# Arntzenius on 'Why ain'cha rich?'

## Abstract

The best-known argument for Evidential Decision Theory (EDT) is the 'Why ain'cha rich?' challenge to rival Causal Decision Theory (CDT). The basis for this challenge is that in Newcomb-like situations, acts that conform to EDT may be known in advance to have the better return than acts that conform to CDT. Frank Arntzenius has recently proposed an ingenious counter argument, based on an example in which, he claims, it is predictable in advance that acts that conform to EDT will do less well than acts that conform to CDT. We raise two objections to Arntzenius's example. We argue, first, that the example is subtly incoherent, in a way that undermines its effectiveness against EDT; and, second, that the example relies on calculating the average return over an inappropriate population of acts.

## *1: Introduction*

On the standard reading of the standard version of Newcomb's problem[1] the relative *efficacy* of your options diverges from their *news value*: taking the transparent box *makes* you richer than not taking it; but people who don't take it typically *end up* richer than those who do. Accordingly Causal Decision Theory or CDT (which values efficacy) and Evidential Decision Theory or EDT (which values news value) make different recommendations: CDT says that you should take the transparent box whereas EDT says that you shouldn't. Many philosophers find grounds in this disparity for declaring against EDT (Gibbard and Harper 1981: 180-184; Lewis 1981a: 377-8; Joyce 1999: 146-54).

[1] On this standard version (Nozick 1970) you have the choice between (i) taking just an opaque box and (ii) taking the opaque box plus a transparent box containing $1,000. You get to keep the contents of whichever box or boxes you take. Yesterday a very powerful predictor of human actions (who does not however 'see' into the future in any way that involves backwards causation) put $1M into the opaque box if and only if it predicted that you would, now, take only the opaque box. Should you (i) 'one-box' or (ii) 'two-box'?

This paper concerns an argument that it is the *causalist* who has got things wrong. Frank Arntzenius states it as follows (2008: 289):

> In a Newcomb type case evidential decision theorists will, on average, end up richer than causal decision theorists. Moreover, it is not as if this is a surprise: evidential and causal decision theorists can foresee that this will happen. Given also that it is axiomatic that money, or utility, is what is strived for in these cases, it seems hard to maintain that causal decision theorists are rational.

The key premise of this argument is that evidential decision theorists will be richer on average than causal decision theorists. That is not quite the best way to put it: disputes between CDT and EDT are not about the relative welfare of *theorists* who champion those theories. They are about the relative return to the *acts* that those theories recommend, whether the actor in question is himself a self-conscious causalist, a self-conscious evidentialist, or—like the vast majority of people to whom decision theoretic recommendations should also apply—someone who has never heard of either.

So the key premise is better put like this: the *act* that EDT recommends in a Newcomb type situation—namely, one-boxing—has a better average return than the act that CDT recommends there—namely, two-boxing. Making this amendment and affixing Lewis's (1981b) title for it we have the following argument:

*Why ain'cha rich*

(1) The average return to one-boxing exceeds that to two-boxing (*premise*)

(2) Everyone can see that (1) is true (*premise*)

(3) Therefore one-boxing foreseeably does better than two-boxing (*by* 1, 2)[2] [3]

(4) Therefore CDT is committed to the foreseeably worse option for anyone facing Newcomb's problem (*by* 3)

So understood it is easy to see that the key premise (1) is true. Let the predictor get it right 95% of the time. That is: he predicts that a player will one-box (and so puts \$1M in the opaque box) on 95% of occasions when that player one-boxes. And he predicts that a player will two-box (and so puts nothing in the opaque box) on 95% of occasions when that player two-boxes. Then assuming linear utility for money and writing M for a million and k for a thousand, the average returns (AR) to one-boxing and two-boxing over many trials are:

(5) AR (One-boxing) = 95%. M + 5%. 0 = 950k

(6) AR (Two-boxing) = 5%. (M + k) + 95%. k = 51k

So clearly (1) is true and everyone can see that. So CDT recommends an act that returns *foreseeably* less than what EDT recommends.

It is no use the causalist's whining that foreseeably, Newcomb problems do in fact reward *irrationality*, or rather CDT-irrationality. The point of the argument is that if

[2] Here and elsewhere expressions like *by* and *from* are not intended to indicate that the steps that they label are in all cases *deductively* valid. It is enough that they indicate that the step is supposed to be rationally compelling: for instance, it is our view that anyone who accepts (1) and (2) is rationally compelled to accept (3). This rational compulsion may however lapse in the presence of some defeater; indeed in our view that is precisely what happens in the case that Arntzenius describes.

[3] Of course there is *a* sense in which compatibly with (1) and (2) one-boxing does *not* foreseeably do better than two-boxing. One-boxing does foreseeably worse than two-boxing in the sense that on *any* particular encounter with a Newcomb problem, a one-boxer *would* have done better to have taken both boxes. In this 'counterfactual' sense of 'foreseeably better', two-boxing is foreseeably the better option.

So distinguish that *counterfactual* sense of 'foreseeably better' from the sense in which it means: does in fact have the greater expected *actual* return. In that second sense—the one that we intend—all parties will agree that or one-boxing does foreseeably better than two-boxing given that the predictor is foreseeably accurate. What is at issue between Arntzenius and us is not *that* point, but whether anything follows *from* that point about the superiority of EDT as a normative theory of rational choice. We say yes: Arntzenius says no. (Thanks to a referee.)

everyone knows that the CDT-irrational strategy will in fact do better on average than the CDT-rational strategy, then it's *rational* to play the CDT-irrational strategy.

