Evidence and Choice

by

Ian Wells

B.A., Cornell University (2011)

Submitted to the Department of Linguistics and Philosophy
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2017

Signature redacted

Author .....
Department of Linguistics and Philosophy
August 7, 2017

Signature redacted

Certified by .....
Roger White
Professor of Philosophy
Thesis Supervisor

Signature redacted

Accepted by .....
Roger White
Chair of the Committee on Graduate Students

Evidence and Choice

by

Ian Wells

Submitted to the Department of Linguistics and Philosophy
on August 7, 2017, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

This dissertation defends causal decision theory and argues against its main rival, evidential decision theory. In Chapter 1, I introduce a decision problem in which evidentialists end up predictably worse off, on average, than causalists. This result is surprising since comparisons of average welfare have traditionally been taken to support evidential decision theory and undermine causal decision theory. In Chapter 2, Jack Spencer and I give a new argument for one of causal decision theory's distinctive recommendations: two-boxing in Newcomb's problem. Unlike arguments based on causal dominance, our argument relies on a more basic principle connecting rational choice to guidance and actual value maximization. In Chapter 3, I take up the issue of rationalization. Is it possible to manipulate the demands of rationality in predictable ways? I argue that it is not. Then I show that if evidential decision theory is true, rationalization is not only possible but sometimes advisable.

Thesis Supervisor: Roger White
Title: Professor of Philosophy

Acknowledgments

I have benefited from helpful discussions with many people, including but not limited to Adam Bales, Nilanjan Das, Ryan Doody, Kevin Dorst, Lyndal Grant, Cosmo Grant, Justin Khoo, Matthew Mandelkern, Milo Phillips-Brown, Kevin Richardson, Said Saillant, Bernhard Salow, Ginger Schultheis and Judy Thomson.

For thoughtful feedback on Chapter 2, Jack Spencer and I thank Arif Ahmed, Sara Aronowitz, Dave Chalmers, Mikaël Cozic, Paul Egré, Branden Fitelson, Brian Hedden, Wes Holliday, Uriah Kriegel, David Nicolas, François Recanati, Brad Skow, Quinn White and the members of the Rutgers Formal Epistemology and Decision Theory Reading Group.

I am greatly indebted to the members of my dissertation committee—Caspar Hare, Bob Stalnaker and Roger White—for the time and effort that they invested in helping improve this dissertation. Early conversations with Caspar were particularly important in the development of Chapters 1 and 3.

I owe a special thanks to Steve Yablo, for countless, invaluable discussions over the years, and to Jack Spencer, who has been not only a coauthor but a mentor, friend, and committee member in everything but name.

My deepest gratitude is to my fiancée, Molly Johnson, for her continuous support and encouragement.

Contents

Chapter 1

Equal Opportunity and Newcomb's Problem

1.1 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.

1.2 '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 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*.

---

[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 (2014a). 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.

[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 2014a, p. 194). Hare and Hedden (2016) appeal to a similar argument against causal decision theory based

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 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

on a non-standard Newcomb case. My criticism of the standard argument applies to their argument as well.

[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).

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*).

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.

[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*.

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.

1.3 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 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 prising 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]

[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 1981, 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.

[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.

1.4 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:

> 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

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

[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 (2015b).

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]

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.

1.5 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 1-1.

> *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

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

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, 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.

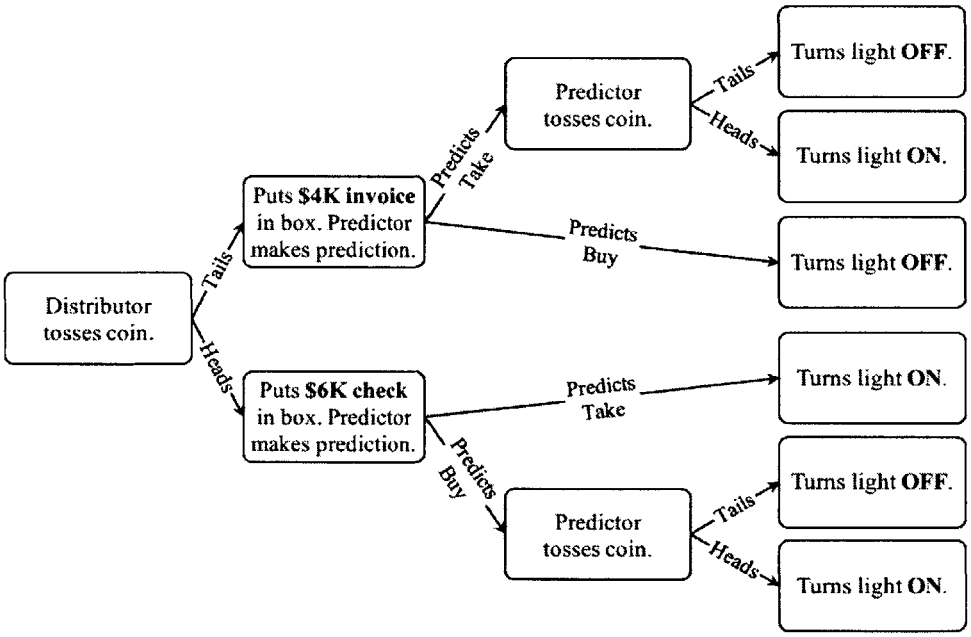

Figure 1-1: The *Newcomb Coin Toss* protocol.

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 1-2. 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.

