Equal Opportunity and Newcomb's Problem

Ian Wells

Forthcoming in *Mind*

Abstract

The 'Why ain'cha rich?' argument for one-boxing in Newcomb's problem allegedly vindicates evidential decision theory and undermines causal decision theory. But there is a good response to the argument on behalf of causal decision theory. I develop this response. Then I pose a new problem and use it to give a new 'Why ain'cha rich?' argument. Unlike the old argument, the new argument targets evidential decision theory. And unlike the old argument, the new argument is sound.

1 Welfare and Rationality

There is a difference between doing well because of the opportunities you have, on the one hand, and doing well because of the decisions you make, on the other. Consider:

> *Bluecomb*. There is a transparent box and an opaque box. You have two options: you can take just the opaque box (one-box) or you can take both boxes (two-box). The transparent box contains $1,000. The opaque box contains either $1,000,000 or nothing, depending on a judgment made yesterday by a highly reliable observer. The observer looked at you. If the observer judged that you have blue eyes, the opaque box contains $1,000,000. If the observer judged that you do not have blue eyes, the opaque box contains nothing.

Two agents, Blue and Green, face *Bluecomb* once a month for many months on end. Each month, blue-eyed Blue one-boxes and green-eyed Green two-boxes. Month after month, Blue makes

$1,000,000 while Green makes $1,000. (Occasionally, when the observer misjudges, Blue makes nothing and Green makes $1,001,000. But these occasions are few and far between.) In the long run, Blue's bankroll dwarfs Green's.

Blue does well because of the opportunities Blue has and in spite of the bad decisions Blue makes. Blue does well but could have done better, had Blue not left behind $1,000 every month. Green does poorly because of the opportunities Green lacks and in spite of the good decisions Green makes. Green does poorly but could not have done any better than Green did. Green is poor but rational. Blue is irrational but rich. When the opportunities are unevenly distributed, the link between welfare and rationality is compromised.

2 'Why Ain'cha Rich?'

The foregoing exposes the mistake in a common argument for one-boxing in Newcomb's problem.[1] The problem:

> *Newcomb*. There is a transparent box and an opaque box. You have two options: you can take just the opaque box or you can take both boxes. The transparent box contains $1,000. The opaque box contains either $1,000,000 or nothing, depending on a prediction made yesterday by a highly reliable predictor. If the predictor predicted that you would one-box, the opaque box contains $1,000,000. If the predictor predicted that you would two-box, the opaque box contains nothing.

Two agents, Eva and Casey, face *Newcomb* once a month for many months on end.[2] Each month, Eva one-boxes and Casey two-boxes. Month after month, Eva makes $1,000,000 while

[1] This section expounds the standard causalist response to the 'Why ain'cha rich?' argument. The response is due to Gibbard and Harper (1978), endorsed by Lewis (1981b) and Joyce (1999), and criticized by Ahmed (2014). Newcomb's problem is introduced in Nozick (1969).

[2] Throughout this essay, I consider agents facing decision problems repeatedly over a large span of time. To forestall confusion, let it be understood that (i) the agent always believes that she is making a one-off decision and (ii) the agent's memory of past decisions and their outcomes is erased between rounds.

Casey makes $1,000. (Occasionally, when the predictor mispredicts, Eva makes nothing and Casey makes $1,001,000. But these occasions are few and far between.) In the long run, Eva's bankroll dwarfs Casey's.

Evidential decision theory commends Eva's decisions to one-box and condemns Casey's decisions to two-box.[3] Causal decision theory commends Casey's decisions and condemns Eva's. Many evidentialists appeal to the welfare gap between Eva and Casey to argue that Eva's decisions are rational and Casey's are irrational.[4] This is the so-called 'Why ain'cha rich?' argument for one-boxing in *Newcomb*.

Sometimes the argument is put in terms of expected returns. Supposing for concreteness that the predictor is believed to be 90% reliable, the expected return on one-boxing is $900,000 while the expected return on two-boxing is only $101,000. The argument then proceeds as follows:

(A1) In *Newcomb*, one-boxing uniquely maximizes expected return.

(A2) Choosing an option is rational if and only if the option maximizes expected return.

(A3) Therefore, in *Newcomb*, one-boxing is the uniquely rational option.

The problem with this way of putting the argument is that (A2) begs the question against the causalist. The expected return on an option *just is* the evidential expected value of the option. Hence, (A2) amounts to the claim that rational decision-making is a matter of maximizing evidential expected value—precisely what the causalist denies.

The argument is better framed as an inference to the best explanation. The fact to be explained is that Eva does better than Casey in the long run, or, more generally, that agents

[3] Here and throughout, I assume for simplicity that the agent's values are linear in dollars. Some authors try to reconcile evidential decision theory with two-boxing. See, for example, Jeffrey (1965), Eells (1982), Price (1986) and Burgess (2004). For criticism of these attempts, see Joyce (1999).

[4] For example, Ahmed says that the 'Why ain'cha rich?' argument 'seems to me to be a strong argument for one-boxing in the standard Newcomb case' (Ahmed 2014, p. 194). Hare and Hedden (2016) appeal to a similar argument against causal decision theory based on a non-standard Newcomb case. My criticism of the standard argument applies to their argument as well.

who follow the advice of evidential decision theory in *Newcomb* are richer, on average, than agents who follow the advice of causal decision theory.[5] The best explanation of this fact, so the argument goes, is that evidential decision theory gives rational advice and causal decision theory does not.

Why Ain'cha Rich? (*Newcomb*)

(B1) In *Newcomb*, agents who follow the advice of evidential decision theory are richer, on average, than agents who follow the advice of causal decision theory.

(B2) The best explanation of (B1) is that evidential decision theory gives rational advice and causal decision theory gives irrational advice.

(B3) Therefore, in *Newcomb*, evidential decision theory gives rational advice and causal decision theory gives irrational advice.

If sound, the argument spells the end of causal decision theory.

But the argument is not sound. Recall *Bluecomb* and suppose that agents who one-box in *Bluecomb* tend to have blue eyes while agents who two-box tend to have non-blue eyes. Consider two fictitious decision theories: the sane theory and the insane theory. Let the sane theory advise two-boxing in *Bluecomb* and let the insane theory advise one-boxing.[6] We can now manufacture a clearly unsound inference to the best explanation that mirrors Why Ain'cha Rich? (*Newcomb*).

[5] There is no need to empirically test this claim. Insofar as we are justified in believing the description of *Newcomb*, we are justified in believing (B1).

[6] The insane theory is not evidential decision theory. For someone who does not know the eye color of the agent, one-boxing is evidence of blue eyes. But I assume that the agent facing *Bluecomb* knows her own eye color. Such knowledge screens off the evidential impact of her decision on her eye color. Hence, unlike the insane theory, evidential decision theory correctly recommends two-boxing in *Bluecomb*.

Why Ain'cha Rich? (*Bluecomb*)

(C1) In *Bluecomb*, agents who follow the advice of the insane theory are richer, on average, than agents who follow the advice of the sane theory.

(C2) The best explanation of (C1) is that the insane theory gives rational advice and the sane theory gives irrational advice.

(C3) Therefore, in *Bluecomb*, the insane theory gives rational advice and the sane theory gives irrational advice.

It is true that agents like Blue who follow the advice of the insane theory are richer, on average, than agents like Green who follow the advice of the sane theory. But the best explanation of this fact is not that the former choose rationally and the latter choose irrationally. It is rather that agents who follow the advice of the insane theory tend to have blue eyes while agents who follow the advice of the sane theory tend not to, and blue-eyed agents are frequently afforded the opportunity to win $1,000,000 while non-blue-eyed agents are frequently denied that opportunity. (C2) is false.

Returning to *Newcomb*, suppose that the predictor's prediction is based on a brain scan: the predictor predicts that the agent will one-box if and only if the scan says that the agent has a certain neural property E. The predictor is reliable because one-boxers tend to have E and two-boxers tend to lack E.

The mistake in Why Ain'cha Rich? (*Newcomb*) is now laid bare. It is true that agents like Eva who follow the advice of evidential decision theory are richer, on average, than agents like Casey who follow the advice of causal decision theory. But the best explanation of this fact is not that the former choose rationally and the latter choose irrationally. It is rather that agents who follow the advice of evidential decision theory tend to have E, while agents who follow the advice of causal decision theory tend to lack E, and those who have E are frequently afforded the opportunity to win $1,000,000 while those who lack E are frequently denied that opportunity. Like (C2), (B2) is false.

The predictor's policy of pre-rewarding E-bearers in *Newcomb* is analogous to the observer's policy of pre-rewarding bearers of blue eyes in *Bluecomb*. In both cases, the link between welfare and rationality is compromised by an uneven distribution of opportunity.

