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, Chloe is led to endorse a novel causalist theory of rational choice. 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. In the final round, they are presented with two boxes: one transparent, one opaque. In the transparent box, there is $1000. In the opaque box, there is either $1,000,000 or nothing. They must choose whether to take only the opaque box ('one-box') or both boxes ('two-box'). Before filming, an AI bot named Newcomb analyzed MRI scans to make a prediction about whether they'd one-box or two-box. If it predicted they'd one-box, then $1,000,000 was placed in the opaque box. Else, it was left empty. These predictions are 90% reliable.[1] Erica decides to one-box, and she walks away with a million dollars. Chloe decides to two-box, and she walks away with a thousand. After filming, they share a coffee and discuss their strategies.

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

Erica: 'It was a simple calculation. The expected value of one-boxing was much higher than the expected value of two-boxing. One-boxing was clearly the better bet.'[2]

Final Draft. Forthcoming in the *Australasian Journal of Philosophy*.

† Thanks to R.A. Briggs, Daniel Drucker, Adam Elga, James Shaw, and two anonymous reviewers for helpful conversations and feedback on this material.

1. That is: the probability Newcomb predicted they'd X, given that they X, is 90%.

2. Neither Erica nor Chloe is risk-averse.

Chloe: 'But, no matter what was predicted, you'd be $1000 richer if you'd taken both!'

Erica agrees with Chloe about this, but still thinks she made the right choice. 'If you're so smart', she asks Chloe, 'why ain'cha rich?'

Chloe: 'It's no fair comparing my performance with yours—you had more money in front of you!'

Erica remains unmoved, so Chloe reminds her of an earlier game, just like the game in the final round, but where both boxes were transparent. You could either take just the left box, or take both. If Newcomb predicted you'd take both, then the left box was empty. If it predicted you'd take only the left, then the left box had $1,000,000 in it. Both Erica and Chloe's left box was empty, but Fred's left box had $1,000,000 in it. Fred took just the left box.

Chloe: 'Surely Fred did the wrong thing, leaving the $1000 behind.' Erica agrees.

Chloe: 'But Fred ended up richer than you. So if you're going to use your riches as evidence of rationality, then shouldn't you also see Fred's riches as evidence of his?'[3]

Erica: 'No fair comparing my performance with Fred's—we were making different choices.'

Chloe: 'Wait...why were the choices different?'

Erica: 'If we're going to make the same choice, then we have to have the same options, evidence, probabilities, and desires. Fred and I had different evidence—I knew my left box was empty, whereas he knew his contained the million. So we were making different choices, and the fact that he ended up richer than I did doesn't mean that I was irrational. But you and I were making the same choice. So the fact that I ended up richer than you does mean that you were irrational.'[4]

Chloe: 'But I didn't have to end up with only $1000. I got unlucky, but I could have ended up with $1,001,000. If I had, would that have shown

3. This argument is made by Gibbard and Harper (1978).

4. Erica's views here agree with Ahmed (2014, §7.3).

that you were irrational?'

Erica: 'Well, we shouldn't be thinking about how much you made this one time. We should instead be thinking about how much money you'd have made on average, were you to make this choice over and over again a 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 or not—write 'M' for 'there's a million' and '$\sim M$' for 'there's not'. The probability of M, given that you one-box ('O') is 90%. So, in the long run, I'll make $900,000 on average. On the other hand, you'll get $1000 90% of the time, and $1,001,000 10% of the time. So you'll get $101,000 on average.' Erica writes:

$$\begin{aligned}\mathbf{ER}(T) &= \Pr(M \mid T) \cdot 1,001,000 + \Pr(\sim M \mid T) \cdot 1000 \\ &= 10\% \cdot 1,001,000 + 90\% \cdot 1000 \\ &= 101,000\end{aligned}$$

'So you didn't just get unlucky—it was predictable that you'd get unlucky, given that you chose as you did. Your expected return was lower than mine. You're expected, in the long run, to end up poorer than me. And that's what makes you irrational.'

Chloe: 'Well, wait—couldn't I also get lucky in the long run, and win $1,001,000 on average? How does talking about the long run change anything?'

Erica: 'Sure, it's *possible* that you get lucky in the long run. But I think the long run still helps to clarify our thinking about what it's rational to choose—for, as we play the game more and more times, the probability that our average return matches our expected return approaches 100%.[5]

5. Erica is appealing to the weak law of large numbers; if she were more careful, she'd have said: 'for any $\epsilon > 0$, the probability that my average return and my expected return differ by no more than ϵ approaches 100% as we play the game more and more times.'

