright: here's an idea. I've *squandered* my riches to the extent that, in the long run I expect to face, the alternative would bring me more riches. If I choose A, then the riches I'll actually get on average in my long run is $\mathbf{CER}(A \mid A)$. And the riches the alternative, $\sim A$, would get me on average in my long run is $\mathbf{CER}(\sim A \mid A)$. So the riches I've squandered by not going for the alternative instead is given by the difference $\mathbf{CER}(\sim A \mid A) - \mathbf{CER}(A \mid A)$. If this difference is positive, then I've squandered riches. If it's negative, then I've not.' Chloe writes out:

$$\mathbf{CL}(A) = \mathbf{CER}(\sim A \mid A) - \mathbf{CER}(A \mid A)$$

'There—'**CL**' for 'Chloe loss'. Your Chloe loss is just the difference between what the alternative would get you, on average, in your long run, and what you actually expect to get, on average, in your long run. If the alternative has a higher expected return than your choice, in your long run, then you have squandered your opportunities, and your Chloe loss will be positive. On the other hand, if the alternative has a lower expected return than your choice, in your long run, then you have not squandered your opportunities for riches, and your Chloe loss will be negative.'[24]

Erica: 'Okay, so maybe *this* is how we should understand our disagreement: I think you've chosen irrationally if the alternative gets more riches in its hypothetical long run than you get in yours. Whereas *you* think that you've chosen irrationally if the alternative squanders fewer opportunities for riches in its long run than you squander in yours.' Erica writes down:

[24] If we confine attention to decisions in which there are only two options to choose between, then Gallow (ms) and Barnett (ms) both say to minimize your Chloe loss. In this special case, the option with minimal Chloe loss will be the option which has the highest probability when your act probabilities are in *equilibrium* (*cf.* the deliberational dynamics of Skyrms (1990), Arntzenius (2008), and Joyce (2012, 2018)).

Chloe's 4th Proposal Choosing A is irrational if the alternative, $\sim A$, is such that

$$\mathbf{CL}(A) > \mathbf{CL}(\sim A)$$

Chloe: 'Yeah, I think that's it. So...I was originally saying that we shouldn't compare my performance in *my* long run with your performance in *yours*. And I guess I'm not saying that anymore. I've realized that there's no way to make fair comparisons between us on a single long run of the game with the doors. (Since, in that long run, there's no way for your strategy to be predictable. If I'm predictable and you're not, that's not a fair comparison. And, if I'm not predictable, then there's no connection between the average riches I get on that merely hypothetical long run and the average riches I would expect to *actually* get in the long run.) So I've decided that we need to compare our performances in our respective long runs—the long runs we each would expect to actually face, were we to make the same choice over and over again. But I've changed my mind about how to evaluate our performances in those long run. I'm not evaluating us by looking at the money we end up with—since that confuses the riches we *earn* with the riches Newcomb *provides*. Instead, I am evaluating us by asking how much we did to actually *earn* those riches.'

Erica: 'Wait, I'm a bit confused by that. You were talking about money *squandered*, but now you're talking about money *earned*.'

Chloe: 'Oh, yeah, maybe that's confusing. Here's what I'm thinking: the money I *earn* in my long run is just the value I've added by choosing A rather than the alternative. In the state K, the value I add by choosing A rather than $\sim A$ is $V(AK) - V(\sim AK)$. So, the value I'll add *on average*, in my long run, is—' Chloe writes:

$$\begin{aligned}
&\sum_K \Pr(K \mid A)(V(AK) - V(\sim AK)) \\
=&\sum_K \Pr(K \mid A)V(AK) \; - \; \sum_K \Pr(K \mid A)V(\sim AK) \\
=&\mathbf{CER}(A \mid A) - \mathbf{CER}(\sim A \mid A)
\end{aligned}$$

'And that's just negative 1 times the Chloe loss of A, $-\mathbf{CL}(A)$. So the money you've *earned* by choosing A is just negative 1 times the money you've *squandered* by choosing A. So, you're irrational if you *earn* less riches than you could otherwise have earned. Or, putting the same point a different way: you're irrational if you *squander* more riches than you had to. So the reason I'm now saying it's irrational to one-box is this: one-boxing is throwing away $10,000 that didn't have to be thrown away. When you one-boxed, you needlessly squandered those riches. And that's irrational.'

Erica: 'But, in the game with the doors, you will end up squandering money too, right?'

Chloe: 'Yeah, I agree. No matter which door I choose, in the long run I expect to face choosing *that* door, choosing the other door would earn me $60 more, on average. So, if I choose to open the black door, then I'm throwing away $60, on average, in the long run. But I don't have any choice *but* to throw that $60 away. *Both* of the available choices squander $60 in their respective long runs. (The Chloe loss of each is $60.) So, while I throw away money in the game with the doors, I don't throw that money away *needlessly*. In the game with the boxes, in contrast, there *was* a choice to not throw the $10,000 away. You could have taken both boxes, thereby *earning* yourself an additional $10,000. That's why I didn't just say that one-boxing squanders money. Squandering money is something for which you can be forgiven, if the squandering is unavoidable. Worse that that: one-boxing is *needlessly* squandering money. It is squandering money when you could have secured $10,000 instead.'