3 Equal Opportunity

Let a *WAR argument* be an argument in the mold of those above: an inference to the best explanation from a premise about the average welfare of agents to a conclusion about the rationality of their decisions. The preceding may seem to suggest that WAR arguments fail across the board—that there is just no connection between doing well and deciding rationally. But that suggestion should be resisted. There *is* a connection, even if it breaks on occasion. Consider:

> *Coin Toss.* A box contains either a $6,000 check or a $4,000 invoice. The content of the box was determined by a distributor, who tossed a fair coin. If the coin landed heads, the box contains the check. If the coin landed tails, the box contains the invoice. You have two options: you can buy the box for $3,000 or you can take it for free.

Two agents, Watson and Dudley, face *Coin Toss* once a month for many months on end. Each month, Dudley buys the box and Watson takes it for free. Dudley loses an average of $2,000 per month, winning $3,000 half of the time and losing $7,000 the other half. Watson gains an average of $1,000 per month, winning $6,000 half of the time and losing $4,000 the other half. Watson's savings steadily grow while Dudley's dwindle. As the years pass, the welfare gap between Watson and Dudley widens.

Watson does well. Dudley does poorly. Crucially, the relative welfare of the agents reflects the rationality of their decisions. Watson does well because he chooses rationally and Dudley does poorly because he chooses irrationally. We do not blame Dudley's poverty on an uneven distribution of opportunity because there is no such distribution to blame. Unlike the policy of the observer in *Bluecomb* or the policy of the predictor in *Newcomb*, the policy of the distributor in

Coin Toss is perfectly impartial. Dudley is afforded the opportunity to win $6,000 just as frequently as is Watson. But Dudley squanders the opportunities while Watson seizes them. Dudley is poor and irrational. Watson is rational and rich.

Coin Toss furnishes a sound WAR argument. Recall the sane theory and the insane theory. Let the sane theory advise taking the box for free and let the insane theory advise buying it.

Why Ain'cha Rich? (*Coin Toss*)

(D1) In *Coin Toss*, agents who follow the advice of the sane theory are richer, on average, than agents who follow the advice of the insane theory.

(D2) The best explanation of (D1) is that the sane theory gives rational advice and the insane theory gives irrational advice.

(D3) Therefore, in *Coin Toss*, the sane theory gives rational advice and the insane theory gives irrational advice.

Unlike (B2) and (C2), (D2) is true. The best explanation of the success of those who follow the advice of the sane theory is that they choose sanely.

It is no mystery why WAR (*Coin Toss*) succeeds but WAR (*Bluecomb*) and WAR (*Newcomb*) fail. WAR arguments try to draw a conclusion about the rationality of a decision from a premise about the welfare of agents who make that decision. But how well an agent fares does not just depend on what decision the agent makes. It also depends on the circumstance in which the agent makes the decision.[7] And the circumstance in which the agent makes the decision—being outside of the agent's control—plays no role in our evaluation of the rationality of the decision. We do not blame people for being dealt a bad hand, nor do we praise them for being dealt a good one. So

[7] By 'circumstance', I mean what Lewis meant by 'dependency hypothesis': 'a maximally specific proposition about how the things [the agent] cares about do and do not depend causally on [the agent's] present actions' (Lewis 1981a, p. 11). As Lewis showed, the dependency hypotheses for an agent are causally independent of the agent's present actions. In other words, the agent has no control over which dependency hypothesis obtains.

if we are to trust the conclusion of a WAR argument, we must be assured that the measurements of welfare under comparison do not reflect systematic differences in the circumstances in which the decisions were made. We must be assured, in other words, that the agents whose welfare we are comparing were given an equal opportunity to succeed. *Coin Toss* provides such assurance. *Bluecomb* and *Newcomb* do not.

My aim in what follows is to present a problem that provides the requisite assurance while prizing apart the advice of evidential and causal decision theory. I will use the problem to give a new 'Why ain'cha rich?' argument. Unlike the old argument, the new argument targets *evidential* decision theory. And unlike the old argument, the new argument is sound.[8]

4 Interlude

By way of building up to the problem, I will begin with a similar, simpler problem.[9]

> *Viewcomb*. The setup is the same as *Newcomb*, only now you can look inside the opaque box before making your decision. However, if you wish to make your decision straightaway, without first looking in the box, you must pay a small fee. As before, the predictor predicted only whether you will one-box or two-box.

From a causalist perspective, the difference between *Viewcomb* and *Newcomb* is immaterial. In *Viewcomb*, the causalist simply looks inside the box and then, no matter what she sees, takes both boxes, just as she would in *Newcomb*.

However, from an evidentialist perspective, the difference between *Viewcomb* and *Newcomb* is significant. A reflective evidentialist reasons as follows:

[8] Arntzenius (2008) also gives a WAR argument against evidential decision theory. The argument is ingenious but, as others have observed, subtly unsound. I discuss Arntzenius' argument and its relation to mine in Appendix B.

[9] Versions of the simpler problem are discussed in Gibbard and Harper (1978), Skyrms (1990), Arntzenius (2008), Meacham (2010), Ahmed (2014) and Hedden (2015).

> Looking in the box brings bad news. For if I look then, no matter what I see, I will two-box.[10] But if I two-box, the predictor probably predicted as much, in which case I will see an empty box and make only $1,000. On the other hand, paying the fee brings good news. For if I pay then, not knowing what is inside the box, I will one-box. And if I one-box, the predictor probably predicted as much, in which case I will make $1,000,000. So I should pay the fee.

In fact, the fee need not be small for the evidentialist to pay it. Supposing that the predictor is believed to be 90% reliable, the evidentialist will pay up to $799,000 to avoid looking in the box![11]

Evidential decision theory's treatment of *Viewcomb* may raise some eyebrows. But the problem does not furnish a sound WAR argument against the theory. First, evidentialists are actually richer, on average, than causalists, even though evidentialists pay the fee. So causalists cannot exploit the problem in a sound WAR argument against evidential decision theory. Can anyone else? Consider the insane theorist, who looks in the box but nevertheless one-boxes thereafter. (This is insane behavior. Everyone agrees that if you see, for example, that the opaque box is empty, you should not take only the empty box. You should rather take the $1,000 too.) Insane theorists are richer, on average, than evidentialists, since they reap the benefits of the predictor's partiality without paying the fee. Still, they cannot exploit the problem in a sound WAR argument against evidential decision theory. After all, in *Viewcomb* as in *Bluecomb*, the best explanation of the insane theorists' success is not that they choose rationally but rather that they are frequently afforded golden opportunities.[12]

[10] Why? If the evidentialist sees that the box is empty, the evidential expected value of one-boxing is 0 and that of two-boxing is 1,000. If she sees that the box is full, the evidential expected value of one-boxing is 1,000,000 and that of two-boxing is 1,001,000. Either way, evidential decision theory recommends two-boxing.

[11] This consequence dramatizes evidential decision theory's violation of Good (1967)'s theorem, the claim that it is always rational to gather more evidence before making a decision, provided that the cost of so doing is negligible. For more on violations of Good's theorem, see Skyrms (1990), Buchak (2010), Buchak (2012) and Hedden (2015).

[12] Reflective evidentialists are also frequently afforded golden opportunities in *Viewcomb*, so it is not as if the insane theorists have a leg up in this respect. But I am inclined to think that so long as *some* decision is disadvantaged,

Ruminating on the success of insane theorists in *Viewcomb* helps extinguish any lingering sympathy for WAR (*Newcomb*). But if we seek a sound WAR argument against evidential decision theory, we must look elsewhere.

5 The Problem

The problem I will present resembles *Viewcomb* in three respects.

First, it is a sequential problem: there are two stages at which the agent must make a decision and the prospects of the options at the first stage depend on what the agent believes she will do at the second stage. Second, the problem involves evidence gathering. Specifically, the choice at the first stage is a choice of whether to gather evidence at no cost or to pay to keep the evidence away. Third, the probabilistic relations in the problem are such that if the evidentialist gathers the evidence then, no matter what she learns, evidential decision theory will require her to make a decision that she antecedently hopes she will not make. For this reason, the evidentialist pays to keep the evidence away.

I will begin by describing the problem in words. Parts of the description are tedious. Those parts are represented more perspicuously below, in figure 1.