Whatever's rational to choose once is rational to choose repeatedly, if you're making the same choice over and over again. So, by imagining ourselves choosing over and over again, a large number of times, we can transform a choice made in conditions of uncertainty into a choice made in conditions of certainty: for we can be certain (or as near to certainty as we like) about what would happen, were we choose an option repeatedly over the long run.'

Chloe: 'Okay, so I see why it's useful to think about your performance in the long run, but it still seems to me that, just like it wasn't fair to compare your and Fred's riches, it's not fair to compare your and my riches—even in the long run. You end up with $1,000,000 in front of you more often on your long run than I do on mine. But that doesn't show that you're choosing more rationally than me—it just shows that you're given more wealth on your long run than I'm given on mine.'[6]

Erica: 'Okay, but then what comparison is fair? I mean, making money is the whole point of this game. Surely you think that there's some connection between playing rationally and the money you get—right?'

Chloe: 'I suppose I could deny that there's any connection between rational choice and the goods you expect to end up with...but that seems drastic, and I'd rather not say that. I'm inclined to think that there's some connection. But I don't think the connection is what you say it is.'

Erica: 'Okay, so what is it, then? I've got my answer: you're choosing rationally iff no one else facing the same choice would be expected to make more money than you, over the long run. In general, that's how I decide what to do. When I face a decision, I list off all the options, $A_1, A_2, ..., A_N$, and I list off all the ways things might be, for all I have any control over, $K_1, K_2, ..., K_N$.[7] For each option, A, I ask myself, firstly, how much I'd like to choose A in each state K, and, secondly, how likely each state 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

6. This reply is offered by Lewis (1981b) and Joyce (1999), among others.

7. Erica forgets to note: these K_i's are what Lewis (1981a) calls 'dependency hypotheses', and they are mutually exclusive and jointly exhaustive. (By the way, Erica doesn't think she has to calculate her expected return using these kinds of states—any other set of mutually exclusive and jointly exhaustive states would do just as well. Even so, this is how she does it.)

napkin:

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

'$\Pr(K \mid A)$ is the probability of K, given A. And $V(AK)$ is how much I'd like to choose A in the state K. (Since I only care about money, $V(AK)$ is how much money I expect to have, if I choose A in the state K.) I choose whichever option has the highest expected return. So long as you and I are making the same choice, the expected return of each option will be the same for you as it is for me. So I'll always end up making at least as much money as you, on average, over the long run.'

Chloe: 'I see.'

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

Chloe thinks for a while, 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. So we need to equalize those opportunities. Maybe we should put it like this—' Chloe writes down:[8]

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

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

Erica: 'Hold on. The probability that you're in state K may depend upon how likely you are to choose each option. During the game with the boxes, the probability that there's a million dollars changes, depending

8. Skyrms (1980) evaluates options with CER. Other versions of causal decision theory use similar expectations. See Lewis (1981a).

upon how likely you are to two-box. As the probability of you two-boxing rises, it gets less likely that the million is there. And as the probability of you two-boxing falls, it gets more likely that it's there. So which unconditional probability do you want to use?'

Chloe: 'Let's try using the probability once I've chosen. If you want to show me that I'm irrational, then you should show me that your way of choosing would make me richer than mine, over the long run I'd expect to face when I make the choice.'

Erica: 'But isn't that just what I said? You're now telling me to calculate your 'Chloe' expected return, conditional on you having chosen, and that's this—' Erica writes:

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

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

Chloe: 'Well...yes and no. I think we agree about how to understand expected returns. What we disagree about is how to *compare* choices in terms of their expected returns. You say that it's fair to compare the amount I expect to make when I choose to two-box, $\mathbf{CER}(T \mid T)$, with the amount you expect to make when you one-box, $\mathbf{CER}(O \mid O)$. But I think it's not always fair to compare people in terms of how much they expect to make when they choose, since it could be that one person expects to have been provided with more wealth than the other does. If you one-box and I two-box, then your hypothetical long run will look better than mine, not because you'll be choosing more rationally than me, but just because there will be more money for the taking on your long run than there is on mine. So I think we need to equalize the opportunities for wealth, 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 \$1000 richer than you. So you don't have a higher expected return than I do, once we've equalized the opportunities.'

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

Chloe: 'Yeah. Here, I just mean this: for any options—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 consider the long run you'd expect to face when you select B, and then we ask about how good it would be, in that long run, to have chosen A instead. Maybe we can think about it like this.' Chloe writes down:

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

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

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

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

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

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

Erica: 'Sure, that's fine with me.'[9]

Chloe: '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'd expect to face, making the same choice over and over again. What's wrong with that?'