But Arntzenius doesn't whine. Instead he objects that if *Why ain'cha rich* works against CDT then an exactly parallel argument works against EDT. So the evidentialist is hardly in a position to wield *Why ain'cha rich* against CDT. The remainder of this paper describes and then criticizes that parallel argument.

*2: Arntzenius's Example*

The Yankees and the Red Sox are going to play a lengthy sequence of games; the Yankees win 90% of such encounters. Before each game Mary has the opportunity to bet on either side. The following table summarizes her payoffs on every such occasion as well as our abbreviations for the relevant acts and states:

| | RED SOX WIN (R) | YANKEES WIN (Y) |
|---|---|---|
| Bet on Red Sox (BR) | 2 | -1 |
| Bet on Yankees (BY) | -2 | 1 |

Table 1

Just before each bet a perfect predictor tells her whether her next bet is going to be a winning bet or a losing bet. Now suppose that Mary knows all this. What does EDT recommend?

Suppose that the predictor says: 'Mary, you will win your next bet.' Then the news value $V_W$ (BR) of betting on the Red Sox is:

(7) $V_W(BR) = 2.Cr(R \mid BR \wedge Win) + -1.Cr(Y \mid BR \wedge Win) = 2.1 + -1.0 = 2$

And the news value $V_W(BY)$ of betting on the Yankees is:

(8) $V_W(BY) = -2.Cr(R \mid BY \wedge Win) + 1.Cr(Y \mid BY \wedge Win) = -2.0 + 1.1 = 1$

It follows from (7) and (8) that $V_W(BR) > V_W(BY)$; and EDT recommends V-maximization. So if Mary knows that she will win her next bet then her EDT-rational bet is on the Red Sox.

Suppose that the predictor says: 'Mary, you will lose your next bet.' Then the news value $V_L(BR)$ of betting on the Red Sox is:

(9) $V_L(BR) = 2.Cr(R \mid BR \wedge Lose) + -1.Cr(Y \mid BR \wedge Lose) = 2.0 + -1.1 = -1$

And the news value $V_L(BY)$ of betting on the Yankees is:

(10) $V_L(BY) = -2.Cr(R \mid BY \wedge Lose) + 1.Cr(Y \mid BY \wedge Lose) = -2.1 + 1.0 = -2$

It follows from (9) and (10) that $V_L(BR) > V_L(BY)$. So if Mary knows that she will lose her next bet then her EDT-rational bet is on the Red Sox.

So it follows from (7)-(10) that Mary's EDT-rational bet is going to be on the Red Sox *for every game*.

So Mary will always bet on the Red Sox. And, if the Yankees indeed win 90% of the time, she will lose money, big time. Now, of course, she would have done

much better had she just ignored the announcements, and bet on the Yankees each time. But, being an evidential decision theorist she cannot do this. (Arntzenius 2008: 289-90)

It is easy to see that she would have done better to bet on the Yankees. The average returns to betting on the Red Sox and the Yankees are respectively:

(11) AR (BR) = 90%. –1 + 10%. 2 = –0.7

(12) AR (BY) = 90%. 1 + 10%. –2 = 0.7

It is also easy to see by contrast that CDT *does* recommend betting on the Yankees every time. Win or lose, Mary's bet on any game is causally irrelevant to its outcome. So the causalist's evaluations of those bets are as follows:

(13) U (BR) = V (R $\wedge$ BR).Cr (R) + V (Y $\wedge$ BR).Cr (Y) = 2. 10% – 1.90% = –0.7

(14) U (BY) = V (R $\wedge$ BY).Cr (R) + V (Y $\wedge$ BY).Cr (Y) = –2.10% + 1.90% = 0.7

(Arntzenius 2008: 290). So the causalist bets on the Yankees every time; and he makes an average 70 cents per game. 'So', Arntzenius concludes, 'there are cases in which causal decision theorists, predictably, will do better than evidential decision theorists' (2008: 290).

The argument against *Why ain'cha rich* is therefore a parity argument: if *Why ain'cha rich* works against CDT then this parallel argument works against EDT. In line with the amendment that I initially proposed to Arntzenius's formulation of *Why ain'cha*

*rich*, I suggest that we rewrite it as an argument about *acts* rather than *persons*: so put it runs as follows:

*Yankees*.

(15) The average return to betting on the Yankees exceeds the average return to betting on the Red Sox (*premise: from* (11), (12))

(16) Everyone can see that (15) is true (*premise*)

(17) Therefore betting on the Yankees will foreseeably do better than betting on the Red Sox (*from* (15), (16))

(18) Therefore EDT is committed to what is now the foreseeably worse option for Mary (*from* (7-10), (17))

The dialectical position is now as follows. The evidentialist might think that *Why ain'cha rich* is an argument for preferring EDT to CDT. But Arntzenius seems to have shown that whatever the argument shows, it doesn't show *that*. For a precisely parallel argument, namely *Yankees*, gives just the *same* reason for preferring CDT to EDT. In short: *Why ain'cha rich* cuts both ways if it cuts either way. So *it* cannot motivate a preference for EDT.