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 1-3. We can imagine Eva reasoning evidentially as follows:

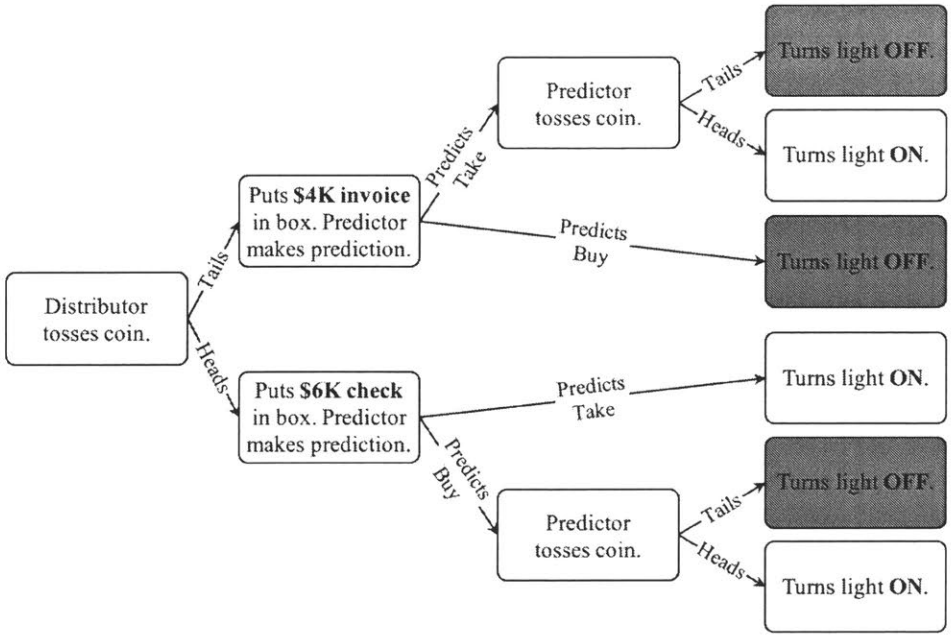

Figure 1-2: The agent's perspective on the *Newcomb Coin Toss* protocol after learning that the light is on. Grayed out nodes represent eliminated possibilities.

> 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.

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.

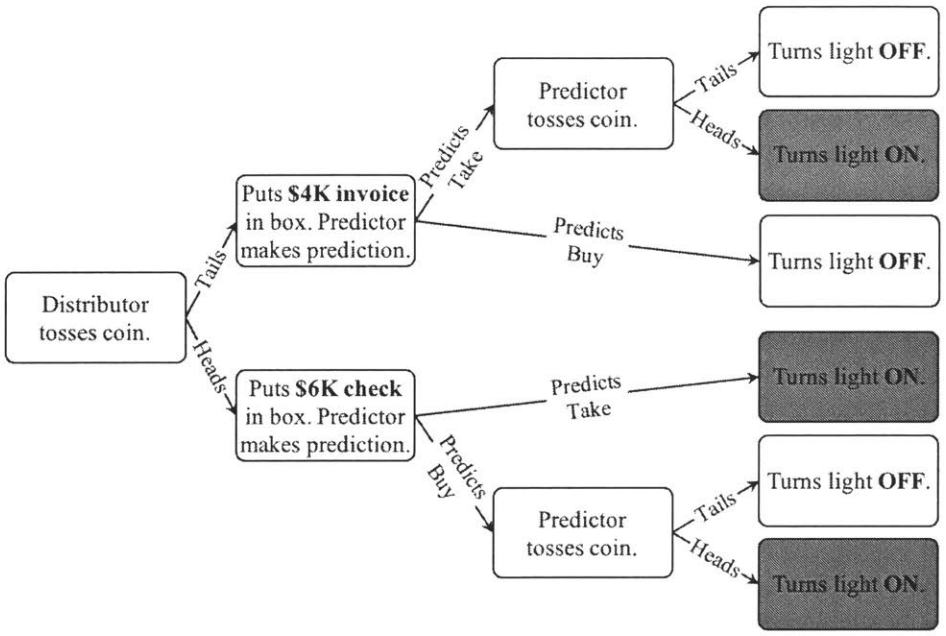

Figure 1-3: The agent's perspective on the *Newcomb Coin Toss* protocol after learning that the light is off.

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 1-2, 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 1-3, 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.

1.6 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.

1.7 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.

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.[13] Hence, in the non-binding version, *do not pay the fee and then take the box for free*

[13]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.

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.**[14] 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.

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?

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

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 motivating the Binding Principle, motivates only a weaker principle:

> **The Weak Binding Principle.**[15] 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 consequence of evidential decision theory in *Newcomb Coin Toss* without invoking binding: namely, causal decision theory.[16]

1.8 '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?'[17] Lewis, a staunch two-boxer, searched for a

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

[16]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.

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

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.[18]

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?'

1.9 Appendix A: Calculations

This appendix confirms the claims of §1.5. 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_jK_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:

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

Evidentialists do not define V in terms of circumstances or dependency hypotheses, since those concepts are explicitly causal. But, as Lewis (1981) 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:

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

[18]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.

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:

$$\begin{aligned} V_L(B) &= (.99)(3000) - (.01)(7000) = 2900. \\ V_L(\overline{B}) &= (.67)(6000) - (.33)(4000) = 2700. \end{aligned}$$

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:

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

The proof proceeds as before.

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

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

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

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,

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

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

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

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

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

By CP and algebra,

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

By TCP and the description of the problem,

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

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

Plugging in the probabilities:

$$V_{\overline{L}}(B) = (.33)(3000) - (.67)(7000) = -3700.$$
$$V_{\overline{L}}(\overline{B}) = (.01)(6000) - (.99)(4000) = -3900.$$

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

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:

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

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, $U(\overline{F}) > U(F)$. ■

1.10 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 (2014a) 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.

Chapter 2

Why Take Both Boxes?

2.1 Introduction

In the classic Newcomb problem, there is an agent, a transparent box, an opaque box, and a predictor, known by the agent to be uncannily good:[1]

> *Classic Newcomb.* The agent has two options: she can take either only the opaque box or 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 the predictor. If the predictor predicted that the agent would take both boxes, the opaque box contains nothing. If the predictor predicted that the agent would take only the opaque box, the opaque box contains \$1,000,000. The agent knows all of this.

One-boxing is the claim that the agent facing *Classic Newcomb* is rationally required to take only the opaque box. *Two-boxing* is the claim that the agent is rationally required to take both boxes. In this paper, we fortify the case for two-boxing.