Erica thinks for a bit, and then says: 'Do you remember that game we played with the two doors?'

Chloe: 'Yeah, that one was strange.'

9. See Gallow (ms) for discussion of additional complications in choices with three or more options.

In the door game, they had to choose between a black door and a white door. If Newcomb predicted they'd take black, $100 was left behind the white door. If Newcomb predicted they'd take white, $100 was left behind the black door. These predictions are 80% reliable.

Erica: 'So I think your proposal says that both white and black are irrational choices in the door game. The Chloe-expected return of black, in the long run you'd expect to face choosing black, is $20, which is less than the Chloe-expected return of white, which is $80.' She writes down:

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

('K_W' says the money's behind the white door, and 'K_B' says it's behind the black door.) So your proposal says black is irrational. And it says the same about white, since the Chloe-expected return of white, in the long run where you choose white, is $20, whereas the Chloe-expected return of black in that same long run is $80. (Black and white are perfectly symmetric.)'

Chloe: 'Shoot. I suppose I could say there's no rational choice here[10]...but I'd rather not. I'm not really sure what to say about the door game, honestly, but surely there's some rational choice to be made.'

Erica: 'So then do you concede that my theory about the connection between rationality and riches makes more 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 we're afforded. What I think you've shown me is just that I had the wrong way of equalizing those opportunities.'

10. Harper (1986, p. 33) says this.

Erica: 'What's the alternative?'

Chloe thinks for a while, 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. When I found myself leaning towards black, white seemed like a better choice—after all, if I'm going to choose black, then Newcomb likely predicted I would, so the money's likely behind white. But then, if I inclined towards white, black appeared the better choice—if I'm likely to take white, then the money's likely behind black. After this vacillating, I ended up stuck in the middle, and I didn't feel like I had anything more to think about, so I just...*picked*. I went with black, but I could have gone for white instead.'[11]

Erica: 'You're a mystery to me, Chloe. What's 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: 'I'm not sure—if I'm more likely to choose black, then the money's more likely behind white, and that may be an important consideration...it breaks the symmetry, in any case. But put that aside. My vacillation gives me an idea about which hypothetical long run equalizes the opportunities. As I was deliberating about what to do, my probability that I would choose black was changing. And when I was just as likely to choose black as not, both doors looked equally good to me. Let's say that my probability distribution was in *equilibrium*. So here's a new idea: we should compare people by looking at how much they earn, on average, in a long run corresponding to this equilibrium probability. If the equilibrium probability for the game with the boxes has me 100% likely to two-box, then this will agree with what I said earlier, but maybe it will let me avoid this trouble you're pointing out with the door game. I guess I'd have to think about how to find this equilibrium probability in general—and I hope that there always is some equilibrium probability—but that's the start of an answer.'[12]

Erica: 'Alright, so is this your proposal? Choosing A is irrational iff the

11. This deliberative vacillation is recommended by Skyrms (1990) and Joyce (2012). See also Arntzenius (2008).

12. Chloe needn't worry—there will always be an equilibrium. See Skyrms (1990) and Arntzenius (2008). (There could be more than one equilibrium. Chloe should worry about that, but it doesn't occur to her.)

alternative has a higher average return in the long run you'd face if you chose...well, wait, what is it?'

Chloe: 'Maybe we shouldn't say that an *act* is rational or irrational. Let's say instead that a *method for selecting acts* is rational or irrational—let's call a method like that a *strategy*. And a strategy needn't tell you to select one act; it might tell you to select each act with a certain probability. So here's the proposal:'

Chloe's 2nd Proposal: A strategy, S, is irrational iff there's another strategy, S^*, with a higher expected return than S in the long run faced by S.

$$\mathbf{CER}(S^* \mid S) > \mathbf{CER}(S \mid S)$$
$$\sum_K \Pr(K \mid S) \cdot V(S^*K) > \Pr(K \mid S) \cdot V(SK)$$

Erica: 'That's helpful, Chloe. So let's think about the strategy of picking black half the time and white half the time. If that's your strategy, then, in the long run, you'd expect the money to be behind white half the time, and the money to be behind black half the time.'

Chloe: 'Right. And then, when I'm evaluating a strategy which says to choose black with some probability—call it 'b'—I'll calculate its expected return, in my long run, like this:'

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

'(Here, 'S_b' is the strategy of choosing black with probability b, and S_e is the *equilibrium* strategy of choosing black with 50% probability.)'

Erica: 'I see—so your Chloe-expected return doesn't depend upon your probability for black at all! In this long run, any strategy is as good as any other. So I guess your new proposal doesn't say that choosing black with 50% probability is irrational.'