### *3: Is the example coherent?*

Our initial concern about *Yankees* is that the example appears to be incoherent, in the sense that it ascribes a belief to the agent that is incompatible, from her own point of view, with the belief that she has a choice. We rely here on a familiar claim about an incompatibility between deliberation, on the one hand, and justified belief about the outcome of that deliberation, on the other. Following Rabinowicz (2002), we shall call the

claim in question the thesis that *deliberation crowds out prediction* (the "DCOP thesis", for short). As Jim Joyce notes, this thesis has wide support, on both sides of the debate between causal and evidential decision theories:

> [M]any decision theorists (both evidential and causal) have suggested that free agents can legitimately *ignore* evidence about their own acts. Judea Pearl (a causalist) has written that while 'evidential decision theory preaches that one should never ignore genuine statistical evidence … actions—by their very definition—render such evidence irrelevant to the decision at hand, for actions change the probabilities that acts normally obey.' (2000: 109) Pearl took this point to be so important that he rendered it in verse:
>
> > Whatever evidence an act might provide
> > On facts that precede the act,
> > Should never be used to help one decide
> > On whether to choose that same act. (2000: 109)
>
> Huw Price (an evidentialist) has expressed similar sentiments: 'From the agent's point of view contemplated actions are always considered to be *sui generis,* uncaused by external factors … This amounts to the view that free actions are treated as probabilistically independent of everything except their effects.' (1993: 261) A view somewhat similar to Price's can be found in Hitchcock (1996).
>
> These claims are basically right: a rational agent, *while in the midst of her deliberations,* is in a position to legitimately ignore any evidence she might possess about what she is likely to do. … A deliberating agent who regards herself

> as free need not proportion her beliefs about her own acts to the antecedent evidence that she has for thinking that she will perform them. (Joyce 2007: 556-7)

Indeed, we think that Joyce here understates the matter. It is not merely that such an agent 'is in a position' to ignore such evidence, as if it were an optional matter. Rather, as Pearl puts it, the choice situation 'render[s] such evidence irrelevant': it ceases to *be* evidence, in other words. The authority that an agent takes herself to have – *qua* agent – over her own future actions, seems inevitably to 'trump' whatever considerations might otherwise have formed the basis for a justified *prediction* (probabilistic or otherwise) about what she will choose to do.

It is true that not all commentators agree with Joyce and the writers he cites on these matters. (One of those who does not is Rabinowicz, from whom we have borrowed the label for the DCOP thesis.) This is not the place to explore the arguments for and against the thesis (though we shall illustrate the flavour of some of the former arguments in a moment). We simply wish to point out that *if* the thesis is accepted, it leads to problems for *Yankees.*

To show why this is so, we begin by noting that the DCOP thesis is closely related to a point at the heart of Dummett's famous (1964) discussion of the coherence of backward causation – a discussion we shall adapt, to illustrate the way in which Yankees is undermined by the thesis. Consider the following example, the *Has Bean Machine:*

> On my office desk, yesterday, there was a box full of beans. The University's bean-counters examined its contents at that time, and assured me that 90% of the beans were Yellow, and 10% Red. How did the beans get there? I'll be sending them there, tomorrow, using my new time transporter (the Has Bean Machine). It

> is yet to be scaled up to human size, but works perfectly for red and yellow beans. It doesn't alter their colour, and I can choose, of course, how many of each colour to send.

Can I trust the bean-counters? At this point, we have Dummett's authority for saying 'no'. Dummett points out that it is coherent for me to believe that a contemplated free action is reliably correlated with some past state of affairs *only* if I do not also believe that I can (in the same circumstances) have knowledge of the state of affairs in question, before I act. So if I am confident of the reliability of the Has Bean Machine, and of my own ability freely to choose what mix of red and yellow beans to send to the past, I cannot *also* take the bean-counters' claims to be reliable.

Dummett's reasoning at this point aligns very closely with the DCOP thesis. Under the assumption that the Has Bean Machine works as advertised – in particular, that it does not change the colour of the beans – the bean-counters' claims amount to a *prediction* about the results of my *deliberation* about which beans to place in the machine. And the thesis assures us that my deliberation crowds out such a prediction: i.e., that it renders it unreliable, from my own epistemic viewpoint, as I deliberate.

Dummett reaches his conclusion by pointing out that familiar proposals to 'bilk' a claimed case of backward causation – i.e., a claimed correlation between a future action and a past states of affairs – rely on arranging matters so that the future action takes place when and only when the relevant past state of affairs does *not* obtain. But as Dummett notes, this requires that the agent in question have epistemic access to the past state of affairs, before she decides whether to perform the future action. In the absence of such access, one cannot bilk.

Conversely, the bilking argument itself provides a way of making vivid the DCOP

thesis: no prediction about one's own future actions can be considered reliable, as one deliberates, because one always has the option to bilk such a prediction. (As we put it above, deliberation thereby *trumps* prediction.) Construed in this general form, the bilking argument is especially salient as an objection to backward causation, because we tend to take for granted that if we could affect the past then we would have access to evidence for our future actions, before we deliberate. (Dummett's contribution is then to highlight the fact that that assumption is crucial, and potentially contestable.) But the underlying point is more basic. Once again, it is the fact that deliberation seems to crowd out prediction.

Let us now apply these considerations to the *Yankees* example. Once Mary knows whether her next bet is a winning bet or a losing bet, she knows that her choice – betting on the Yankees, or betting on the Red Sox – is reliably correlated with the outcome of the game. By a direct application of the DCOP thesis, this means that she cannot take herself to have reliable evidence about the outcome of the game, as she deliberates about how to bet. In particular, therefore, she cannot take herself to be justified in assigning credence 0.9 to a Yankees victory.