> *Newcomb Coin Toss.* The basic setup is the same as *Coin Toss.* A box contains either a $6,000 check or a $4,000 invoice. The content of the box was determined by a distributor, who tossed a fair coin. If the coin landed heads, the box contains the check. If the coin landed tails, the box contains the invoice. You have two options: you can buy the box for $3,000 or you can take it for free.
>
> But there are some additional details involving a predictor and a light. After the distributor determined the content of the box, the predictor predicted whether you would buy the box. So there are four possible cases. In the case in which the box contains the check and the predictor predicted that you would take the box for free,

as is two-boxing in *Viewcomb*, no conclusions about rationality can be drawn from facts about average welfare.

the predictor turned the light on. In the case in which the box contains the invoice and the predictor predicted that you would buy the box, the predictor turned the light off. In the other two cases, the predictor tossed a fair coin, turning the light on if the coin landed heads and off if it landed tails.

You can look at the light before deciding whether to buy the box. However, if you wish to make your decision straightaway, without first looking at the light, you must pay a fee of $2,000. So your options at stage one are to look at the light or pay the fee. And your options at stage two are to buy the box or take it for free.

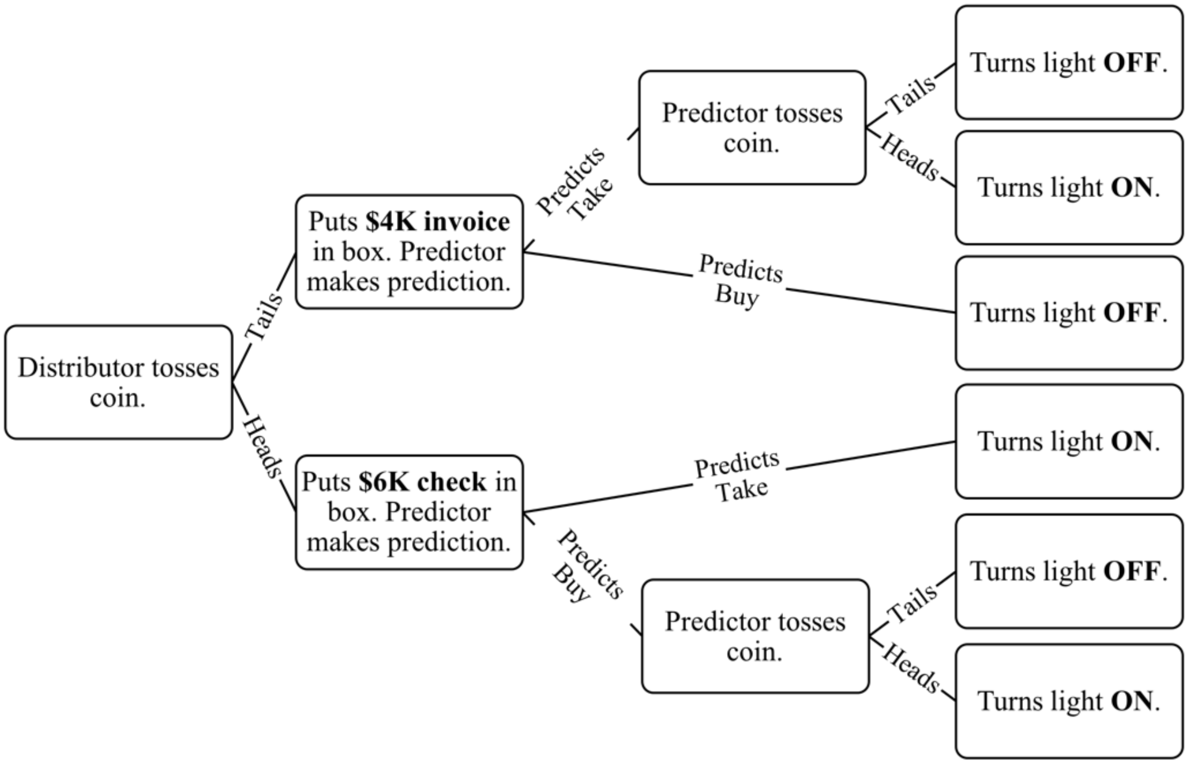

Figure 1: The *Newcomb Coin Toss* protocol. Causation flows from left to right.

Notice that the content of the box is settled, once and for all, before the predictor arrives on the scene. Moreover, the content of the box is determined entirely by the toss of a fair coin, just as it is in *Coin Toss*. So we can be assured that everyone who faces the problem is given an equal opportunity to succeed.

The remainder of this section draws out the consequences of evidential and causal decision theory in *Newcomb Coin Toss*. The discussion is informal. Those who wish to confirm the claims of this section may consult Appendix A, wherein all of the relevant expected values are calculated.

Let us begin with evidential decision theory. Recall Eva, the one-boxer in *Newcomb*. Let us assume that Eva is a reflective evidentialist in the following sense: at each stage of the problem she follows the advice of evidential decision theory and, at stage one, she believes that she will follow the advice of evidential decision theory at stage two. What does Eva do at stage one? Does she pay the fee or save her money and look at the light?

What Eva does at stage one depends on what she believes that she will do at stage two. Now, she knows that if she pays the fee, then at stage two she will take the box for free. After all, at stage one, she prefers that she takes the box for free at stage two, and she knows that if she pays the fee, her beliefs and desires will not change in any relevant way between the two stages. But what if she does not pay the fee? In that case, either she will learn that the light is on or she will learn that it is off.

Suppose that she learns that the light is on. Then her epistemic perspective on the outcome of the protocol is as pictured in figure 2. Let us assume for simplicity that Eva believes the predictor to be perfectly reliable. (This assumption is not essential to the problem, but it simplifies the presentation. It is relaxed in Appendix A.) We can imagine Eva reasoning evidentially as follows:

> Buying the box brings great news: a certain $3,000 gain. After all, buying the box signals that the predictor predicted as much, and the only open possibility in which the predictor made that prediction is one in which the box contains the check. Taking the box for free brings worse news: a possible $4,000 loss. After all, taking the box for free signals that the predictor predicted as much, and there is an open (1/3 likely) possibility in which the predictor made that prediction and put the invoice in the box. So I should buy the box.

Hence, if Eva learns that the light is on, she buys the box.

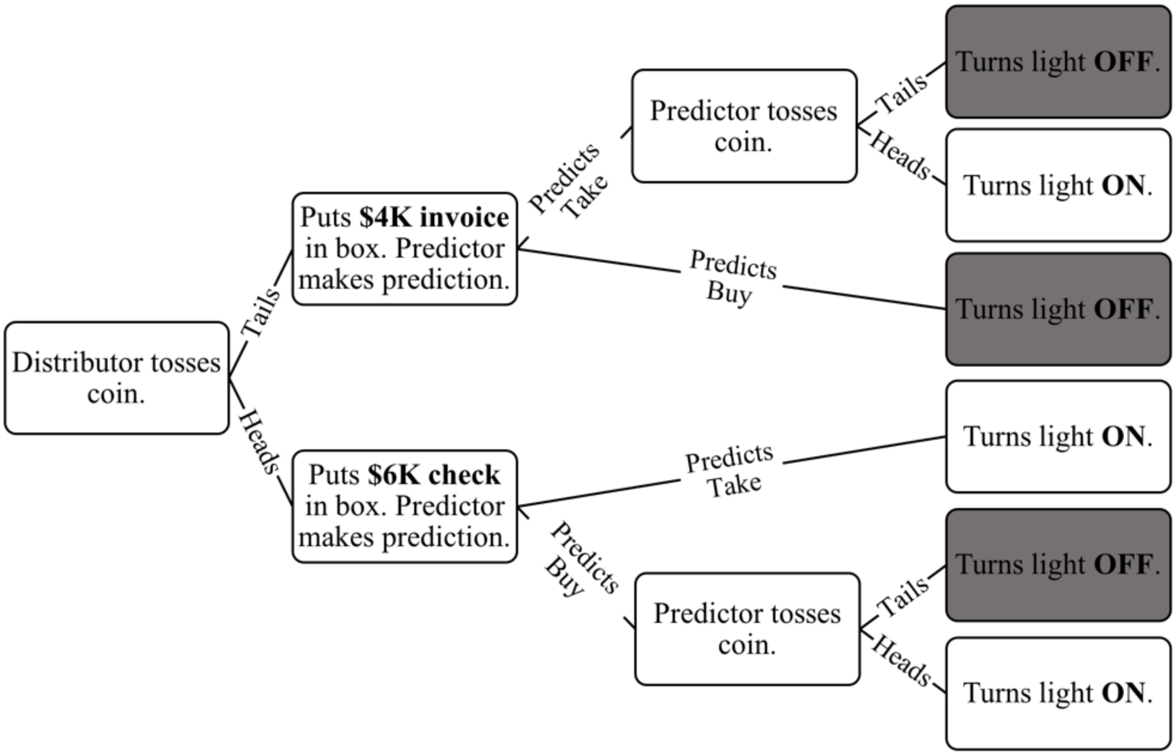

Figure 2: The agent's perspective on the *Newcomb Coin Toss* protocol after learning that the light is on. Grayed out nodes represent eliminated possibilities.