Fortification is needed because the standard argument for two-boxing—a causal dominance argument—fails. The crucial premise of the standard argument is a causal dominance principle that prohibits choosing causally dominated options. The argument fails because the principle is false. As the examples that we present below establish, it is sometimes rationally permissible to choose a causally dominated option.

After presenting counterexamples to the causal dominance principle, we offer a metaethical explanation for why the counterexamples arise. The explanation reveals a new and superior argument for two-boxing, one that eschews

This chapter is co-authored with Jack Spencer. Section titles are my own.

[1] First discussed by Nozick (1969).

the causal dominance principle in favor of a principle linking rational choice to guidance and actual value maximization.

2.2 Actual Value

The actual value of an option (sometimes called the value, utility, or actual utility of the option) is the value of the outcome that would result if the agent were to choose the option. For example, imagine that there are several boxes, each containing a sum of money. The agent is to choose one of the boxes. The outcome that would result if the agent were to choose a particular box is the agent receiving the sum of money contained therein. If money is all that matters, and more money is linearly better, then the actual value of choosing the box can be identified with the number of dollars contained therein.

The main task of decision theory is to identify the options, among those available to the agent, that the agent is rationally permitted to choose. The task is easy when the agent knows the actual values of her options, for then an option is rationally permissible to choose if and only if the option maximizes actual value.[2] The task is more interesting and more difficult when the agent does not know the actual values of her options.

2.3 Two Rules of Rational Choice

Following Savage and Jeffrey,[3] many decision theorists believe that an option is rationally permissible for an agent to choose if and only if the option maximizes *expected* value, where the expected value of an option is the agent's expectation of the actual value of the option.[4] There are many well-defined expected value quantities and there is considerable disagreement about which of them, if any, is tied to rational choice. We will focus on two: causal expected value (hereafter *c-expected value*) and evidential expected value (hereafter *e-expected value*). Both can be defined in a common conceptual framework, which centers on the concept of a decision problem.

A decision problem is characterized by a set of options, a set of possible outcomes, and a decision-making agent. The options $\mathcal{A} = \{A_1, A_2, \ldots, A_n\}$ are the things the agent is choosing between. We take options to be propositions

[2]Cf. Ramsey (1990 [1926], p. 70): "Let us begin by supposing that our subject has no doubts about anything, but certain opinions about all propositions. Then we can say that he will always choose the course of action which will lead in his opinion to the greatest sum of good."

[3]Savage (1954), Jeffrey (1965).

[4]See note 9.

that the agent can make true by choosing.[5] We assume that options are finite in number, mutually exclusive, and jointly exhaustive. The possible outcomes $\mathcal{O} = \{O_1, O_2, \ldots, O_m\}$ are the objects of non-instrumental desire. We take outcomes to be propositions that fully specify the desirable and undesirable consequences that might result from the choice. Like options, outcomes are assumed to be finite in number, mutually exclusive, and jointly exhaustive. We associate the agent both with a credence function C and with a valuation function V. The credence function, a probabilistically coherent function that maps propositions to the unit interval, represents the agent's beliefs. The agent's credence in P, $C(P)$, is the degree to which the agent believes that P. The valuation function, which maps outcomes to real numbers, represents the agent's desires.[6] The value of outcome O, $V(O)$, is the degree to which the agent finds O non-instrumentally desirable.

Given this conception of a decision problem, the e-expected value of an option $A \in \mathcal{A}$ can be written as a credence-weighted sum, wherein the relevant credences are conditional on the option in question:

$$eev(A) = \sum_O C(O \mid A)V(O).$$

The *rule of e-expected value* states that agents are always rationally required to choose so as to maximize e-expected value.

Let '$\square\!\!\rightarrow$' be the non-backtracking counterfactual conditional. If we assume that, for each option, there is a fact of the matter about which outcome would result if the agent were to choose the option,[7] then the c-expected value of an option can be written as a credence-weighted sum, wherein the relevant

[5]We follow Jeffrey (1965) and take options to be among the propositions to which the agent assigns credences. Some philosophers are skeptical of assigning credences to options, since they think that deliberation crowds out prediction. See, among others, Spohn (1977) and Levi (1997). We are convinced by the arguments in Joyce (2002), Rabinowicz (2002) and Hájek (2016) that deliberation does not crowd out prediction.

For reasons discussed in, among other places, Hedden (2015a) and Pollock (2002), an agent must be certain that she will choose an option if she decides to do so. We therefore identify options and decisions. Each option, besides the null decision, is a proposition of the form: S decides to ϕ.

[6]The valuation function is unique up to positive affine transformation.

[7]This assumption is tantamount to counterfactual excluded middle. For a discussion of causal decision theory without counterfactual excluded middle, see, for example, Lewis (1981), Sobel (1994, p. 141-73) and Joyce (1999).

credences are unconditional credences in counterfactual conditionals:[8]

$$cev(A) = \sum_O C(A \mathrel{\square\!\!\rightarrow} O)V(O).$$

The *rule of c-expected value* states that agents are always rationally required to choose so as to maximize c-expected value.[9]

It is generally agreed that, in *Classic Newcomb*, the rule of e-expected value entails one-boxing,[10] and the rule of c-expected value entails two-boxing.[11] But appealing to the rule of e-expected value or the rule of c-expected value cannot settle the debate between one-boxers and two-boxers, for, as you might

[8]See, for example, Gibbard and Harper (1978) and Stalnaker (1981). Some philosophers are suspicious of assigning probabilities to counterfactuals. (Thanks to an anonymous referee here.) For the purposes of this paper, alternative characterizations of c-expected work equally. We could define c-expected value using expected chance, as Skyrms (1984) does, or using dependency hypotheses, as Lewis (1981) does, or using the epistemic (as opposed to stochastic) probabilities of Kyburg (1980).

[9]So long as every option has an actual value (see note 15), both e-expected value and c-expected value can be defined as expectations of actual value. An $av(A)$-level proposition has the form $[av(A) = v]$, and is true just if v is the actual value of A. The e-expected value of an option is the agent's conditional expectation of the actual value of the option and can be written $\sum_v vC([av(A) = v] \mid A)$. The c-expected value of an option is the agent's unconditional expectation of the actual value of the option and can be written $\sum_v vC([av(A) = v])$.