Thus the DCOP thesis suggests that there is an incoherence at the heart of the *Yankees* example. The assertion that *Yankees* is a case in which EDT leads to predictable loss depends on the information that the Yankees win 90% of games. According to the DCOP thesis, however, a free agent with the additional knowledge assumed by the example – knowledge, in advance, about whether each bet will win or lose – cannot take this claim about the frequency of Yankees wins to have evidential relevance to her own situation, as she makes her decision. Why not? Because if taken this way, and combined with the information about whether the present bet is a winning bet or a losing bet, it amounts to evidence about what she will choose, which is precisely what the DCOP thesis disallows.

The Has Bean Machine makes this point by analogy. As she decides to bet (and after she finds out whether the bet will win or lose), Mary's epistemic relation to the outcome of the next ball game is exactly like my epistemic relation to the colour of the next bean I place in the Has Bean Machine, to be sent back in time. (We could even add an analogue of the Win/Lose information to the Has Bean Machine, by having the bean selection mechanism sometimes malfunction, in a manner completely predictable in advance.) So Mary's situation, as she contemplates a season of betting on ball games, is exactly like my situation, as I contemplate selecting a series of beans, one at time, to be sent into the past.

As we noted, Dummett's argument shows that to make my beliefs coherent, I must mistrust the University's bean-counters, who assured me that 90% of the beans sent back in time were actually Yellow. By parity of reasoning, Mary's beliefs are incoherent, unless she, too, mistrusts the information that the Yankees will win 90% of games. No matter if the analogue of the bean-counters in this case is none other than Chance itself, stoutly offering a prediction of the percentage of Yankees wins. If deliberation crowds out prediction, then Mary cannot take herself to be justified in believing that prediction, as she decides how to bet; and hence cannot coherently take herself to be facing a certain loss. This objection goes to the heart of Arntzenius's example, for it is the agent's knowledge of the frequency of Yankees wins which is supposed to sustain the conclusion that she knows that she will do less well by EDT than by CDT.[4]

[4] The same point applies to Arntzenius's other example (2008: 290), which resembles Newcomb's problem, except that *both* boxes are transparent, and the predictor has placed $10 in the left-hand box iff he predicted that the agent would not take the right-hand box, which contains $1. Evidential and Causal Decision theories *both* advise taking the contents of both boxes. Arntzenius claims that agents who heed this advice will foreseeably make less money than those who—insanely—take only the box containing $10.

Our complaint about the Yankees case transposes to this case as follows. If the agent knows that she is going to be able to choose what boxes she takes then she knows in advance that she can so contrive her choices as to make the predictor's accuracy arbitrarily close to zero. (She can do this by taking both boxes on any occasion if and only if the predictor has on that occasion left $10 in the left-hand box.) But if she knows in advance that that is an option for her, then she cannot assume in advance that the predictor is

To put this conclusion in proper perspective, we emphasize again that it depends on the DCOP thesis, which is not entirely uncontroversial. Opponents of the thesis (e.g., again, Rabinowicz 2002) seek to undermine it by pointing out that in some circumstances, agents can adopt what amounts to a third-person perspective on their own deliberations – they can stand outside their own deliberative process, as it were, and make reliable predictions about their own decisions within that process. (The crucial issue then becomes whether, and in what sense, this 'third-person' perspective is available *in* deliberation.)

It might seem that a similar move will rescue *Yankees* from our charge of incoherence. That is, it might be objected that even if deliberation crowds out the evidential significance of the fact that the Yankees win 90% of games *as Mary deliberates how to bet in any particular case,* it does not prevent her from appreciating the disastrous consequences of EDT from a more detached perspective – say, from the one she occupies before the start of the baseball season. At that stage, before she is offered the first bet, cannot she take note of what the upshot will be if she makes the individual bets according to EDT, in the light of the fact that the Yankees win 90% of games?

Indeed she can, in our view, but the objection backfires. From this detached perspective, evidential reasoning alone is sufficient to show Mary that she will do much better to treat the entire season's bets as one decision, and to follow the policy of always betting on the Yankees. This ensures that 90% of the time, she will receive the welcome information that she is to make a winning bet. If she is allowed this detached perspective, in other words, then evidential reasoning does as well as causal reasoning. If she is not allowed it, we have seen that the DCOP thesis implies that the information on which the

---

going to be accurate; so she cannot after all foresee that the strategy endorsed by CDT (and by EDT) will be relatively unprofitable.

This case also illustrates especially clearly why the incoherence that it shares with the Yankees example does not arise in the standard Newcomb case. In the standard Newcomb case the one box is opaque; and the only way to discover its contents is to make the very decision whose return depends upon them. So there is no way of knowing in advance what on any occasion of choice you have been predicted to choose. Nor therefore is there any identifiable strategy for systematically falsifying those predictions.

conclusion that EDT leads to loss compared to CDT is based is simply not salient to her, as she makes each individual choice. In neither case, then, can she be in the situation claimed by Arntzenius, of being justified in believing that EDT will do less well than CDT.

Notice that to take advantage of this detached perspective, Mary must be capable of `binding herself to the mast,' so that her resolution at the beginning of the season is not overridden by new evidential circumstances she finds herself in as she makes each individual bet (at which stage, as we saw, the DCOP thesis implies that she is not entitled to a credence 0.9 to a Yankees victory). *Yankees* thus belongs to an interesting class of decision problems in which an agent's beliefs and/or desires change in a predictable way, with predictable implications for rational decision – implications such that a rational agent will deprive his (equally rational) later self of a choice, if he has the means to do so.