Erica: 'Well, I think that I'm securing myself that $1,000,000 by one-boxing. One-boxing tells me that I'm in a long run which has the $1,000,000 in it nine times out of ten.'

Chloe: 'We're going to have to agree to disagree about that—I don't think merely *reassuring yourself* that the money is there counts as *securing* it in any serious sense. But let's put that debate to the side.'

Erica: 'Okay, fine. So can we see what this proposal says about the game with the envelopes? If I'm understanding your new proposal correctly, taking envelope X is squandering $10—since, in the long run in which you take X each time, you'll make $10 on average, when you could have instead had the guaranteed $20 by taking envelope Y.'

Chloe: 'Yeah, I think that's right. But taking envelope Y squanders $70—since, in the long run in which you take Y each time, you'll make $20 on average, when you could have instead had $100 nine times out of ten, or an average of $90. So, even though taking envelope X squanders money, taking envelope Y squanders *even more* money. So I say it's irrational to take Y.'

Erica: 'But this is silly! In my long run, I'm sitting pretty with an average of $20, whereas you only get an average of $10 in your long run. I still don't see how you've dealt with my original objection: *if you're so smart, why ain'cha rich?*'

Chloe: 'Good, let's get back to that objection. With this new proposal, I think I should give the following reply: that rhetorical question presupposes that rationality is always rewarded with riches, and that, therefore, poverty is a symptom of irrationality. But, on my view, that's just not so. Rational choosers don't prize long run wealth *per se*. Instead, they distinguish between the wealth the world affords them, and the wealth which they themselves *create* with their choices. Only the latter speaks in favor of a choice. So, on my view, it's not wealth but wealth *creation* which is a symptom of rationality. Likewise, it's not poverty but wealth *destruction* which is a symptom of irrationality. Since this is how I think about rational choice, I think that the world can predictably punish rationality, and it can predictably reward irrationality. And that's what I think happens in the game with the envelopes. By choosing X, I end up poorer than you in the long run, but I squander less money than you do. By choosing Y, you end up richer in the long run, but you squander more money than I do. Similarly for the game with the boxes. As a two-boxer, I end up poorer than you in the long run, but I *earn* more money than you do. As a one-boxer, you end up richer in the long run, but you *earn* less money than me. Really, I should turn your rhetorical question back around: *if you're so smart, why'd you lose so much money?*.'

References

Ahmed, Arif. 2014. *Evidence, Decision and Causality*. Cambridge University Press, Cambridge, UK. [3], [5]

Arntzenius, Frank. 2008. "No regrets, or: Edith Piaf revamps decision theory." *Erkenntnis*, vol. 68: 277–297. [3], [11], [12], [16], [21]

Barnett, David James. ms. "Graded Ratifiability." [21]

Egan, Andy. 2007. "Some Counterexamples to Causal Decision Theory." *Philosophical Review*, vol. 116 (1): 93–114. [10], [17]

Gallow, J. Dmitri. ms. "The Causal Decision Theorist's Guide to Managing the News." [21]

Gibbard, Allan & William L. Harper. 1978. "Counterfactuals and Two Kinds of Expected Utility." In *Foundations and Applications of Decision Theory*, A. Hooker, J.J. Leach & E.F. McClennan, editors, 125–162. D. Reidel, Dordrecht. [3], [10]

Greene, Preston. 2018. "Success-First Decision Theories." In *Newcomb's Problem*, Arif Ahmed, editor. Cambridge University Press. [2]

Harper, William. 1986. "Mixed Strategies and Ratifiability in Causal Decision Theory." *Erkenntnis*, vol. 24: 25–36. [11]

Joyce, James M. 1999. *The Foundations of Causal Decision Theory*. Cambridge University Press, Cambridge. [2]

—. 2012. "Regret and instability in causal decision theory." *Synthese*, vol. 187 (1): 123–145. [10], [11], [16], [21]

—. 2018. "Deliberation and Stability in Newcomb Problems and Pseudo-Newcomb Problems." In *Newcomb's Problem*, Arif Ahmed, editor. Oxford University Press, Oxford. [11], [21]

Lewis, David K. 1981a. "Causal Decision Theory." *Australasian Journal of Philosophy*, vol. 59 (1): 5–30. [7]

—. 1981b. "'Why ain'cha rich?'." *Noûs*, vol. 15 (3): 377–380. [2]

Nozick, Robert. 1969. "Newcomb's Problem and Two Principles of Choice." In *Essays in Honor of Carl G. Hempel*, Nicholas Rescher, editor, 114–146. D. Reidel, Dordrecht. [2]

Richter, Reed. 1984. "Rationality Revisited." *Australasian Journal of Philosophy*, vol. 62 (4): 392–403. [10]

Skyrms, Brian. 1990. *The Dynamics of Rational Deliberation*. Harvard University Press, Cambridge, ma. [11], [12], [16], [21]

Soares, Nate & Benjamin Anders Levinstein. ms. "Cheating Death in Damascus." Available at `https://intelligence.org/files/DeathInDamascus.pdf`. [2]

Weirich, Paul. 1985. "Decision Instability." *Australasian Journal of Philosophy*, vol. 63 (4): 465–478. [10]

Wells, Ian. forthcoming. "Equal Opportunity and Newcomb's Problem." *Mind*. [2], [7]