[10]Let A_{1B} be the option of taking only the opaque box. Let A_{2B} be the option of taking both boxes. Conditional on A_{1B}, the agent is highly confident that the opaque box contains \$1,000,000, so, equating dollars and units of value, $eev(A_{1B}) \approx 1,000,000$. Conditional on A_{2B}, the agent is highly confident that the opaque box contains \$0, so $eev(A_{2B}) \approx 1,000$. Since $eev(A_{1B}) > eev(A_{2B})$, the rule of e-expected value entails one-boxing.

[11]Let O_0, O_T, O_M, and O_{M+T} be the outcomes of receiving \$0, \$1,000, \$1,000,000, and \$1,001,000, respectively. The agent knows that either O_0 or O_M will result if she takes only the opaque box and that either O_T or O_{M+T} will result if she takes both boxes. Moreover, she knows that her choice has no causal bearing on what sum of money is contained in the opaque box, so $C([A_{1B} \mathrel{\square\!\!\rightarrow} O_0]) = C([A_{2B} \mathrel{\square\!\!\rightarrow} O_T])$ and $C([A_{1B} \mathrel{\square\!\!\rightarrow} O_M]) = C([A_{2B} \mathrel{\square\!\!\rightarrow} O_{M+T}])$. Hence, no matter what credence function she has, the c-expected value of taking both boxes is exactly 1,000 greater than the c-expected value of taking only the opaque box:

$$\begin{aligned}
cev(A_{2B}) &= \sum_O C([A_{2B} \mathrel{\square\!\!\rightarrow} O])V(O) \\
&= C([A_{2B} \mathrel{\square\!\!\rightarrow} O_T])V(O_T) + C([A_{2B} \mathrel{\square\!\!\rightarrow} O_{M+T}])V(O_{M+T}) \\
&= (1 - C([A_{2B} \mathrel{\square\!\!\rightarrow} O_{M+T}]))(1,000) + C([A_{2B} \mathrel{\square\!\!\rightarrow} O_{M+T}])(1,001,000) \\
&= 1,000 + C([A_{2B} \mathrel{\square\!\!\rightarrow} O_{M+T}])(1,000,000) \\
&= 1,000 + C([A_{1B} \mathrel{\square\!\!\rightarrow} O_0])(0) + C([A_{1B} \mathrel{\square\!\!\rightarrow} O_M])(1,000,000) \\
&= 1,000 + \sum_O C([A_{1B} \mathrel{\square\!\!\rightarrow} O])V(O) \\
&= 1,000 + cev(A_{1B}).
\end{aligned}$$

suspect, one-boxers typically reject the rule of c-expected value, and two-boxers typically reject the rule of e-expected value.[12] To move the debate forward, we need an independent argument, one that nowhere appeals to an expected value quantity.

Many two-boxers believe that they have an independent argument: namely, a causal dominance argument.[13]

2.4 Causal Dominance

A natural way to argue for two-boxing is by disjunctive syllogism. We can imagine running through the argument from the agent's point of view:

> The opaque box contains either \$1,000,000 or nothing. If it contains \$1,000,000, then both boxes together contain \$1,001,000, and hence I would make more money if I took both boxes. If it contains nothing, then both boxes together contain \$1,000, and hence I would make more money if I took both boxes. Either way, I would make more money if I took both boxes. So I should take both boxes.

This argument, although unregimented, seems compelling and nowhere invokes an expected value quantity.

A preliminary attempt to regiment the argument appeals to states and (strict) dominance. A set of propositions $\mathcal{S} = \{S_1, S_2, ..., S_n\}$ is a set of *states* if its members are mutually exclusive and jointly exhaustive, and each $S \in \mathcal{S}$ is compossible with each $A \in \mathcal{A}$. Let AS be the conjunction of option A and state S. If the options and states are sufficiently fine-grained (and let us choose them so that they are), then every AS necessitates a unique outcome. If AS necessitates O, we set $V(AS)$ equal to $V(O)$. Option A_i *dominates* option A_j, then, if and only if there is a set of states $\mathcal{S}$ such that, for every $S \in \mathcal{S}$, $V(A_iS)$ exceeds $V(A_jS)$.

One might allege the following connection between dominance and rational choice:

> **Dominance**: If option A_i dominates option A_j, then it is not rationally permissible for the agent to choose A_j.

[12]Heterodoxically, Eells (1982) argues that the rule of e-expected value entails two-boxing, and Spohn (2012) argues that the rule of c-expected value entails one-boxing.

[13]See, for example, Joyce (1999, p. 152-54), Lewis (1981, p. 309-12), Skyrms (1984, p. 67) and Sobel (1994).

But it is common ground between one-boxers and two-boxers that Dominance is false. It is sometimes rationally permissible to choose dominated options, as cases like the following make vivid:[14]

> *The Extortionist.* A moviegoer parks her car in the lot. An extortionist, who the moviegoer has excellent reason to trust, says to the moviegoer, "If you pay me $10, I'll ensure that your windshield is unbroken when you return. But I'll smash your windshield if you don't pay me."

Let the set of states be $\{S_B, S_{\overline{B}}\}$, where S_B is the proposition that the windshield is broken when the moviegoer returns and $S_{\overline{B}}$ is the proposition that the windshield is not broken when the moviegoer returns. Let A_P be the option of paying the extortionist and let $A_{\overline{P}}$ be the option of not paying. $V(A_{\overline{P}}S_{\overline{B}}) > V(A_PS_{\overline{B}})$, since it would be better by the moviegoer's lights not to pay the extortionist and return to an unbroken windshield than to pay the extortionist and return to an unbroken windshield. $V(A_{\overline{P}}S_B) > V(A_PS_B)$, since it would better by the moviegoer's lights not to pay the extortionist and return to a broken windshield than to pay the extortionist and return to a broken windshield. Dominance therefore entails that the agent is rationally required to not pay the extortionist—which is absurd. The moviegoer is rationally required to pay the extortionist: paying $10 is much better than paying $1,000 for a new windshield.