We shall return to this aspect of *Yankees* below. For the moment, we emphasize that neither of our two Marys is in the situation claimed by Arntzenius, of being rationally confident that EDT will lead to a loss, in the light of the information that Yankees win 90% of their games. Pre-season Mary can take account of this information. Accordingly, she takes EDT to recommend binding herself to the policy of always betting on the Yankees, and expects that this policy will lead to a net gain. But pre-game Mary, once she has been told whether she faces a winning bet or a losing bet, cannot rationally take information about the usual frequency of Yankees wins to be applicable to her case, on pain of conflict with the DCOP thesis. So although EDT now leads her to bet on the Red Sox, she, too, is not in the situation claimed by Arntzenius.

### *4: Restoring the disparity*

Our second objection also turns on the fact that *Yankees* involves a shift in epistemic perspective, which Arntzenius's argument ignores. It approaches the point from a different direction, however, and does not assume the DCOP thesis. Once again, our aim is to show that *Yankees* suffers from flaws that do not affect *Why ain'cha rich*; and hence that one can consistently maintain the latter against CDT whilst denying that the former has any weight against EDT. We shall do this by examining arguments in which the relevant flaw in *Yankees* appears more clearly.

Here is one. Every Monday morning everyone has an opportunity to pay $1 for a medical check-up at which a prescription is issued should the doctor deem it necessary. Weeks in which people take this opportunity are much more likely to be weeks in which they fall ill than weeks in which they pass it up. In fact on average, 90% of Mondays on which someone *does* go in for a check-up fall in weeks when he or she is subsequently ill; whereas only 10% of Mondays on which someone *doesn't* go for a check-up fall in weeks when he or she is subsequently ill. There is nothing surprising or sinister about this correlation: what explains it is rather the innocuous fact that one is more likely to go for a check-up when one already has reason to think that one will fall ill.

All weekend you have suffered from fainting and dizzy spells. You're pretty sure that there is something wrong. Should you go for the check-up on Monday morning? Clearly if you *are* ill this week, it will be better to have the prescription than not, so the check-up will have been worth your while. But if you are *not* ill this week then the check-up will have been a waste of money. Your payoffs are therefore as stated in the following table, which also gives our abbreviations for the relevant states and acts:

| | Well this week (W) | Ill this week (~W) |
|---|---|---|
| Check-up (C) | 1 | 0 |

| | | |
|---|---|---|
| No Check-up (~C) | 2 | -1 |

Table 2

Given this table and the statistical facts already mentioned we may compute the average return to going and to not going for a check-up:

(19) AR (C) = 10%. 1 + 90%. 0 = 0.1

(20) AR (~C) = 90%. 2 + 10%. –1 = 1.7

So the average return to going for a check-up exceeds that of not going for a check-up. We may therefore construct the following argument against going for a check-up:

*Why ain'cha well.*

(21) The average return to going for a check-up exceeds the average return to *not* going for a check-up (*premise: from* (19), (20))

(22) Everyone can see that (21) is true (*premise*)

(23) Therefore going for a check-up is now a foreseeably worse option for you than not going for one (*from* (21), (22))

Should you then not go for your check-up? That would be insane: *of course* you should given the dizzy spells etc. So what is wrong with the argument?

What is wrong with it is the inference from (21) and (22) to (23). Taken over *every* opportunity for a check-up for *anyone*, it is true that those opportunities that are taken shortly precede illness much more often than those that are not taken. But this is not the

relevant basis on which *you* should compute the average returns to your options *now*. What you should rather compute are the average returns to your options *given what you now know about yourself*. That is: you should compute the average returns to C and ~C, not amongst all opportunities for check-ups but amongst *occasions on which the subject is suffering from your symptoms*. That is: you should look at what happens to people when they are suffering from fainting and dizziness. Is subsequent illness *amongst these people on these occasions* any more frequent amongst those who go for check-ups than amongst those who do not? Common sense suggests that amongst such people on such occasions, the subsequent incidence of illness is high in both groups and that it is equal in both groups. In that case it is easily verified that:

(24) Amongst people with the symptoms that you now have, the average return to going for a check-up exceeds that of *not* going for a check-up.

So *for you*, *now*, going for a check-up is foreseeably the *better* option.

The fallacy of *Why ain'cha well* is that of applying an overly broad statistical generalization to a single case: in this case, yourself. The generalization is overly broad because it is not limited to cases that resemble yours in relevant respects that you know about. Knowing that you are suffering from dizziness and fainting, the statistical generalization that you should apply to yourself is not (21); it is one that covers only that sub-population that resembles your present stage in that respect i.e. (24). Hence applying (21) rather than (24) to yourself involves a failure to consider evidence that is both relevant and available.

Whatever its other faults *Why ain'cha rich* does not commit *this* error. The inference of (4) from (3), and ultimately from (1) and (2), is not an application of an

*overly* broad statistical generalization. Anyone facing Newcomb's problem has *no* evidence that relevantly distinguishes him or her now from anyone else whom the statistical generalization (1) covers, that is, all other persons who ever face this problem.[5] The application of (1) to anyone facing Newcomb's problem is therefore not illegitimate in the way that the application of (21) to your present stage is illegitimate.