Suppose, on the other hand, that Eva learns that the light is off. Then her epistemic perspective on the outcome of the protocol is as pictured in figure 3. We can imagine Eva reasoning evidentially as follows:

> Taking the box for free brings bad news: a certain $4,000 loss. After all, taking the box for free signals that the predictor predicted as much, and the only open possibility in which the predictor made that prediction is one in which the box contains the invoice. Buying the box brings better news: a possible $3,000 gain. After all, buying the box signals that the predictor predicted as much, and there is an open (1/3 likely) possibility in which the predictor made that prediction and put the check in the box. So I should buy the box.

Hence, if Eva learns that the light is off, she buys the box.

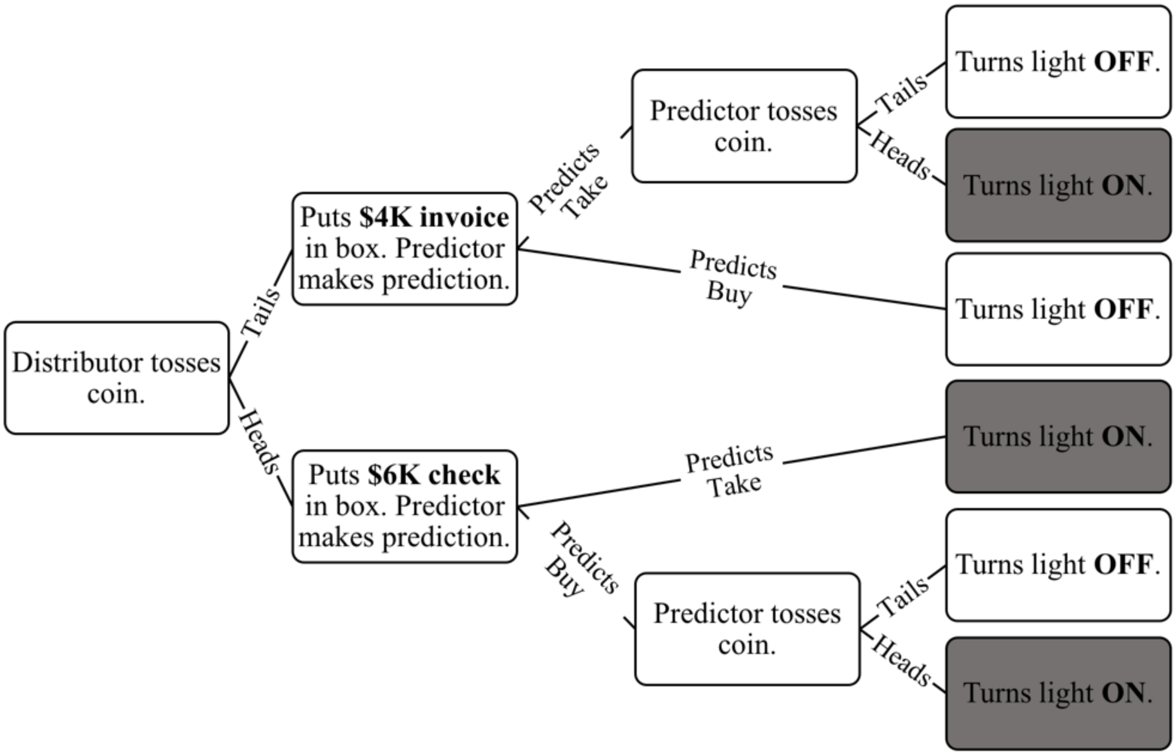

Figure 3: The agent's perspective on the *Newcomb Coin Toss* protocol after learning that the light is off. Grayed out nodes represent eliminated possibilities.

So if Eva does not pay the fee then, no matter what she learns about the light, she buys the box at stage two. Moreover, being reflective, Eva is in a position to know this about herself by reasoning in the way just described. Since at stage one Eva prefers that she does *not* buy the box at stage two, she believes that if she does not pay the fee, she will end up doing something that she hopes she will not do. So she has reason to pay the fee. The upshot: Eva pays $2,000 so that she will not fork over even more later.

Turn now to causal decision theory. Recall Casey, the two-boxer in *Newcomb*. Let us assume that Casey is a reflective causalist. What does Casey do at stage one?

What Casey does at stage one depends on what she believes that she will do at stage two. Like Eva, she knows that if she pays the fee, then at stage two she will take the box for free. After all, at stage one, she prefers that she takes the box for free at stage two, and she knows that if she pays the fee, her beliefs and desires will not change in any relevant way between the two stages.

But what if she does not pay the fee? In that case, either she will learn that the light is on or she will learn that it is off.

Suppose that Casey learns that the light is on. Then her epistemic perspective on the outcome of the protocol is as pictured in figure 2, and she reasons causally as follows:

> Since the light is on, the box probably contains the check. So if I were to take the box for free, I would probably gain $6,000 (and possibly lose $4,000). But if I were to buy the box, I would probably gain only $3,000 (and possibly lose $7,000). In any case, no matter what the box contains, I do better saving my money than giving it away. So I should take the box for free.

Hence, if Casey learns that the light is on, she takes the box for free.

Now suppose that Casey learns that the light is off. Then her epistemic perspective on the outcome of the protocol is as pictured in figure 3, and she reasons causally as follows:

> Since the light is off, the box probably contains the invoice. So if I were to take the box for free, I would probably lose $4,000 (and possibly gain $6,000). But if I were to buy the box, I would probably lose $7,000 (and possibly gain only $3,000). Again, no matter what the box contains, I do better saving my money than giving it away. So I should take the box for free.

Hence, if Casey learns that the light is off, she takes the box for free.

So if Casey does not pay the fee then, no matter what she learns about the light, she takes the box for free at stage two. Moreover, being reflective, Casey is in a position to know this about herself by reasoning in the way just described. Since at stage one Casey prefers that she takes the box for free at stage two, she believes that if she does not pay the fee, she will end up doing exactly what she now hopes she will do. So she has no reason to pay the fee. The upshot: at both stages, Casey keeps her money.

6 The Argument

Eva and Casey face *Newcomb Coin Toss* once a month for many months on end. Each month, Eva pays the fee at stage one and takes the box for free at stage two. Casey also takes the box for free at stage two, but she keeps her money at stage one. Eva loses an average of $1,000 per month, winning $4,000 half of the time and losing $6,000 the other half. Casey gains an average of $1,000 per month, winning $6,000 half of the time and losing $4,000 the other half. Casey's savings steadily grow while Eva's dwindle. As the years pass, the welfare gap between Eva and Casey widens.

Casey does well. Eva does poorly. Eva can anticipate all of this. We can imagine her thinking, 'If only I was a causalist like Casey, I could save my money at stage one and turn a profit in the long run. But I am not, so I must pay up and absorb a loss'. Eva envies Casey much as Casey envies Eva in *Newcomb*. The difference is that Casey's envy of Eva in *Newcomb* is like Green's envy of Blue in *Bluecomb*. Green wishes that she was Blue so that she could reap the benefits of the observer's partiality, but she does not wish that she made decisions as Blue did. Likewise, in *Newcomb*, Casey wishes that she was Eva so that she could reap the benefits of the predictor's partiality, but she does not wish that she made decisions as Eva did. The situation is different in *Newcomb Coin Toss*, where Eva envies Casey precisely because of the decisions Casey makes.

Our problem furnishes a new WAR argument.

Why Ain'cha Rich? (*Newcomb Coin Toss*)

(E1) In *Newcomb Coin Toss*, reflective agents who follow the advice of causal decision theory are richer, on average, than reflective agents who follow the advice of evidential decision theory.

(E2) The best explanation of (E1) is that causal decision theory gives rational advice and evidential decision theory gives irrational advice.

(E3) Therefore, in *Newcomb Coin Toss*, causal decision theory gives rational advice and evidential decision theory gives irrational advice.

I see no way out of the argument, so I say goodbye to evidential decision theory.

7 'Why Ya Poor?'

In 1981, at the height of Newcombmania, David Lewis reported that one-boxers had taken to taunting two-boxers in a now-familiar refrain: 'If you're so smart, why ain'cha rich?'[13] Lewis, a staunch two-boxer, searched for a riposte to his hecklers—a case in which the tables were turned—but came up empty. He regretfully concluded that there was none to be had.[14]

Lewis would be happy to hear that his conclusion was premature. Armed with *Newcomb Coin Toss*, causalists have a damning reply to their evidentialist opponents: 'You may be rich when the game is fixed, but when it's fair, why ya poor?'