Chloe: 'No, I don't think so. But I think it does say choosing black with 100% probability is irrational. If you're sure to choose black, then in

the long run, the money will be behind white 80% of the time. So you'll expect to make $20 on average; whereas, in that long run, the alternative strategy of choosing white every time would make $80, on average. And I guess the same would be true for any strategy other than the 50/50 split. If your probability for black is greater than 50%, then just taking white every time would get you a higher return, on average; if it's less than 50%, then just taking black every time would get you a higher return, on average. So the equilibrium strategy is the only rational strategy.'

Erica: 'I see...that's cool, Chloe.' Erica scribbles on the napkin for a bit, then says: 'Well, wait. I'm a bit confused. You said that, in the long run you expect to face playing the equilibrium strategy, your strategy (like every other strategy) has an average return of $50. But, as I calculate it, your average return is $20.'

Chloe: 'No, I don't think so. Here's my expected return—'

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

'And the value of the equilibrium strategy when the money's behind the black door is...'

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

'...and similarly, the value of S_e when the money's behind the white door is $50. So my expected return is $50.'

Erica: 'No, I think you're miscalculating the value of your strategy when the money's behind the black door. You're ignoring important correlations between which door you end up choosing with the strategy and where the money awaits. We agree that, in your long run, the money's behind black 50% of the time, and it's behind white 50% of the time. But you're assuming that the door you end up choosing is independent of where the money is. Here—' Erica draws two squares like these.

'You're assuming the distribution on the left—you're assuming that you'll pick black 50% of the time, independent of where the money is. But if that were so, Newcomb wouldn't be a reliable predictor. You *should* be assuming the distribution on the right. When the money's behind black, you'll pick black only 20% of the time. So the value of your strategy, when the money's behind black, is actually this:'

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

'Likewise, the value of your strategy, when the money's behind white, is \$20. So your expected return is \$20, not \$50.'

Chloe: 'Oh...well, couldn't Newcomb just be good at predicting which strategy I adopt? After that, whether I pick black or white is just a roll of the dice, and surely Newcomb can't predict how those dice land.'

Erica: 'I don't know how Newcomb works, but I know this: given that you pick black, it's 80% likely to have predicted that you'd pick black. That's what we were told, and that's all I'm assuming. And, you know, we're not talking about actual dice here (remember, the producers were very insistent that we couldn't use coin flips or dice rolls). So whatever corresponds to 'rolling the dice' is something up in your brain. And Newcomb saw scans of our brains. So maybe it is able to predict how the dice will hand. Anyhow, let's suppose it can, since that's the interesting case.'

Chloe: 'Okay, I guess that's right. So, if Newcomb can predict what I pick, then the equilibrium strategy will have an expected return of \$20...so what's going on here?' She thinks, and then says: 'I see...when you calculated my expected return, you took for granted that Newcomb could predict what I'd choose—you held fixed the probability that the money was behind black, given that I choose black, $\Pr(K_B \mid B)$, and the

probability that it's behind black, given that I choose white, $\Pr(K_B \mid W)$. But, when I was calculating the expected returns of the other strategies on my long run, I wasn't holding those conditional probabilities fixed. In fact, I *couldn't* have held them fixed. Since I'm asking about *my* long run, I'm holding fixed the probability that the money's behind black, $\Pr(K_B)$. But this, together with $\Pr(K_B \mid B)$ and $\Pr(K_B \mid W)$, determines how likely I am to choose black,[13] so it determines what my strategy is. So considering how any *other* strategy fares in *my* long run requires us to suppose that Newcomb isn't 80% reliable at predicting that strategy.'

Erica: 'Yeah, that seems right—so doesn't this show that the whole exercise of trying to 'equalize the opportunities' and see how someone else would fare in your long run is confused from the get-go? 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, then both should be predictable if either is. And that's just what my proposal accomplishes.'

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

Erica: 'Is that a new proposal?'

Chloe: 'I think so. How to put it? Let's not ask about the Chloe-expected return of my strategy—that's tantamount to assuming that I'm predictable. Instead, let's ask about what we can call its *unpredictable* expected return.' She writes:[14]

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

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

14. Chloe assumes that, for any proposition ϕ and any partition $\{Z_1, Z_2, \ldots, Z_N\}$, $V(\phi) = \sum_i \Pr(Z_i \mid \phi) \cdot V(\phi Z_i)$. That's why she allows herself to replace '$V(S^*K)$' in $\mathbf{CER}(S^* S)$ with $\sum_A \Pr(A \mid S^*K) \cdot V(AS^*K)$. Chloe doesn't intrinsically value strategies, so $V(AS^*K) = V(AK)$.