What about *Yankees*? It turns out that whether it commits this fallacy depends upon what 'now' in (18) is supposed to denote. Consider first any moment *after* she has learnt whether her next bet will win or lose but *before* she has decided how to bet. It would be fallacious for Mary to apply *Yankees* to herself then, because it would be fallacious for her then to apply (15) to herself. For at any such moment she has relevant information that puts her in a narrower sub-population than that over which (15) generalizes. It puts her not only in the population of bettors but in the sub-population of *winning* bettors (if she has just learnt that she will win), or in the sub-population of *losing* bettors (if she has just learnt that she will lose).

Thus suppose that the predictor has just said to Mary: 'Mary, you will win your next bet.' Then the statistical generalization that she should apply to herself is not the one that compares the average return to placing a bet on the Red Sox with the average return to placing a bet on the Yankees (i.e. (15)). It is the one that compares the average return to placing a *winning* bet on the Red Sox with the average return to placing a *winning* bet on

[5] Here we slide over an important distinction within the class of Newcomb scenarios. In some such cases it is either stipulated or allowed that prior to choosing the agent is directly aware of a 'tickle'—an inclination to choose in one direction or the other —whose presence screens off his act from the earlier prediction of it and so also from the contents of the opaque box (Eells 1982 ch. 6).

In these 'tickle' cases it is of course false that the agent has no evidence that relevantly distinguishes him from anyone else facing the problem, so in tickle case *Why Ain'cha Rich* does not support one-boxing. But then neither does EDT support one-boxing in tickle cases: on the contrary, the presence of a screening-off inclination in *either* direction makes the agent's act evidentially irrelevant to the contents of the opaque box and hence also entails the unique EDT-rationality of two-boxing.

So the defender of EDT should be comfortable with this distinction and also with the consequent qualification of the statement in the text. His position will continue to be that *Why Ain'cha Rich* supports EDT over CDT because it mandates one-boxing in just *those sorts* of Newcomb cases where EDT recommends one-boxing and CDT does not. (Thanks to a referee.)

the Yankees. Now we know from Table 1 that the average return to placing a winning bet on the Red Sox is 2 and the average return to placing a wining bet on the Yankees is 1. Hence the appropriate generalization is not (15) but:

(25) The average return to placing a winning bet on the Red Sox exceeds the average return to placing a winning bet on the Yankees.

Inferring (18) ultimately from premises including (15) rather than its opposite from ones including (25) is just the same fallacy as that of *Why ain'cha well*: the fallacy of ignoring available and relevant evidence. So if 'now' in (18) refers to a time *after* Mary learns that she will win her next bet then *Yankees* is fallacious.

With appropriate adjustments the argument of the foregoing paragraph will apply if 'now' in (18) refers to any time at which Mary has just learnt that she will *lose* her next bet. Hence it is fallacious to apply *Yankees* to Mary once she has learnt the outcome of her next bet, *whatever* she has learnt.[6]

[6] It is reasonable to wonder whether this diagnosis of the error in Arntzenius's argument is not sensitive to the way in which we are here applying the principle of 'total evidence'. Our objection is that more specific information is available to Mary, on any occasion, than is used in Arntzenius's calculation of the average return to each of her options on that occasion. But how are we supposed to incorporate this information?

In the present framework the additional information (that 'the bet is losing' or that 'the bet is winning') is used as a description of the action whose expected utility is thus calculated. But why is that the right way of incorporating the additional evidence? In the simpler context of inductive reasoning—without considering actions as yet—the principle of total evidence would say: Given that the statistical probability of H (x) given E (x) is r, and given that one's total evidence about the individual a is that E (a) is the case, one's subjective probability that H (a) is the case ought to be r. So one's evidence figures as a proposition on which one then conditionalizes.

Applying the principle in this way yields the result that in any case a bet on the Red Sox is the better bet. For instance: since the statistical probability that x is a bet on a game that the Red Sox win, given that x is a bet on the Red Sox and x is a winning bet, is 1, one's credence that the Red Sox will win this game given that this bet is a winning bet on the Red Sox should be 1. That yields one of the conditional probabilities figuring in (7); by similar means we arrive at the rest and so conclude that in any case Red Sox is the rational bet. But that is exactly what EDT implies and what we are here proposing: given the information that Mary has on any particular occasion, she is indeed rational on that occasion *to bet on the Red Sox*, regardless of (15). So our argument about Mary's case is indeed robust to variations in the exact manner in which you are supposed to apply the principle of total evidence to it.

A related objection is that conditionalizing on the information that, say, this bet is going to win, does nothing to affect Mary's confidence that in the long run and taken over all bets, bets on the Yankees will do better than bets on the Red Sox. So even if she learns that she will win her next bet, is she not still

What about the time just *before* Mary has learnt the outcome of her next bet? At those times she does not *have* the evidence that is supposed to vitiate the inference from (15) to (18). So isn't the argument then just as plausible as *Why ain'cha rich*?

It's true that it doesn't then commit the *same* fallacy as *Why ain'cha well*. The trouble is that now we cannot infer (18) from (7)-(10) and (17) because it no longer follows from (7)-(10) that EDT recommends betting on the Red Sox. *Before* Mary has learnt whether she will win her bet, the news values of betting on the Red Sox and on the Yankees are:

(26) $V(BR) = 2.Cr(R \mid BR) + -1.Cr(Y \mid BR) = 2.10\% + -1.90\% = -0.7$

(27) $V(BY) = -2.Cr(R \mid BY) + 1.Cr(Y \mid BY) = -2.10\% + 1.90\% = 0.7$

Hence at this time EDT recommends betting on the *Yankees*, so once again its preferred option is the one that foreseeably does better.