A bit of reflection reveals why Dominance fails. Reasoning by Dominance is supposed to put the agent in a position to conclude a fact about the ordinal ranking of options vis-à-vis actual value. The disjunctive syllogism mentioned at the outset of this section, for example, is supposed to put the agent in a position to conclude that the actual value of taking both boxes exceeds the actual value of taking only the opaque box. But actual value does not respect dominance: the fact that A_i dominates A_j does not entail that the actual value of A_i exceeds the actual value of A_j. Since we are assuming that, for each option, there is a fact of the matter about which outcome would result if the agent were to choose the option, we can characterize actual value as a sum. Where T is an indicator function that assigns one to truths and zero to falsehoods, the actual value of an option, $av(A)$, can be written:

$$av(A) = \sum\nolimits_O T(A \mathrel{\square\!\!\rightarrow} O)V(O).$$

Given our assumptions, there is exactly one $O \in \mathcal{O}$ for which $[A \mathrel{\square\!\!\rightarrow} O]$ is true. If $[A \mathrel{\square\!\!\rightarrow} O]$ is true, then O is the outcome that would result if the agent

[14] An adaptation of an example from Joyce (1999, p. 114-19). Jeffrey (1965, p. 9-10) uses the example of nuclear disarmament.

were to choose A, and hence $av(A)$ equals $V(O)$.[15] Let $S_@$ be the state that actually obtains. The fact that A_i dominates A_j entails that $V(A_iS_@)$ exceeds $V(A_jS_@)$. It might be tempting to identify the actual values of A_i and A_j with $V(A_iS_@)$ and $V(A_jS_@)$, respectively. But that temptation must be resisted. The actual value of A is equal to $V(AS)$ only if S would have obtained had the agent chosen A. If the agent chooses A_i, then the actual value of A_i is equal to $V(A_iS_@)$. But the actual value of an unchosen option A_j need not be equal to $V(A_jS_@)$.

To illustrate, return to *The Extortionist*, and suppose that the extortionist is trustworthy. The moviegoer irrationally chooses to not pay the extortionist and returns to a broken windshield. S_B is true, and the actual value of not paying is equal to $V(A_{\overline{P}}S_B)$. $V(A_{\overline{P}}S_B)$ exceeds $V(A_PS_B)$, of course, since not paying dominates paying. But the actual value of paying is not $V(A_PS_B)$; rather, the actual value is $V(A_PS_{\overline{B}})$, since the outcome that would result if the agent were to pay the extortionist is that she would have \$10 fewer and an unbroken windshield. Moreover, $V(A_PS_{\overline{B}})$ exceeds $V(A_{\overline{P}}S_B)$.

But, importantly, while actual value does not respect dominance, it does respect causal dominance. A state is *causally act-independent* for an agent if and only if the agent knows that she has no causal influence over whether the state obtains. (More formally, S is causally act-independent for an agent if and only if the agent knows, for each $A \in \mathcal{A}$, $S \leftrightarrow [A \mathrel{\Box\!\!\rightarrow} S]$.) If there is a set of causally act-independent states $\mathcal{S}$ such that, for every $S \in \mathcal{S}$, $V(A_iS)$ exceeds $V(A_jS)$, then option A_i *causally dominates* option A_j.[16] The alleged connection between causal dominance and rational choice is structurally identical to the alleged connection between dominance and rational choice:

> **Causal Dominance**: If option A_i causally dominates option A_j, then it is not rationally permissible for the agent to choose A_j.

But Causal Dominance is more plausible than Dominance. Causal Dominance avoids the absurd recommendation, in *The Extortionist*, that the moviegoer rationally ought to not pay.[17]

---

[15] If counterfactual excluded middle fails, unchosen options might fail to have actual values. When $[A \mathrel{\Box\!\!\rightarrow} O]$ is true, the chance of O conditional on A, i.e. $Ch(O \mid A)$, is one, so we could broaden the notion of actual value by setting $av(A)$ equal to $\sum_O Ch(O \mid A)V(O)$. But it is unclear whether the broadened notion of actual value can do the metaethical work done by the narrower notion.

[16] If A_i causally dominates A_j, then $V(A_iS_@) > V(A_jS_@)$. Since $S_@$ is causally act-independent, $[A_i \mathrel{\Box\!\!\rightarrow} A_iS_@]$ and $[A_j \mathrel{\Box\!\!\rightarrow} A_jS_@]$ both are true, so $av(A_i) = V(A_iS_@)$ and $av(A_j) = V(A_jS_@)$. Hence, $av(A_i) > av(A_j)$.

[17] The moviegoer knows that she exerts causal influence over the future state of her windshield, so neither S_B nor $S_{\overline{B}}$ is causally act-independent.

Causal Dominance is weaker than Dominance but still strong enough to entail two-boxing. Let the set of states be $\{S_{\$0}, S_{\$M}\}$, where $S_{\$0}$ is the proposition that the opaque box contains \$0 and $S_{\$M}$ is the proposition that the opaque box contains \$1,000,000. Since

$$V(A_{2B}S_{\$0}) = 1,000 > 0 = V(A_{1B}S_{\$0}), \text{ and}$$
$$V(A_{2B}S_{\$M}) = 1,001,000 > 1,000,000 = V(A_{1B}S_{\$M}),$$

A_{2B} dominates A_{1B}. Moreover, the agent knows that she has no causal influence over the amount of money in the opaque box, so $S_{\$0}$ and $S_{\$M}$ are causally act-independent states. Hence, taking both boxes causally dominates taking only the opaque box, a fact exploited in the *Causal Dominance Argument* for two-boxing:

(P1) If option A_i causally dominates option A_j, then it is not rationally permissible for the agent to choose A_j.

(P2) In *Classic Newcomb*, taking both boxes causally dominates taking only the opaque box.

(C) Therefore, an agent facing *Classic Newcomb* is rationally required to take both boxes.