'The difference is that, when we calculate your Chloe-expected return, we pay attention to correlations between the act you choose, A, and the state of the world, K; but, when we calculate your *unpredictable* expected return, we ignore those correlations. So the new proposal is:'[15]

Chloe's 3rd Proposal: A strategy, S, is irrational iff there's another strategy, S^*, with a higher unpredictable expected return than S in the long run faced by S.

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

Erica thinks about this for a while, then says: 'I'm a bit confused by this proposal, Chloe. I was asking you what connection you saw between rational choice and long-run riches. Now, you're telling me the rational choice will make you richer, not in the long run you'd actually expect to face, making the choice over and over again, but rather in a different long run, one you wouldn't expect to face, and in which you're not predictable?'

Chloe: 'I'm just making the comparison fair. The other strategies can't be predictable in my long run, so I shouldn't be either.'

Erica: 'I guess, but this proposal still just seems wrong to me. You're pretending you're not predictable, when you know that you are. Maybe it would help if we could think about that game with the envelopes.'

In the envelope game, there were two envelopes, labeled 'X' and 'Y'. They had to choose one, and only one, of the envelopes. There was a guaranteed \$20 in Y. If Newcomb predicted they'd take Y, then \$100 was placed in X. If it predicted they'd take X, X was left empty. Newcomb's predictions in this game are 90% reliable.[16]

Chloe: 'Yeah, let's think that game through. If I'm sure to take X, then X is only 10% likely to have \$100 in it. Since that's worse than a guaranteed \$20, Y will be a more attractive option, if I'm sure to take X. And, if I'm

15. In a choice between two options, a strategy will be rational according to **Chloe's 3rd Proposal** iff that strategy is an *equilibrium*, in the sense of Skyrms (1990)'s deliberational causal decision theory. This theory is also endorsed by Arntzenius (2008) and Joyce (2012).

16. Similar decisions are discussed in Egan (2007).

sure to take Y, then X is 90% likely to have \$100 in it, which is better than a guaranteed \$20. So X will be a more attractive option, if I'm sure to take Y. Both options will be equally attractive when...' Chloe scribbles on the napkin,[17] and concludes: '...I'm 7/8ths likely to take X.'

Erica: 'Okay, so, if that's your strategy, then you should expect the money to be in X...I guess, 20% of the time?' She jots down some equations to check.[18] 'Yeah, 20% of the time.'

Chloe: 'Yeah, right. Actually, I think this is just like the game with the doors. In the long run I'll expect to face playing my 7/8ths strategy, the *unpredictable* expected return of every strategy is going to be as good as every other, since...' She writes out:

$$\begin{aligned}\mathbf{UER}(S_x \mid S_e) &= \Pr(K_X \mid S_e)\cdot V(K_X S_x) + \Pr(\sim K_X \mid S_e)\cdot V(\sim K_X S_x)\\ &= 20\%\cdot(100x + 20(1-x)) + 80\%\cdot(0x + 20(1-x))\\ &= 20\%\cdot(20 + 80x) + 80\%\cdot(20 - 20x)\\ &= 20\end{aligned}$$

(Chloe uses 'S_x' for the strategy of taking X with probability x, and she uses 'S_e' for the equilibrium strategy of taking X with probability 7/8ths. 'K_X' says that the \$100 is in X.)

Erica: 'Okay, this clarifies things for me. So here's what I'm confused about. Suppose that you and I actually play this game a large number of times. You take X 7/8ths of the time, and I take Y every time. I get \$20 every time, and you walk away with an average of \$11.25, as you'd expect.[19] At the end of the game, I ask you '*Why ain'cha rich?*' What are you going to say to me?'

Chloe: '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 well as anybody could have.'

Erica: 'But that's just not true, Chloe. You clearly could have done better,

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

18. She writes: $\Pr(K_X \mid S_e) = \Pr(K_X \mid X S_e)\cdot\Pr(X S_e) + \Pr(K_X \mid Y S_e)\cdot\Pr(Y \mid S_e) = 10\%\cdot(7/8) + 90\%\cdot(1/8) = 1/5$.

19. This is what you'd expect because $\mathbf{ER}(S_e) = \Pr(Y S_e)\cdot 20 + \Pr(X K_X S_e)\cdot\Pr(K_X S_e)\cdot 100 = (1/8)\cdot 20 + (7/16)\cdot(1/5)\cdot 100 = 11.25$.

even on your own long run, if only you'd taken *Y* every time. What's true is that you *would* have done as well as anyone could have, if Newcomb hadn't been any good at predicting your choice. But since it *is* good at predicting your choice, and you know this, I don't see why your riches in that unpredictable long run, that you *wouldn't* expect to face, should tell you anything about how to choose in the long run that you *would* expect to face.'