8 Binding

Objection: Evidentialists lose money in *Newcomb Coin Toss* only if they lack the ability to bind themselves to a sequence of decisions. For suppose that the following option is available: *do not pay the fee and then take the box for free.* That option maximizes evidential expected value. Of course, evidentialists who take that option are just as well off, on average, as causalists, since they make the same decisions at every stage. So if binding is an option, (E1) is false.

[13] See Lewis (1981b). The term 'Newcombmania' is from Levi (1982).

[14] More carefully, Lewis concluded that the evidentialist can never be in the same position that the causalist is in, when she faces *Newcomb*: namely, the position of being certain at the time of decision that 'the irrational choice will, and the rational choice will not, be richly pre-rewarded'. If the evidentialist is certain that the putatively irrational choice will be richly pre-rewarded, then, by the design of evidential decision theory, that choice is the evidentially rational choice after all. Put this way, Lewis' conclusion is not at odds with the conclusion of this paper, since *Newcomb Coin Toss* is not a problem in which any one choice is richly pre-rewarded.

Reply: Distinguish two versions of *Newcomb Coin Toss*: a binding version and a non-binding version. To make the binding version vivid, let us imagine that the agent is to write her decisions at each stage on a card, give the card to a proxy, and then watch from within a soundproof glass holding cell as the proxy carries out the specified decisions for her. In the non-binding version, there is no proxy; the agent must carry out the decisions herself. Moreover, let us imagine that the agent has powerful evidence that at stage two she will make what she then takes to be the rational decision. For reasons articulated in Pollock (2002), a proposition represents an option for an agent only if the agent is certain that she will make the proposition true if she tries.[15] Hence, in the non-binding version, *do not pay the fee and then take the box for free* is not an option for the evidentialist, since she doubts that she will make it true if she tries. Although the binding version of *Newcomb Coin Toss* does not furnish a sound WAR argument against evidential decision theory, the non-binding version does. And one sound argument against the theory is one too many.

Rejoinder: I concede that evidentialists who lack the ability to bind themselves do poorly in *Newcomb Coin Toss*. But that is no mark against evidential decision theory. It is rather just a dramatization of the fact that the ability to bind oneself is sometimes a very helpful ability to have. This is the position of Arntzenius et al. (2004) and it is codified in the following principle:

> **The Binding Principle**.[16] If a decision theory has counterintuitive consequences that only arise for agents who lack the ability to bind themselves, these consequences are not a mark against the theory.

The Binding Principle suggests that (E2) is false: the best explanation of why evidentialists do poorly in the non-binding version of *Newcomb Coin Toss* is not that they make irrational decisions but rather that they lack the ability to bind themselves.

[15] Does *any* proposition meet this requirement? If not, classical decision theory is in trouble. But this problem is orthogonal to the debate between evidentialists and causalists, so we need not settle it here. See Pollock (2002) for a possible solution.

[16] This statement of the principle paraphrases Meacham (2010). My discussion of the principle follows Meacham's closely. For related discussion of binding, see Hedden (2015).

Reply: The proposed explanation strikes me as deeply unsatisfactory. The explanandum is not the claim that evidentialists do poorly. It is the comparative claim that evidentialists do *worse* than causalists. The explanans should therefore make reference to a difference between the two groups: e.g. evidentialists do worse because they, unlike causalists, lack the ability to bind themselves. But in the non-binding version of the problem causalists *also* lack the ability to bind themselves. Yet they still do better. It is implausible that the difference in welfare between the two groups is explained by a property that both groups have in common.

What of the Binding Principle? The principle is false. A decision problem is identified in part with a set of available options. Binding-enabled agents have different options available to them than do binding-disabled agents. So to say that evidential decision theory has counterintuitive consequences that only arise for binding-disabled agents is just to say that evidential decision theory has counterintuitive consequences in some decision problems but not others. Why should the fact that a theory lacks counterintuitive consequences in one problem do anything to mitigate the counterintuitive consequences of the theory in a different problem?

That said, there is a more plausible principle in the vicinity of the Binding Principle. Suppose that we are evaluating some decision theories by examining their consequences in a variety of decision problems. And suppose that we identify a problem in which *every* theory under evaluation has the same counterintuitive consequence. In that case, it may be unreasonable to hold one theory in particular accountable. But this consideration, rather than militating in favor of the Binding Principle, motivates a weaker principle:

> **The Weak Binding Principle.**[17] If *no* decision theory can avoid a given counterintuitive consequence without invoking binding options, then, for any *particular* decision theory, that consequence is not a mark against the theory.

But the Weak Binding Principle cannot save evidential decision theory from the threat of WAR (*Newcomb Coin Toss*). After all, there *is* a decision theory that avoids the counterintuitive

[17] This is the principle that Meacham (2010) calls '*Ought Implies Can (Binding)*'.

consequence of evidential decision theory in *Newcomb Coin Toss* without invoking binding: namely, causal decision theory.[18]

References

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

Ahmed, A. and Price, H. 2012. "Arntzenius on 'Why ain'cha rich?'." *Erkenntnis* 77:15–30.

Arntzenius, F. 2008. "No Regrets, or: Edith Piaf Revamps Decision Theory." *Erkenntnis* 68:277–297.

Arntzenius, F., Elga, A., and Hawthorne, J. 2004. "Bayesianism, Infinite Decisions, and Binding." *Mind* 113:251–283.

Briggs, R. 2010. "Decision-Theoretic Paradoxes as Voting Paradoxes." *The Philosophical Review* 119:1–30.

Buchak, L. 2010. "Instrumental Rationality, Epistemic Rationality, and Evidence-Gathering." *Philosophical Perspectives* 24:85–120.

[18] Binding versions of *Newcomb Coin Toss* and *Viewcomb* highlight a separate worry for evidential decision theory: the worry that the theory is not 'self-recommending' (Skyrms (1982)). After all, if the agent is able to ensure that she follows the advice of causal decision theory in *Newcomb Coin Toss*, evidential decision theory recommends that she do so. And if the agent is able to ensure that she follows the advice of the insane theory in *Viewcomb*, evidential decision theory recommends that she do so. Now, in some cases, failures of self-recommendation are to be expected. For example, if an agent believes that she will become irrational, lose information, or change her values, then it may make sense for her to ensure that she does not follow the advice of her preferred theory. Similarly, Arntzenius et al. (2004) describe cases involving infinite sequences of decisions in which it seems rational for the agent to ensure that she does not behave rationally. The worry for evidential decision theory is that neither *Newcomb Coin Toss* nor *Viewcomb* is anything like these problems. In both, the agent faces a finite sequence of decisions. And in both, the agent believes that she will remain rational, that she will not lose any information, and that her values will remain unchanged throughout the sequence of decisions. Therefore, evidential decision theory's failure of self-recommendation in these cases is anomalous.

—. 2012. "Can it be Rational to have Faith?" In J. Chandler and V. Harrison (eds.), *Probability in the Philosophy of Religion*. Oxford University Press.

Burgess, S. 2004. "The Newcomb Problem: An Unqualified Resolution." *Synthese* 138:261–287.

Eells, E. 1982. *Rational Decision and Causality*. Cambridge University Press.

Gibbard, A. and Harper, W. 1978. "Counterfactuals and Two Kinds of Expected Utility." In Leach J. Hooker, A. and E. McClennen (eds.), *Foundations and Applications of Decision Theory*, 125–162. Reidel.

Good, I. J. 1967. "On the Principle of Total Evidence." *British Journal for the Philosophy of Science* 17:319–321.

Hare, C. and Hedden, B. 2016. "Self-Reinforcing and Self-Frustrating Decisions." *Noûs* 50:604–628.

Hedden, B. 2015. *Reasons without Persons*. Cambridge University Press.

Jeffrey, R. 1965. *The Logic of Decision*. University of Chicago Press.

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

Levi, I. 1982. "A Note on Newcombmania." *The Journal of Philosophy* 79:337–342.

Lewis, D. 1981a. "Causal Decision Theory." *Australasian Journal of Philosophy* 59:5–30.

—. 1981b. "'Why ain'cha rich?'." *Noûs* 15:377–380.

Meacham, C. 2010. "Binding and Its Consequences." *Philosophical Studies* 149:49–71.

Nozick, R. 1969. "Newcomb's Problem and Two Principles of Choice." In N. Rescher (ed.), *Essays in Honor of Carl G. Hempel*, 114–146. Reidel.