The Causal Dominance Argument is the aforementioned standard argument for two-boxing.[18]

Note the intimate relation between Causal Dominance and the rule of c-expected value. Given a set of causally act-independent states $\mathcal{S}$, the c-expected value of an option can be characterized as a function of the agent's unconditional credences in the members of $\mathcal{S}$:

$$cev(A) = \sum_O C(A \mathrel{\square\!\!\rightarrow} O)V(O) = \sum_S C(S)V(AS).$$

As the last sum in the equation makes clear, a causally dominated option cannot maximize c-expected value: the rule of c-expected value entails Causal Dominance.

[18]Some prefer an informational variant; see, for example, Pollock (2010, p. 57-82). Not every two-boxer relies on dominance reasoning. See, for example, Levi (1975).

2.5 Two Counterexamples to Causal Dominance

We believe that Causal Dominance is false, and hence that the Causal Dominance Argument is unsound. We will offer two counterexamples to the rule of c-expected value, and then transform them into counterexamples to Causal Dominance.

The first counterexample is non-ideal. An ideal agent is both introspective—she knows all of the facts about her own beliefs and desires—and logically omniscient. A non-ideal agent is introspective but not logically omniscient. An ideal counterexample features an ideal agent, and a non-ideal counterexample, like the following, features a non-ideal agent:

> *The Fire.* The fire alarm rings and the agent, a firefighter, hurries onto the truck. On the ride over she deliberates. She has three options: she can enter the building through the left door, the middle door, or the right door. Since she does not know the exact distribution of residents in the building, she does not know which option will result in the most rescues. Based on her credences about the distribution of residents, she calculates the c-expected value of each option and writes the value on a notecard. After exiting the truck and attaching the water hose, she races toward the building. She reaches into her pocket, but the notecard is gone! Time is of the essence. She knows that all of the residents will die in the time it would take her to recalculate the c-expected values. Her credences about the distribution of residents are unchanged, so she knows that her current c-expected values are what they were when she calculated them. But she cannot fully remember the results of her calculations. She remembers that the c-expected value of entering through the middle door is 9. Of the other two options, she remembers that one has a c-expected value of 0 and that the other has a c-expected value of 10, but she cannot remember which c-expected value goes with which option. (In fact, entering through the right door has a c-expected value of 10, as the lost notecard attests.)[19]

[19] *The Fire* is an elaboration of a case discussed by Kagan (MS). The fact that non-ideal agents are not always able to access expected value is also discussed in, among other places, Feldman (2006) and Weirich (2004, ch. 5). Some philosophers believe that decision theory applies only to ideal agents, and hence that examples like *The Fire* cannot be relevant to decision theory. We think that decision theory should extend to non-ideal agents. But, even if decision theory applies only to ideal agents, we think that non-ideal cases, such as *The Fire*, help shed light on how an agent must be related epistemically to a value quantity in

We say that the agent facing *The Fire* is rationally required to enter through the middle door, even though it is true, by hypothesis, that the option that uniquely maximizes c-expected value is entering through the right door.[20]

The second counterexample is ideal.

> *The Frustrater.* There is an envelope and two opaque boxes, A and B. The agent has three options: she can take box A, box B, or the envelope. (The three options may be labeled A_A, A_B, and A_E, respectively.) The envelope contains \$40. The two boxes together contain \$100. How the money is distributed between the boxes depends on a prediction made yesterday by the Frustrater, a reliable predictor who seeks to frustrate. If the Frustrater predicted that the agent would take box A, box B contains \$100. If the Frustrater predicted that the agent would take box B, box A contains \$100. If the Frustrater predicted that the agent would take the envelope, each box contains \$50. The agent knows all of this.[21]

We say that an ideal agent facing *The Frustrater* is rationally required to choose the envelope. But the options that maximize c-expected value are A_A and/or A_B, depending on the agent's credences.[22] (*Proof*: No matter what credence function the agent has, $cev(A_E) = 40$ and $cev(A_A) + cev(A_B) = 100$. Two numbers smaller than 40 cannot sum to 100.)[23]

---

order to be rationally required to maximize that value quantity, an issue that we discuss in more detail in §2.5, §2.6, and §2.7.

[20]We think that the intuition in this case speaks for itself. But we offer a theoretical account of why the agent facing *The Fire* is rationally required to enter through the middle door in note 31.

[21]This example is inspired by other purported counterexamples to the rule of c-expected value: Bostrom (2001), Egan (2007) and especially Ahmed (2014b). Lewis (1981) shows that there are realistic cases that have the same structure as *Classic Newcomb*. There are also realistic cases that have the same structure as *The Frustrater*. A commuter wants to get home as soon as possible. There are three routes home: highway A, highway B, and the ferry. The commuter knows that the ferry is second-fastest. The commuter does not know which highway is faster—that varies depending on the day—but the commuter knows that one of the highways is slightly faster than the ferry and that the other is much, much slower. Moreover, the commuter reasonably believes that commuters are like-minded. Conditional on taking highway A/B, the commuter is highly confident that highway B/A is the fastest route.

[22]We assume that the agent facing *The Frustrater* cannot play a mixed strategy. Perhaps the agent is unable to randomize her choice, or perhaps it is simply unwise to play a mixed strategy, since the Frustrater is very good at detecting whether an agent is playing a mixed strategy and punishes the agent severely for doing so.

[23]If we transform *The Frustrater* into a sequence of choices—first a choice between A_E and eliminating A_E, and then a choice, if A_E is eliminated, between A_A and A_B—the rule of c-expected value as applied to the sequence recommends A_E. We note three things. First,

With a few alterations, both *The Fire* and *The Frustrater* can be transformed into counterexamples to Causal Dominance. Start with a variation on *The Fire*:

> *The Dominating Fire.* Everything is the same as in *The Fire*, except that, unbeknownst to the agent, the option of entering through the right door causally dominates the other two options.