Chloe: 'Yeah, that's 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?' She gets quiet and thinks. After a while, she says: 'I guess I've been trying to find some way to compare us after equalizing the opportunities—so I've been looking for some single long run in which to compare our performances. But I think what I've learned is that that was the wrong way to go. In the game with the doors, there's no fair way to compare our performance in my long run, since it's impossible for you to be predictable in my long run. I tried to make the comparison fair by making us *both* unpredictable, but then I'm not comparing *my* long-run riches with yours, but rather the long-run riches I *would* have had, *were* I unpredictable. But that doesn't tell me anything about *my* long-run riches, since I *am* predictable.'

Erica: 'Right.'

Chloe: 'But there's another thought I had, with doesn't have anything to do with equalizing opportunities, and which still seems right to me: in the game with the boxes, you squandered your riches, needlessly throwing away $1,000. So maybe I should change tack, and agree with you that we can compare your performance in your long run with my performance in mine. But perhaps I should just evaluate those performances differently. I shouldn't evaluate our long-run performances by asking how much money we end up with, on average. That confuses the riches which were provided to us, through no effort of our own, and the riches which came to us *as a consequence of our choices*. Instead, we should evaluate our performance by asking how much wealth our choices brought us. In the long run I expect to face two-boxing, my choices win me an additional $1000, on average. Whereas, in the long run you expect to face one-boxing, your choices lose you $1000, on average. And that's why I think you're irrational. You end up richer than I do, but that's just because, before you made your choices, you inher-

ited millions from Newcomb's favorable prediction. Though gifted with a fortunate inheritance, your choices did nothing to increase your fortune; instead, they lost you $1000, on average.'

Erica: 'I'm not sure I understand. Could you explain how this helps with the game with the doors? In that game, won't your choices lose you money in the long run, in whatever sense my one-boxing lost me money?'

Chloe: 'Right, I don't want to say that both choices are irrational in that game. But maybe I don't have to. Maybe we should think about it like this: in the game with the doors, both choices will, in the long run, squander some riches—in the sense that, no matter what you choose, you'll expect a hypothetical long run in which the alternative would bring you more money. But we can still compare us by asking *how much* money we squander in the long run. And if you squander more money on your long run than I do on mine, I'll say that you've chosen irrationally.'

Erica: 'That's interesting, Chloe. So, in the game with the doors, I guess that both black and white squander the same amount of money in the long run, so either is permissible. Is that right? How should I measure the money squandered, on average, in the long run?'

Chloe: 'Let me think about it for a bit.' She scribbles on the napkin for a while, then 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 would expect to get, on average, is $\mathbf{CER}(A \mid A)$. And the riches I'd expect the alternative to bring me, on average, 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 my choice has squandered riches. If it's negative, then my choice has brought me riches.' Chloe writes:

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

'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 your choice gets you, on average, in your long run. If you expect the alternative would get you more money than your choice, then you're squandering your riches, and your Chloe loss is positive. If you

expect your choice to get you more money than the alternative would, then you're not squandering your riches, and your Chloe loss is negative.'

Erica: 'Okay, so maybe this is how we should understand our disagreement: I think you've chosen irrationally if the alternative has more riches in its long run than you have in yours. Whereas you think that you've chosen irrationally if the alternative squanders less riches in its long run than you do in yours.' Erica writes:[20]

Chloe's 4th Proposal: Choosing A is irrational iff the alternative, $\sim A$, has a lower Chloe loss.

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

Chloe: 'Yeah, I think that's it. So what we disagree about is how to evaluate our performances in our respective long runs. I don't look at the money we end up with, since that confuses the money we *earn* with the money Newcomb *provides*. Instead, I just look at how much money we *earn* in the long run.'

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

Chloe: 'Here's what I'm thinking: the money my choice *earns* me is just the value I add by choosing A instead of $\sim A$. In the state K, the value I add by choosing A instead of $\sim A$ is $V(AK) - V(\sim AK)$. So the value I'll add *on average*, in the long run, is:'

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

'And that's just -1 times the Chloe loss of A,$-\mathbf{CL}(A)$. So the money I

20. Given a choice between two options, both Gallow (ms) and Barnett (ms) say to minimize your Chloe loss. Due to complications which Erica and Chloe don't have time to discuss, Gallow (ms) accepts a slightly different theory for choices between three or more options.

say your choice has *earned* you is just negative 1 times the money I say your choice has *squandered*. So when I said you are irrational iff you *squander* more money than you need to, over the long run, I might just as well have said that you are irrational iff you *earn* less money than you could have, over the long run. The reason I think two-boxing is rational is that it earns you $1000, whereas one-boxing needlessly throws $1000 away.'