*Yankees* is therefore unsustainable for reasons that do nothing to undercut *Why ain'cha rich*. Neither after nor before Mary has learnt whether her bet is a winner does *Yankees* support an option that diverges from EDT in the way that *Why ain'cha rich* supports an option that diverges from CDT. Not after, because then it doesn't *support* a divergent option; not before, because then it doesn't support a *divergent* option. And this restores the disparity between EDT and CDT. *Why ain'cha rich* does not cut both ways: *it*

---

entitled to be just as confident in (15) as she was before? And in that case doesn't Arntzenius's argument still go through?

But the point is then not that Mary's information makes (15) false but that it makes it inappropriate to apply (15) to her present situation. For sure, her next bet belongs to a population of bets of which (15) is true. But the oracle's predication also puts it in a *narrower* population of which (25) is true. And the principle of total evidence tells us that she should be applying the generalization about the *narrower* population to her present bet rather than the (equally true) generalization about the broader population. Otherwise it would be rational not to visit the doctor, even given these rather serious symptoms, on the grounds that in the general population people who visit doctors fall sick more often than those who do not. (Thanks to a referee.)

tells against CDT but not EDT, and we have been given no parallel argument that tells against EDT but not CDT.

### *5: Objection: Preference Instability*

Both replies to Arntzenius involve some *difference* between Mary's relevant *early* informational state—before the season begins—and her relevant *late* information states—just before some particular bet, and after learning whether it is a winner. It is explicit in s3 that the information that Yankees win 90% of the time is available in the early state but not in any late state. It is implicit in s4 that the information that whether this bet will win is available and relevant in each later state but not available in the early state. Both replies therefore commit us to saying that before the series begins she will

(28) —rationally prefer betting on the Yankees every time to betting on the Red Sox every time.

(29) —foresee that her informational state just prior to each bet will be different from what it is now.

(30) —foresee that in light of that new informational state, whatever it is, she will rationally prefer betting on the Red Sox.

But is (28)-(30) really a coherent combination? We can see two reasons to worry that it isn't.

The first worry arises in connection with binding. Suppose that before the season begins we offer Mary the chance to bind herself to a single betting policy for the whole season. As we've already seen, EDT recommends taking this opportunity; in particular it

recommends that she bind herself to the policy of always betting on the Yankees. So if she follows EDT then that is what Mary will do, even though she *knows* that before each bet she will get (free) information in the light of which EDT will recommend betting on the Red Sox.

But how can this be? How can it be rational now to bind yourself to a policy that you *know* it will be rational to reverse in the light of future information? Well, isn't it rational for Ulysses to bind himself to the mast? It is: but then Ulysses knows that his future preference for a different option will be caused solely by an exogenous shock to his *desires* (the sirens' singing). Whereas in the present case Mary knows that she will be getting new *information* in the light of which her unchanged desire for money will make it rational to bet on the Red Sox, not the Yankees. So Mary does, but Ulysses does not, violate the following plausible principle:

(31) If free and relevant information is available before acting then you should take the information before acting rather than binding yourself now to some course of action.

So in Mary's case EDT and (31) are incompatible. Does this mean that EDT gives her bad advice?

It doesn't, because (31) is false. Its application to any case depends on whether the binding policy makes any (evidential) *difference* to *what* that information is in that case. In particular, if binding yourself now makes it more likely that the information will be *good* news then it may indeed *not* be worth waiting for the information, free and relevant though it is. And so it is in Mary's case with the option of pre-seasonal binding. If she binds herself now to bet on the Yankees all season then she will on 70% of betting

occasions get the good news that her next bet will win; if she does not do this then she will get that good news on only 10% of betting occasions.[7]

The second worry about (28)-(30) is not that it is decision-theoretically implausible but that it seems to violate a plausible constraint on rational preferences whether or not these can or do direct the agent's behaviour. (28)-(30) implies that Mary has some preference that she *knows* will be reversed in the light of information that she *knows* that she is going to get. But then wouldn't the *ex post* preference have been rational all along?[8]

It would not. The pattern of preferences that (28)-(30) realizes is simply the perfectly rational upshot of an unusual but by no means fantastic statistical pattern of which there follows a more realistic example.

---

[7] But couldn't we make Mary's early and soon-to-be reversed preference for a bet on the Yankees practically harmful to her? Suppose she knew that we were going to offer her: (i) a choice between betting on the Yankees or on the Red Sox *before* she learnt whether her next bet was going to be a winner; and then (ii) the option to switch bets for a fee, *after* she had learnt whether her next bet was going to be a winner. EDT seems to commit her to (i) a bet on the Yankees and (ii) paying the fee—as long as it is less than \$1—and betting instead on the Red Sox. But this is irrational: when offered the choice (i) she could *foresee* that she would get information that would lead her to prefer a bet on the Red Sox, so the more rational thing to do would be to take the bet on the Red Sox *then* and save herself the fee.