From the standpoint of rationality, *The Dominating Fire* is no different than *The Fire*. A non-ideal agent might not be in a position to know which options causally dominate which others. (We can imagine that the $V(AS)$'s are stored in the agent's brain, in the form of a payoff matrix, and that it takes the agent a non-trivial amount of time to survey the matrix.) If an agent is not in a position to know that an option is causally dominated, then the fact that the option is causally dominated is not relevant to what the agent rationally ought to choose. Therefore, as in *The Fire*, an agent facing *The Dominating Fire* is rationally required to enter through the middle door, even though entering through the middle door is causally dominated by entering through the right door.

Causal Dominance is an elimination principle, which marks options as rationally impermissible to choose. But it entails the following selection principle:

> **Causal Dominance Selection**: If option A_i causally dominates all other options, then the agent is rationally required to choose A_i.

The Dominating Fire is a counterexample not just to Causal Dominance, but also to Causal Dominance Selection.

There are no ideal counterexamples to Causal Dominance Selection, a fact that we will return to, and explain, later. But there are ideal counterexamples to Causal Dominance:

> *The Semi-Frustrater.* There are two buttons, a white button and a black button. The agent has four options: she can press either button with either hand. (The four options may be labeled $A_{RH:W}$,

this is a different decision problem. *The Frustrater* remains a counterexample to the rule of c-expected value. Second, it may not be rationally permissible for the agent to choose between A_E and eliminating A_E—perhaps because the Frustrater punishes agents who do so. Third, not all of the counterexamples to the rule of c-expected value can be transformed into a sequence of choices, cf. Egan (2007). Thanks to Bernhard Salow and Caspar Hare for discussion on this point.

> $A_{LH:W}$, $A_{RH:B}$, and $A_{LH:B}$.) The white button connects to the white box, the black button connects to the black box, and the agent will receive the contents of whatever box is connected to the button she presses. One of the boxes contains \$0 and the other contains \$100. Which box contains which sum depends on a prediction made yesterday by the Semi-Frustrater. The Semi-Frustrater seeks to frustrate. If the Semi-Frustrater predicted that the agent would press the black button, the white box contains \$100. If the Semi-Frustrater predicted that the agent would press the white button, the black box contains \$100. There are two left-right asymmetries. First, the agent will receive an extra \$5 if she presses a button right-handedly. Second, because the Semi-Frustrater bases her prediction on a scan of merely half of the agent's brain, the Semi-Frustrater is a 90% reliable predictor of right-handed button pressings but only a 50% reliable predictor of left-handed button pressings. The agent knows all of this.

We say that *The Semi-Frustrater*, like *The Frustrater*, is an ideal counterexample to the rule of c-expected value. In our view, an ideal agent facing *The Semi-Frustrater* is rationally required to choose $A_{LH:W}$ or $A_{LH:B}$, and rationally permitted to choose either, even though the options that maximize c-expected value are, depending on the agent's credences, $A_{RH:W}$ and/or $A_{RH:B}$.[24] What is more surprising is that we have an ideal counterexample to Causal Dominance. The claim that an (ideal) agent is never rationally permitted to choose a (strictly) causally dominated option is a staple of game theory, where it appears in textbooks as the injunction against playing strategies that can be iteratively eliminated by (strict) causal domination,[25] and is regarded as sacrosanct by many expert decision theorists.[26] But $A_{RH:W}$ causally dominates $A_{LH:W}$, and $A_{RH:B}$ causally dominates $A_{LH:B}$, so an ideal

[24]Either S_W, the white box contains \$100, or S_B, the black box contains \$100. Since the agent knows that her choice has no causal influence over the contents of the boxes, $\{S_W, S_B\}$ is a set of causally act-independent states. Equating dollars and units of value, $V(S_W A_{RH:W}) = 105 = 5 + V(S_W A_{LH:W})$; $V(S_B A_{RH:W}) = 5 = 5 + V(S_B A_{LH:W})$; $V(S_W A_{RH:B}) = 5 = 5 + V(S_W A_{LH:B})$; and $V(S_B A_{RH:B}) = 105 = 5 + V(S_B A_{LH:B})$. So $cev(A_{RH:W})$ maximizes if $C(S_W) \geq 0.5$, and $cev(A_{RH:B})$ maximizes if $C(S_B) \geq 0.5$.

[25]See, for example, Fudenberg and Tirole (1991, ch. 2) and Myerson (1991, s. 3.1).

[26]Briggs (2015, p. 836): "The following is an independently compelling claim about rationality: if it is knowable a priori that strategy a yields a better result than strategy b, then it is pragmatically irrational to choose strategy b when strategy a is available." Pettigrew (2015, p. 806): "the so-called Dominance Principle, which says that an option is irrational if there is an alternative that is guaranteed to be better than it, and if there is nothing that is guaranteed to be better than that alternative [...] is an uncontroversial principle of decision theory." Also see, for example, Buchak (2015), Briggs (2010), Gibbard and Harper (1978), Lewis (1981), Joyce (1999), Nozick (1969), Sobel (1994) and Skyrms

agent facing *The Semi-Frustrater* is rationally required to choose a causally dominated option.

One might object that *The Semi-Frustrater* is really no different than *Classic Newcomb*. In both examples there is some intuitive pull toward choosing a causally dominated option, since in both examples consistently choosing a causally dominated option results in greater long run wealth. Consistent one-boxers end up wealthier than do consistent two-boxers. Consistent left-handers end up wealthier than do consistent right-handers. Two-boxers resist the one-boxing intuition, so, if the intuition that an agent rationally ought to press a button left-handedly in *The Semi-Frustrater* is really no different than the intuition that an agent rationally ought to take only the opaque box in *Classic Newcomb*, two-boxers should also resist the left-handed intuition.