Erica: 'Okay, I think I understand how you're using the term 'earn'—but why should I think that how much money you 'earn', in your sense, is relevant to rational choice? Why shouldn't we just look at the total amount of money you have at the end of the day?'

Chloe: 'Well, just because Trump has a lot of money, that doesn't make me think that he's a wise investor—after all, he just inherited most of that wealth from his father. When we think about whether Trump's investments were rational or not, we're not just interested in the total size of his fortune. We also want to know how his choices contributed to that fortune—what the fortune could have been, had he chosen differently. And I'm just trying to apply those same standards here.'

Erica: 'I guess...but why can't we just look at the amount by which our fortunes *change* as a result of our choices? My fortune was raised by $1,000,000—yours was only raised by $1000. Why isn't that enough to say that I earned more than you did?'

Chloe: 'Here's another game we could have played: on the basis of its prediction, Newcomb deposits the contents of the opaque box into our bank accounts the day before we make our choice (though we aren't allowed to check our accounts). Then: we have to choose whether to take the transparent box with the $1000 or not. In all relevant respects, this game is just like the one we actually played—right?'

Erica: 'Yeah, I think so. I'd also want to leave the transparent box behind in that game.'

Chloe: 'Okay, but in that game, before making your choice, your fortune would *already include* the $1,000,000, so leaving the transparent box behind wouldn't change your fortune at all. On the other hand, taking it raises my fortune by $1000. So if we think about 'earnings' in terms of how your overall fortune changes in the way you suggested, then I would

earn more than you in this alternative version of the game. I think my choice did earn me $1000, but I don't think we should treat these two versions of the game differently. We should say that I earned the same amount of money in both of these games. So, to see how much money we earned, I don't think we should compare our fortunes *before* and *after* we choose. Instead, I think we should compare our *actual* fortunes with the fortunes we *would* have had, had we chosen the alternative. Since two-boxing left me with $1000 more than I'd have had one-boxing, I think I earned $1000. And since one-boxing left you with $1000 less than you'd have had two-boxing, I think that you squandered $1000.'

Erica: 'But, in the game with the doors, you're going to squander money, too, right?'

Chloe: 'Yeah, that's right. No matter which door I choose, in the long run I'll 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'll throw away $60, on average, over the long run. But I didn't have any choice *but* to throw $60 away. In that game, both choices squander $60. So, while I throw away money, I don't *needlessly* throw the money away. In contrast, when you one-box, you *did* have a choice about whether to throw the $1000 away. You could have taken both boxes, and earned yourself an additional $1000. That's why I didn't just say that one-boxing squanders money. Squandering money is a sin for which you can be forgiven, if the squandering is unavoidable. Worse than that, one-boxing *needlessly* squanders money.'

Erica: 'Can we see what this proposal says about the game with the envelopes? If I'm understanding, then taking X squanders $10—since, if you take X, you'll expect to make $10 on average, in the long run, which is $10 less than what you could have had by taking Y.'

Chloe: 'Yeah, that's right. But taking Y squanders $70—since, if you take Y, you'll certainly get $20, but you'll expect that taking X instead would get you $90 on average, over the long run. So both options here squander some money, but since X squanders less money, I say that you're required to take X.'

Erica: 'But this is silly, Chloe! In my long run, I'm sitting pretty with $20, while you're only getting $10, on average. 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. With this new proposal, I think I should respond like this: your rhetorical question presupposes that rationality is always rewarded with riches, and that, therefore, my poverty is a symptom of my irrationality. But on my view, that's not so. Rational choosers don't prize long-run wealth *per se*. Instead, they distinguish between the wealth the world brings them unbidden and the wealth they bring about with their choices. Only the latter speaks in favor of a choice. So, on my view, it's not long-run wealth but long-run wealth *creation* which is symptomatic of rationality. Likewise, it's not long-run poverty but long-run wealth *destruction* which is symptomatic of irrationality. Since this is how I think about rational choice, I think the world can punish rationality with poverty and reward irrationality with riches. And that's what I think happens in the game with the envelopes. Because I choose X, I expect to face a long run in which there is, on average, at most $20 for the taking (there's only $10 on average in X, and $20 in Y). Because you choose Y, you expect to face a long run in which there is, on average, $90 for the taking (since there's an average of $90 in X). When I think about whether your choices were rational, I don't care how much money the world bequeathed you—I only care about what you *did* with those riches. And you squandered $70 of the wealth you were bequeathed, whereas I only squandered $10 of mine. It's the same as in the game with the boxes. As a two-boxer, I expect to face a long run which has, on average, $101,000 for the taking. And I take it all. As a one-boxer, you expect to face a long run which has, on average, $901,000 for the taking. But you deliberately leave some of those riches behind. You end up richer than me, but even so, you *squander* more money that I do. Really, I should turn your rhetorical question back around: *if you're so smart, why'd you lose so much money?*'

Erica: 'I lost a thousand dollars, but I did that in order to get the million.'