Pollock, J. 2002. "Rational Choice and Action Omnipotence." *Philosophical Review* 111:1–23.

Price, H. 1986. "Against Causal Decision Theory." *Synthese* 67:195–212.

Skyrms, B. 1982. "Causal Decision Theory." *Journal of Philosophy* 79:695–711.

—. 1990. "The Value of Knowledge." *Minnesota Studies in the Philosophy of Science* 14:245–266.

Appendix A: Calculations

This appendix confirms the claims of Section 5. I associate an agent at a time with a probability function P, representing the agent's rational degrees of belief at the time, and a utility function u, representing the agent's non-instrumental desires at the time. Let $\mathcal{A} = \{A_1, ..., A_n\}$ be a finite partition of propositions representing the agent's options at the time. Let $\mathcal{K} = \{K_1, ..., K_m\}$ be a finite partition of propositions representing the agent's possible circumstances (i.e. dependency hypotheses). I assume that for each $A_j \in \mathcal{A}$ and each $K_i \in \mathcal{K}$ the conjunction $A_j K_i$ entails a unique outcome—the outcome that would result if the agent were to choose option A_j in circumstance K_i—and that the set of all such outcomes is the domain of u. The evidential expected value V of an option $A_j \in \mathcal{A}$ can now be defined as follows:

$$V(A_j) = \sum_{i=1}^{m} P(K_i \mid A_j) u(A_j K_i).$$

Evidentialists do not define V in terms of circumstances or dependency hypotheses, since those concepts are explicitly causal. But, as Lewis (1981a) observes and Briggs (2010) proves, the evidentialist definition is equivalent to the one above.

The causal expected value U of an option $A_j \in \mathcal{A}$ is defined as follows:

$$U(A_j) = \sum_{i=1}^{m} P(K_i) u(A_j K_i).$$

Some abbreviations will help streamline the presentation. Let F be the proposition that the agent pays the fee at stage one. Let B be the proposition that the agent buys the box at stage two. Let C be the proposition that the check is in the box. Let L be the proposition that the light is on. Let $\mathscr{B}$ be the proposition that B was predicted. I will assume that learning goes by conditionalization and that the predictor is believed to be 99% reliable.

Claim 1. After Eva learns that the light is on, her evidential expected value of buying the box exceeds her evidential expected of taking it for free.

Proof: Let P be Eva's probability function at stage one, before she sees the light. Let V_L be Eva's evidential expected value function after she sees the light and learns L. Then we have:

$$V_L(B) = P(C \mid BL)(3000) - P(\overline{C} \mid BL)(7000).$$
$$V_L(\overline{B}) = P(C \mid \overline{B}L)(6000) - P(\overline{C} \mid \overline{B}L)(4000).$$

Next we calculate the probabilities.

Subclaim 1.1: $P(C \mid BL) = .99$. *Proof*: By the law of total conditional probability (hereafter, TCP),

$$P(C \mid BL) = P(C \mid BL\mathscr{B})P(\mathscr{B} \mid BL) + P(C \mid BL\overline{\mathscr{B}})P(\overline{\mathscr{B}} \mid BL).$$

By the description of the problem, $P(C \mid BL\mathscr{B}) = 1$ and $P(C \mid BL\overline{\mathscr{B}}) = .67$. By the definition of conditional probability (hereafter, CP) and algebra,

$$P(\mathscr{B} \mid BL) = \frac{P(\mathscr{B}L \mid B)}{P(L \mid B)}.$$

By TCP and the description of the problem,

$$\begin{aligned}P(\mathscr{B}L \mid B) &= P(\mathscr{B}L \mid B\mathscr{B})P(\mathscr{B} \mid B) + P(\mathscr{B}L \mid B\overline{\mathscr{B}})P(\overline{\mathscr{B}} \mid B)\\ &= (.25)(.99) + (0)(.01) = .2475.\\ P(L \mid B) &= P(L \mid B\mathscr{B})P(\mathscr{B} \mid B) + P(L \mid B\overline{\mathscr{B}})P(\overline{\mathscr{B}} \mid B)\\ &= (.25)(.99) + (.75)(.01) = .255.\end{aligned}$$

Hence, $P(\mathscr{B} \mid BL) = \frac{.2475}{.255} = .97$. Substituting values,

$$P(C \mid BL) = (1)(.97) + (.67)(.03) = .99.$$

Subclaim 1.2: $P(C \mid \overline{B}L) = .67$. *Proof*: By TCP,

$$P(C \mid \overline{B}L) = P(C \mid \overline{B}L\mathscr{B})P(\mathscr{B} \mid \overline{B}L) + P(C \mid \overline{B}L\overline{\mathscr{B}})P(\overline{\mathscr{B}} \mid \overline{B}L).$$

By the description of the problem, $P(C \mid \overline{B}L\mathscr{B}) = 1$ and $P(C \mid \overline{B}L\overline{\mathscr{B}}) = .67$. By CP and algebra,

$$P(\mathscr{B} \mid \overline{B}L) = \frac{P(\mathscr{B}L \mid \overline{B})}{P(L \mid \overline{B})}.$$

By TCP and the description of the problem,

$$\begin{aligned} P(\mathscr{B}L \mid \overline{B}) &= P(\mathscr{B}L \mid \overline{B}\mathscr{B})P(\mathscr{B} \mid \overline{B}) + P(\mathscr{B}L \mid \overline{B}\overline{\mathscr{B}})P(\overline{\mathscr{B}} \mid \overline{B}) \\ &= (.25)(.01) + (0)(.99) = .0025. \\ P(L \mid \overline{B}) &= P(L \mid \overline{B}\mathscr{B})P(\mathscr{B} \mid \overline{B}) + P(L \mid \overline{B}\overline{\mathscr{B}})P(\overline{\mathscr{B}} \mid \overline{B}) \\ &= (.25)(.01) + (.75)(.99) = .745. \end{aligned}$$

Hence, $P(\mathscr{B} \mid \overline{B}L) = \frac{.0025}{.745} = .003$. Substituting values,

$$P(C \mid \overline{B}L) = (1)(.003) + (.67)(.997) = .67.$$

Plugging in the probabilities:

$$V_L(B) = (.99)(3000) - (.01)(7000) = 2900.$$

$$V_L(\overline{B}) = (.67)(6000) - (.33)(4000) = 2700.$$

Hence, $V_L(B) > V_L(\overline{B})$. ■

Claim 2. After Eva learns that the light is off, her evidential expected value of buying the box exceeds her evidential expected value of taking it for free.

Proof: Eva's evidential expected values after learning $\overline{L}$ are:

$$V_{\overline{L}}(B) = P(C \mid B\overline{L})(3000) - P(\overline{C} \mid B\overline{L})(7000).$$

$$V_{\overline{L}}(\overline{B}) = P(C \mid \overline{BL})(6000) - P(\overline{C} \mid \overline{BL})(4000).$$

The proof proceeds as before.

Subclaim 2.1: $P(C \mid B\overline{L}) = .33$. *Proof*: By TCP,

$$P(C \mid B\overline{L}) = P(C \mid B\overline{L}\mathscr{B})P(\mathscr{B} \mid B\overline{L}) + P(C \mid B\overline{L\mathscr{B}})P(\overline{\mathscr{B}} \mid B\overline{L}).$$

By the description of the problem, $P(C \mid B\overline{L}\mathscr{B}) = .33$ and $P(C \mid B\overline{L\mathscr{B}}) = 0$. By CP and algebra,

$$P(\mathscr{B} \mid B\overline{L}) = \frac{P(\mathscr{B}\overline{L} \mid B)}{P(\overline{L} \mid B)}.$$

By TCP and the description of the problem,

$$\begin{aligned} P(\mathscr{B}\bar{L} \mid B) &= P(\mathscr{B}\bar{L} \mid B\mathscr{B})P(\mathscr{B} \mid B) + P(\mathscr{B}\bar{L} \mid B\overline{\mathscr{B}})P(\overline{\mathscr{B}} \mid B) \\ &= (.75)(.99) + (0)(.01) = .7425. \\ P(\bar{L} \mid B) &= P(\bar{L} \mid B\mathscr{B})P(\mathscr{B} \mid B) + P(\bar{L} \mid B\overline{\mathscr{B}})P(\overline{\mathscr{B}} \mid B) \\ &= (.75)(.99) + (.25)(.01) = .745. \end{aligned}$$