But if she is going to be offered (i) *and* (ii) then EDT will *not* recommend, at the time of (i), that she take the bet on the Yankees. That recommendation relied on the assumption, implicit in (27), that the news value of a win for the Yankees, given that she bets before learning the outcome of her bet, is 1. But if Mary *knows* that she will change her mind and hence her bet (as she must do for an initial bet on the Yankees to be irrational), then this assumption no longer holds: at the time of (i) the value of a Yankees win given that Mary *now* bets on the Yankees is rather –1, because she knows that when the Yankees win she'll be holding a Red Sox ticket. In fact in that situation EDT *will* prescribe betting early on the Red Sox and saving the fee.

[8] A similar but not quite identical situation arises in Newcomb's problem itself: the follower of EDT begins with a preference that he takes only the opaque box in the knowledge that whatever its contents, he will later think that he would have done better to take both boxes. The difference is that in the Newcomb case it is not the relative *news values* of one-boxing and two-boxing that foreseeably fluctuate—for once the agent has taken one box his *ex post* news value for taking two is undefined—; rather it is that the agent can foresee regretting, so to speak counterfactually, what he currently prefers to do. Foreseeable regret is a much discussed phenomenon that has little bearing on our dispute with Arntzenius; what is important is that we distinguish it from the phenomenon of foreseeable preference instability, which is both relevant and relatively little discussed in these contexts.

On the other hand the fact that EDT violates the principle of dominance in the Newcomb case certainly implies that a *modification* of that case accurately simulates Mary's situation. Suppose that before acting the evidentialist agent gets to peek into the opaque box. Then he knows before peeking that (a) he now prefers one-boxing to two-boxing; and that (b) whatever he sees in the opaque box he will after seeing it prefer two-boxing to one-boxing. So this modified Newcomb case is also a case of foreseeable preference instability. (Thanks to a referee.)

The admissions statistics for the English and Mathematics Departments at Simpson's Paradox University (SPU) are what you would expect. Male applicants are less successful than female ones overall: 14% of men who apply for admission on to a graduate course in one of these Departments are successful but 20% of women who so apply are successful. But in each Department the discrepancy is reversed: 5% of male applicants for Mathematics are successful as against 1% of female applicants; 50% of male applicants for English are successful as against 25% of female applicants. The explanation is that male candidates are more likely than female candidates to apply to the more competitive Mathematics Department.

Your best friend has just told you that he or she has applied to graduate school at SPU. For some reason it matters greatly to you that your friend's application to this particular university is successful. You know that your friend would have applied to the English Department or to the Mathematics Department (but not both). But being very absent-minded you have forgotten (a) which of these it is and (b) whether your friend is male or female. You ask your friend about (a).

Before you hear the answer you reflect that *now*, the news value of the information that your friend is a girl exceeds the news value of the information that your friend is a boy. After all, female applicants to SPU do better than male ones. You then reflect that *after* you have heard the answer to (a) and *whatever that answer is*, the news value of the information that your friend is a *boy* will now exceed the news value of the information that your friend is a *girl*. After all, male applicants to SPU's Mathematics Department do better than female ones, and male applicants to SPU's English Department do better than female ones. Finally, you reflect that now, before you know the answer to (a), you have a preference over the possible answers to (b) that you *know* is going to be reversed in the light of information that you are about to receive.

But there is nothing *irrational* about your (news-) preferences in this situation. At any point in time you have just the preference that is appropriate in the light of all of the information that you then possess. The only *peculiarity* of the situation is the foreseeable fluctuation in your preference; but that fluctuation is itself a perfectly rational response to an undeniably real statistical phenomenon i.e. Simpson's Paradox.

It follows that even in the more general setting where one's preferences may be purely passive, preferential patterns analogous to (28)-(30) may be perfectly rational. So too are the actions based upon them that *are* available in Mary's case and which Evidential Decision Theory there recommends.[9]

References


Arntzenius, F. 2008. No regrets, or: Edith Piaf revamps decision theory. *Erkenntnis* 68: 277-297.

Dummett, M.1964. Bringing about the past. *Philosophical Review* 73, 338–359.

Gibbard, A. and W. Harper 1981. Counterfactuals and two kinds of expected utility. In Harper, W., R. Stalnaker and G. Pearce, eds: *Ifs: Conditionals, Belief, Decision, Chance and Time*. Dordrecht: D. Reidel: 153-192.

Hitchcock, C. 1996: Causal decision theory and decision-theoretic causation. *Noûs* 30: 508-526.

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

———. 2007. Are Newcomb problems really decisions? *Synthese* 156:537-562.



[9] We are grateful to Frank Arntzenius and to two referees for helpful comments on earlier drafts of this paper.

Lewis, D. 1981a. Causal decision theory. In Gardenfors, P. and N.-E. Sahlin, eds: *Decision, Probability and Utility: Selected Readings*. Cambridge: Cambridge University Press: 377-405.

———. 1981b. Why ain'cha rich? *Nous* 15: 377-80.

Nozick, R. 1970. Newcomb's problem and two principles of choice. In Rescher, N., ed. *Essays in honor of Carl G. Hempel*. Dordrecht: D. Reidel: 114-46.

Pearl, J. 2000. *Causality: Models, Reasoning, and Inference.* Cambridge: Cambridge University Press.

Price, H. 1993. The direction of causation: Ramsey's ultimate contingency. In Hull, D., Forbes, M. and Okruhlik, K., eds, *PSA 1992,* Volume 2. East Lansing, Michigan: Philosophy of Science Association: 253-267.

Rabinowicz, W. 2002. Does practical deliberation crowd out self-prediction? *Erkenntnis* 57: 91–122.