Chloe: 'I don't understand that, Erica. The million was there for you whether or not you lost the thousand. Losing the thousand didn't get you anything.'

Erica: 'Look, there's a reason that my long run is so much richer than yours—it's because I left the thousand behind. Two-boxing gives you a long run which has the million only one time out of ten. One-boxing

gives me a long run which has the million nine times out of ten. Surely getting that kind of long run is worth $1000.'

Chloe: 'Let's think about this hypothetical long run. We're playing the game with the boxes over and over again. I assume that how you choose the first time we play doesn't causally affect what Newcomb predicts the second time. In particular, one-boxing the first time doesn't cause Newcomb to predict that you'll one-box the second time. If it did, I'd rethink two-boxing. Also, this wouldn't be a long run in which we're making the same choice that we actually made—since, if your choice causes Newcomb to predict differently the second time, then the causal consequences of your choice would be different.'

Erica: 'Okay, I guess that's right. How I choose the first time doesn't causally affect what prediction gets made about the second time.'

Chloe: 'So then Newcomb could have made all its predictions about how you'll choose each time at the very beginning—right?'

Erica: 'I suppose it could have.'

Chloe: 'But then, the first time we play, it's already determined how much money's going to be sitting in front of you in the long run. Nothing you do can change that. So, in this hypothetical, I don't think it's true that losing the thousands got you a long run that had the million nine times out of ten. You already had that long run before you made your choice.'

Erica: 'Okay, I guess that's technically correct. But I still think that, in the sense that's relevant here, the fact that I had a long run filled with millions is *attributable* to my leaving behind the thousands—I should get *credit* for that long run.'

Chloe: 'I have such a hard time understanding that, Erica. Newcomb left you those millions before you made any choices—you don't think those millions are *causally* attributable to you leaving behind the thousands, do you?'

Erica: 'No, I agree with you that leaving the thousands behind didn't *cause* Newcomb to give me the millions. Even so, leaving behind the thousands is evidence that the millions are there, and that's enough to attribute the millions to my leaving behind the thousands, in the relevant

sense.'

Chloe: 'Wow, this seems like a whole can of worms. We'll have to get into it some other time. I don't suppose I've persuaded you of anything today?'

Erica: 'No, I'm afraid not. I'm still perfectly happy with my original proposal.'

Chloe: 'Well, I'm sorry about that. But thanks for helping me figure out what to think about the relationship between rational choice and long-run wealth. And thanks for paying for my coffee.'

Erica: 'Don't mention it—I know how poor all those rational choices have left you.'

References

Ahmed, Arif. 2014. *Evidence, Decision, and Causality*. Cambridge: Cambridge University Press.

Arntzenius, Frank. 2008. "No regrets, or: Edith Piaf revamps decision theory". *Erkenntnis*. 68(2): 277-297.

Barnett, David James. ms. "Graded Ratifiability".

Egan, Andy. 2007. "Some Counterexamples to Causal Decision Theory". *Philosophical Review*. 116(1): 93-114.

Gallow, J. Dmitri. ms. "Manage the Improvement News".

Gibbard, Allan and William L. Harper. 1978. "Counterfactuals and Two Kinds of Expected Utility", in *Foundations and Applications of Decision Theory*, ed. A. Hooker, J.J. Leach, and E.F. McClennan. Dordrecht: D. Reidel: 125-162.

Harper, William. 1986. "Mixed Strategies and Ratifiability in Causal Decision Theory". *Erkenntnis*. 24(1): 25-36.

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

Joyce, James M. 2012. "Regret and instability in causal decision theory".

Synthese 187(1): 123-145.

Lewis, David K. 1981a. "Causal Decision Theory". *Australasian Journal of Philosophy*. 59(1): 5-30.

Lewis, David K. 1981b. "Why ain'cha rich?" *Noûs*. 15(3): 377-380.

Skyrms, Brian. 1982. "Causal Decision Theory". *The Journal of Philosophy*. 79(11): 695-711.

Skyrms, Brian. 1990. *The Dynamics of Rational Deliberation*. Cambridge: Harvard University Press.