Hence, $P(\mathscr{B} \mid B\bar{L}) = \frac{.7425}{.745} = .997$. Substituting values,

$$P(C \mid B\bar{L}) = (.33)(.997) + (0)(.003) = .33.$$

Subclaim 2.2: $P(C \mid \overline{BL}) = .01$. *Proof*: By TCP and the description of the problem,

$$\begin{aligned} P(C \mid \overline{BL}) &= P(C \mid \overline{BL}\mathscr{B})P(\mathscr{B} \mid \overline{BL}) + P(C \mid \overline{BL\overline{\mathscr{B}}})P(\overline{\mathscr{B}} \mid \overline{BL}) \\ &= P(C \mid \overline{BL}\mathscr{B})P(\mathscr{B} \mid \overline{BL}). \end{aligned}$$

By CP and algebra,

$$P(\mathscr{B} \mid \overline{BL}) = \frac{P(\mathscr{B}\bar{L} \mid \bar{B})}{P(\bar{L} \mid \bar{B})}.$$

By TCP and the description of the problem,

$$\begin{aligned} P(\mathscr{B}\bar{L} \mid \bar{B}) &= P(\mathscr{B}\bar{L} \mid \bar{B}\mathscr{B})P(\mathscr{B} \mid \bar{B}) + P(\mathscr{B}\bar{L} \mid \overline{B\overline{\mathscr{B}}})P(\overline{\mathscr{B}} \mid \bar{B}) \\ &= (.75)(.01) + (0)(.99) = .0075. \\ P(\bar{L} \mid B) &= P(\bar{L} \mid \bar{B}\mathscr{B})P(\mathscr{B} \mid \bar{B}) + P(\bar{L} \mid \overline{B\overline{\mathscr{B}}})P\overline{\mathscr{B}} \mid \bar{B}) \\ &= (.75)(.01) + (.25)(.99) = .255. \end{aligned}$$

Hence, $P(\mathscr{B} \mid \overline{BL}) = \frac{.0075}{.255} = .03$. Substituting values, $P(C \mid \overline{BL}) = (.33)(.03) = .01$.

Plugging in the probabilities:

$$V_{\bar{L}}(B) = (.33)(3000) - (.67)(7000) = -3700.$$

$$V_{\bar{L}}(\overline{B}) = (.01)(6000) - (.99)(4000) = -3900.$$

Hence, $V_{\bar{L}}(B) > V_{\bar{L}}(\overline{B})$. ■

Claim 3. Before seeing the light, Eva prefers that she takes the box for free at stage two.

Proof: Note that $P(C \mid B) = P(C \mid \overline{B}) = .5$. After all, by TCP and the description of the problem,

$$\begin{aligned} P(C \mid B) &= P(C \mid B\mathscr{B})P(\mathscr{B} \mid B) + P(C \mid B\overline{\mathscr{B}})P(\overline{\mathscr{B}} \mid B) \\ &= (.5)(.99) + (.5)(.01) = .5. \end{aligned}$$

Analogous reasoning shows that $P(C \mid \overline{B}) = .5$ as well. Then we have:

$$\begin{aligned} V(B) &= P(C \mid B)(3000) - P(\overline{C} \mid B)(7000) \\ &= (.5)(3000) - (.5)(7000) = -2000. \\ V(\overline{B}) &= P(C \mid \overline{B})(6000) - P(\overline{C} \mid \overline{B})(4000) \\ &= (.5)(6000) - (.5)(4000) = 1000. \end{aligned}$$

Hence, $V(\overline{B}) > V(B)$. ■

Claim 4. At stage one, Eva's evidential expected value of paying the fee exceeds her evidential expected value of not paying the fee.

Proof: Since Eva is reflective, she knows that she will buy the box at stage two if and only if she does not pay the fee at stage one. And she knows that she will take the box for free if and only if she pays the fee. Hence, $P(C \mid F) = P(C \mid \overline{B})$ and $P(C \mid \overline{F}) = P(C \mid B)$. We already showed that $P(C \mid B) = P(C \mid \overline{B}) = .5$. Hence, $P(C \mid F) = P(C \mid \overline{F}) = .5$. So we have:

$$\begin{aligned} V(F) &= P(C \mid F)(4000) - P(\overline{C} \mid F)(6000) \\ &= (.5)(4000) - (.5)(6000) = -1000. \\ V(\overline{F}) &= P(C \mid \overline{F})(3000) - P(\overline{C} \mid \overline{F})(7000) \\ &= (.5)(3000) - (.5)(7000) = -2000. \end{aligned}$$

Hence, $V(F) > V(\overline{F})$. ■

Claim 5. After Casey learns that the light is on, her causal expected value of taking the box for free exceeds her causal expected value of buying the box.

Proof: Let P be Casey's probability function at stage one, before she sees the light. Let U_L be Casey's causal expected value function after she sees the light and learns L. Then we have:

$$\begin{aligned} U_L(B) &= P(C \mid L)(3000) - P(\overline{C} \mid L)(7000). \\ U_L(\overline{B}) &= P(C \mid L)(6000) - P(\overline{C} \mid L)(4000). \end{aligned}$$

It is straightforward to see that, given any probability function, $U_L(\overline{B})$ exceeds $U_L(B)$ by exactly 3,000. Let $P(C \mid L) = x$. Then we have:

$$\begin{aligned} U_L(B) &= 3000x - (1 - x)(7000) \\ &= 10000x - 7000. \\ U_L(\overline{B}) &= 6000x - (1 - x)(4000) \\ &= 10000x - 4000. \end{aligned}$$

Hence, $U_L(\overline{B}) > U_L(B)$. ■

Claim 6. After Casey learns that the light is off, her causal expected value of taking the box for free exceeds her causal expected value of buying the box.

Proof: Let $U_{\overline{L}}$ be Casey's causal expected value function after she sees the light and learns $\overline{L}$. Then we have:

$$U_{\overline{L}}(B) = P(C \mid \overline{L})(3000) - P(\overline{C} \mid \overline{L})(7000).$$
$$U_{\overline{L}}(\overline{B}) = P(C \mid \overline{L})(6000) - P(\overline{C} \mid \overline{L})(4000).$$

As before, $U_L(\overline{B})$ exceeds $U_L(B)$ by 3,000, given any probability function. ■

Claim 7. At stage one, Casey prefers that at stage two she takes the box for free.

Proof: By the description of the case, $P(C) = .5$. The causal expected values follow straightforwardly:

$$\begin{aligned} U(B) &= P(C)(3000) - P(\overline{C})(7000) \\ &= (.5)(3000) - (.5)(7000) = -2000. \\ U(\overline{B}) &= P(C)(6000) - P(\overline{C})(4000) \\ &= (.5)(6000) - (.5)(4000) = 1000. \end{aligned}$$

Hence, $U(\overline{B}) > U(B)$. ■

Claim 8. At stage one, Casey's causal expected value of not paying the fee exceeds her causal expected value of paying the fee.

Proof: Since Casey is reflective, she knows that she will take the box for free at stage two no matter what she does at stage one. Hence,

$$\begin{aligned} U(F) &= P(C)(4000) - P(\overline{C})(6000) \\ &= (.5)(4000) - (.5)(6000) = -1000. \\ U(\overline{F}) &= P(C)(6000) - P(\overline{C})(4000) \\ &= (.5)(6000) - (.5)(4000) = 1000. \end{aligned}$$

Hence, $U(\overline{F}) > U(F)$. ■

Appendix B: Arntzenius

Arntzenius (2008) poses a problem in which, he claims, causalists are richer, on average, than evidentialists. Here is his problem:

> *Red Sox vs Yankees.* The Red Sox and Yankees will play each other in a long series of games. The Yankees win 90% of such games. Before each game, you will be offered two bets, of which you must pick exactly one: you can either bet on the Yankees at 1:2 odds (risk \$2 to win \$1) or bet on the Red Sox at 2:1 odds (risk \$1 to win \$2). However, before placing your bet, a perfect predictor of your decisions and the outcomes of the games will tell you whether you will win or lose your bet.

Consider an agent facing *Red Sox vs Yankees.* Let P be her probability function before the series begins and let u be her utility function at that time. Consider some arbitrary game in the series and let $\mathscr{Y}$ be the proposition that the Yankees win that game. Let Y be the proposition that the agent bets on the Yankees. Let W be the proposition that the agent wins her bet: $W \equiv (Y\mathscr{Y} \vee \overline{Y}\overline{\mathscr{Y}})$.

Let V_W be the agent's evidential expected value function after learning W. Then we have:

$$
\begin{aligned}
V_W(Y) &= P(\mathscr{Y} \mid YW)u(Y\mathscr{Y}) + P(\overline{\mathscr{Y}} \mid YW)u(Y\overline{\mathscr{Y}}) \\
&= (1)(1) + (0)(-2) = 1. \\
V_W(\overline{Y}) &= P(\mathscr{Y} \mid \overline{Y}W)u(\overline{Y}\mathscr{Y}) + P(\overline{\mathscr{Y}} \mid \overline{Y}W)u(\overline{Y}\overline{\mathscr{Y}}) \\
&= (0)(-1) + (1)(2) = 2.
\end{aligned}
$$

Since $V_W(\overline{Y}) > V_W(Y)$, evidential decision theory advises betting on the Red Sox.

Let $V_{\overline{W}}$ be the agent's evidential expected value function after learning $\overline{W}$. Then we have:

$$\begin{aligned}
V_{\overline{W}}(Y) &= P(\mathscr{Y} \mid Y\overline{W})u(Y\mathscr{Y}) + P(\overline{\mathscr{Y}} \mid Y\overline{W})u(Y\overline{\mathscr{Y}}) \\
&= (0)(1) + (1)(-2) = -2. \\
V_{\overline{W}}(\overline{Y}) &= P(\mathscr{Y} \mid \overline{Y}\overline{W})u(\overline{Y}\mathscr{Y}) + P(\overline{\mathscr{Y}} \mid \overline{Y}\overline{W})u(\overline{Y}\overline{\mathscr{Y}}) \\
&= (1)(-1) + (0)(2) = -1.
\end{aligned}$$

Since $V_{\overline{W}}(\overline{Y}) > V_{\overline{W}}(Y)$, evidential decision theory advises betting on the Red Sox. Hence, evidential decision theory always advises betting on the Red Sox. But, by hypothesis, the Yankees win 90% of the games. So it seems that agents who follow the advice of evidential decision theory lose \$1 90% of the time and win \$2 10% of the time, for an average loss of 70¢ per game.

Turn now to causal decision theory. Arntzenius says that causal decision theory always advises betting on the Yankees. That it does depends, as Arntzenius acknowledges in a footnote, on the assumption that the agent has no beliefs about which bet she will make. I will return to this assumption shortly. For now, let us take it on board.

Let U be the agent's causal expected value function after receiving the predictor's information. On its own, the information provided by the predictor is evidentially irrelevant to the outcome of the game. So we have:

$$\begin{aligned}
U(Y) &= P(\mathscr{Y})u(Y\mathscr{Y}) + P(\overline{\mathscr{Y}})u(Y\overline{\mathscr{Y}}) \\
&= (.9)(1) + (.1)(-2) = .7. \\
U(\overline{Y}) &= P(\mathscr{Y})u(\overline{Y}\mathscr{Y}) + P(\overline{\mathscr{Y}})u(\overline{Y}\overline{\mathscr{Y}}) \\
&= (.9)(-1) + (.1)(2) = -.7.
\end{aligned}$$

Since $U(Y) > U(\overline{Y})$, causal decision theory always advises betting on the Yankees. And since the Yankees win 90% of the games, agents who follow the advice of causal decision theory gain \$1 90% of the time and lose \$2 10% of the time, for an average gain of 70¢ per game. So *Yankees* *vs Red Sox* seems to furnish a WAR argument against evidential decision theory.

Why Ain'cha Rich? (*Red Sox vs Yankees*)

(F1) In *Red Sox vs Yankees*, agents who follow the advice of causal decision theory are richer, on average, than agents who follow the advice of evidential decision theory.

(F2) The best explanation of (F1) is that causal decision theory gives rational advice and evidential decision theory gives irrational advice.

(F3) Therefore, in *Red Sox vs Yankees*, causal decision theory gives rational advice and evidential decision theory gives irrational advice.

However, as Ahmed and Price (2012), Ahmed (2014) and Hare and Hedden (2016) have shown, the argument breaks down on closer examination. The advice of evidential decision theory depends on the epistemic perspective of the agent being advised. There are two possible perspectives: that of the agent before receiving the predictor's information and that of the agent after receiving the information. *Ex ante*, the agent expects betting on the Red Sox to fare worse than betting on the Yankees, but evidential decision theory does not advise betting on the Red Sox. The theory only gives such advice once the agent receives the predictor's information. *Ex post*, evidential decision theory advises betting on the Red Sox, but the agent expects betting on the Red Sox to fare better than betting on the Yankees. After all, among those who learn that they will win their bet, those who bet on the Red Sox always gain $2 while those who bet on the Yankees always gain $1, and among those who learn that they will lose their bet, those who bet on the Red Sox always lose $1 while those who bet on the Yankees always lose $2. Either way, then, (F1) is false. The premise only appears to be true if we illicitly assume both perspectives at the same time.

We might try to revive Arntzenius' argument by modifying *Red Sox vs Yankees* so that it more closely resembles *Newcomb Coin Toss*.

> *Red Sox vs Yankees (Fee Version).* Everything is the same as in *Red Sox vs Yankees*, only now you can pay a small fee before the series begins to silence the predictor for the duration of the series.

Consider Eva, the reflective evidentialist. Eva pays the fee and then bets on the Yankees every game. After all, before the series begins, she expects that if she eschews the fee she will always bet on the Red Sox, losing an average of 70¢ per game, while if she pays it she will always bet on the Yankees, gaining an average of 70¢ per game. The one-time fee is a small price to pay to replace a future of failure with a future of success.

Now consider Casey, the reflective causalist. If Casey could anticipate always betting on the Yankees, even in the event that she eschews the fee, she would eschew the fee. And if she did, a version of Arntzenius' argument (focused on the *ex ante* perspective of a reflective agent) would be revived. But, being reflective, Casey cannot anticipate always betting on the Yankees.

Proof: Suppose that Casey eschews the fee. Then there is a possibility in which she learns that she will lose her bet. Let $U_{\overline{W}}$ be Casey's causal expected value function after learning $\overline{W}$. Suppose, for reductio, that $U_{\overline{W}}(Y) > U_{\overline{W}}(\overline{Y})$. It is straightforward that $U_{\overline{W}}(Y) > U_{\overline{W}}(\overline{Y})$ if and only if $P(\mathscr{Y} \mid \overline{W}) > 2/3$. Hence, $P(\mathscr{Y} \mid \overline{W}) > 2/3$. Now, Casey's beliefs about the outcome of the game are related to her beliefs about which bet she will take by the following equation:

$$\begin{aligned} P(\mathscr{Y} \mid \overline{W}) &= P(\mathscr{Y} \mid \overline{W}Y)P(Y \mid \overline{W}) + P(\mathscr{Y} \mid \overline{W}\overline{Y})P(\overline{Y} \mid \overline{W}) \\ &= (0)P(Y \mid \overline{W}) + (1)P(\overline{Y} \mid \overline{W}) \\ &= 1 - P(Y \mid \overline{W}). \end{aligned}$$

Hence, $P(\mathscr{Y} \mid \overline{W}) > 2/3$ if and only if $P(Y \mid \overline{W}) < 1/3$. Hence, $P(Y \mid \overline{W}) < 1/3$. But Casey is reflective: she believes that she will choose the option that maximizes causal expected value—which, by hypothesis, is betting on the Yankees. So $P(Y \mid \overline{W})$ is high, for a contradiction. Discharging the assumption, $U_{\overline{W}}(Y) \not> U_{\overline{W}}(\overline{Y})$. ■

So if Casey eschews the fee, she does not always bet on the Yankees.[19] Moreover, being reflective, Casey is in a position to know this about herself by reasoning in the way just described. So she pays the fee to ensure that she will always bet on the Yankees. And she ends up no better off than Eva.

The reason that Arntzenius' argument cannot be made to work is that, after the agent learns that her bet will lose, causal decision theory gives different advice depending on the agent's beliefs about which bet she will make. The same problem does not arise in *Newcomb Coin Toss*. As we saw in Appendix A, causal decision theory advises taking the box for free no matter what probability function the agent has after learning about the light. So the assumption that the agent is a reflective causalist is in perfect harmony with the claim that causal decision theory always advises taking the box for free.

[19] Does that mean that Casey sometimes bets on the Red Sox? Not necessarily. After Casey learns that she will lose her bet, the advice of causal decision theory becomes unstable, much as it does in Gibbard and Harper (1978)'s famous *Death in Damascus* problem: if Casey believes that she will bet on the Yankees, causal decision theory advises betting on the Red Sox; and if she believes that she will bet on the Red Sox, causal decision theory advises betting on the Yankees. Perhaps, then, Casey should expect that if she eschews the fee, she will sometimes perpetually oscillate between betting on the Yankees and betting on the Red Sox. No such instability arises in *Newcomb Coin Toss* because taking the box for free dominates buying the box.