But, in at least two respects, *The Semi-Frustrater* and *Classic Newcomb* are importantly different. The first and most important difference is metaethical—that is, it concerns how the agent is epistemically related to the relevant value quantities. For an agent facing *Classic Newcomb*, maximizing c-expected value is non-accidentally doable. If the agent seeks to maximize c-expected value, she can be confident both about which option she will choose and that she will choose an option that maximizes c-expected value. She will be confident that she will take both boxes and confident that by doing so she will choose an option that maximizes c-expected value. By contrast, for an agent facing *The Semi-Frustrater*, maximizing c-expected value is doable only accidentally. An agent facing *The Semi-Frustrater*, who seeks to maximize c-expected value, cannot be confident both about which option she will choose and that she will choose an option that maximizes c-expected value. If she is confident about which option she will choose, then she is confident that by choosing that option she will *fail* to maximize c-expected value. As we say in more detail below, in our view, the fact that the c-expected value of an option exceeds the c-expected value of another option is relevant to what an agent rationally ought to choose only if the agent is appropriately related to c-expected value maximization. An agent facing *Classic Newcomb* is appropriately related to c-expected value maximization, but an agent facing *The Semi-Frustrater* is not. Ultimately it is this epistemological and metaethical difference that marks the crucial divide between *The Semi-Frustrater* and *Classic Newcomb*.

But even before we get into metaethics, there is a simple descriptive difference between *The Semi-Frustrater* and *Classic Newcomb*. In *Classic Newcomb*, there is unequal environmental fortune. Imagine that the choices are made in a room containing only the agent and the boxes. Consistent one-boxers almost always make their choices in lucrative rooms: they almost always choose

(1984). In epistemic decision theory, too, the claim that (strictly) dominated options are *ipso facto* irrational is relied upon heavily. See, for example, Joyce (1998).

between two options, each worth $1,000,000 or more, in a room that contains more than $1,000,000. Consistent two-boxers almost always make their choices in impoverished rooms: they almost always choose between two options, each worth no more than $1,000, in a room that contains $1,000. The argument for one-boxing, based on the claim that consistent one-boxers are wealthier than consistent two-boxers, is undermined by the unequal environmental fortune. What explains why consistent one-boxers are wealthier than consistent two-boxers is that one-boxers make their choices in lucrative rooms, not that one-boxers choose more wisely.[27] Consistently choosing unwisely in lucrative rooms leads to greater long run wealth than does consistently choosing wisely in impoverished rooms. Notice, in *The Semi-Frustrater*, however, that there is no difference in environmental fortune. Like consistent left-handers, consistent right-handers always make their choices in rooms that contain exactly $105. What explains why consistent left-handers are wealthier than consistent right-handers is a difference of rationality, not a difference of environmental fortune. Consistent right-handers end up poorer than do consistent left-handers because they choose irrationally.

2.6 Guidance

Although the Causal Dominance Argument is unsound, a successful, independent argument for two-boxing is in the nearby vicinity. The successful argument relies on a metaethical principle connecting guidance and actual value maximization to rational choice.

There are two 'ought's of decision-making, an objective 'ought' and a rational 'ought'. Decision theory, being consequentialist in nature, takes both to be reducible to value quantity maximization.

The objective 'ought' reduces to actual value maximization: agents are always objectively required to choose so as to maximize actual value.

The objective 'ought' is not our main concern. Our main concern is the rational 'ought', which can, and often does, come apart from the objective 'ought'. For example:

> *Boxes like Miners.* There are three opaque boxes: the left box, the middle box, and the right box. The agent must choose exactly one box. The agent knows that the middle box contains $9. Of the other two boxes, the agent knows that one contains $0 and that the other contains $10, but does not know which box contains which sum. (In fact, the right box contains $10.)

[27] For more on this point, see Wells (Forthcoming).

An agent facing *Boxes like Miners* is, though objectively required to choose the right box, rationally required to choose the middle box.

At the metaethical level, the most important difference between the objective 'ought' and the rational 'ought' is a difference of guidance. The objective 'ought' is not always capable of guiding the agent's choice. Actual value is the value quantity the maximization of which makes options objectively permissible for the agent to choose, but agents are not always capable of being guided by actual value. A necessary condition on being capable of being guided by actual value is being in a position to know of some option that it maximizes actual value, and agents often are in no such position. An agent facing *Boxes like Miners*, for example, is in no such position.

The rational 'ought', by contrast, is always capable of guiding the agent's choice. An agent is always capable of being guided by the value quantity the maximization of which makes options rationally permissible for the agent to choose.

It is here that we break with the metaethical orthodoxy. The orthodoxy has it that a value quantity is choice-guiding if and only if the facts about which options maximize the value quantity supervene on the facts about the agent's beliefs and desires.[28] Actual value fails this supervenience test. The actual value of an option is a function of the truth-values of certain counterfactual claims, and such truth-values float free of the agent's psychology. By contrast, e-expected value and c-expected value pass the supervenience test. Both are functions of the agent's beliefs and desires.

In our view, the orthodoxy is mistaken twice over. First, it is a mistake to try to divide value quantities into those that are, and those that are not, choice-guiding. Whether a value quantity is capable of guiding an agent's choice is settled, in our view, occasion by occasion, not once and for all. Second, it is a mistake to identify choice-guidance with supervenience on the agent's beliefs and desires. On some occasions an agent is capable of being guided by a value quantity that does not supervene on her beliefs and desires, and on some occasions an agent is incapable of being guided by a value quantity that supervenes on her beliefs and desires.

We claim that a value quantity is capable of guiding an agent's choice on an occasion if and only if the agent has stable access to the value quantity on that occasion. Stable access is defined in terms of being in a position to know. An agent is *in a position to know* a proposition if and only if there is no obstacle blocking her from knowing the proposition.[29] An agent has *access* to a value quantity Q if and only if there is an option $A \in \mathcal{A}$ such that the

[28]Or, more generally, supervene on the agent's internal mental states. See, for example, Conee and Feldman (2004).

[29]Cf. Williamson (2000, p. 95).

agent is in a position to know of A that it maximizes Q. An agent has *stable access* to Q if and only if there is an option $A \in \mathcal{A}$ such that (i) the agent is in a position to know of A that it maximizes Q, and (ii) conditional on A, the agent still is in a position to know of A that it maximizes Q.[30] If an agent has stable access to Q, then the agent is *stably* in a position to know of some option A that it maximizes Q.

The stability is crucial. An agent who chooses option A is guided by Q only if she can know both that she will choose A and that A is Q-maximizing. It is for this reason that access alone is not sufficient for guidance. An agent who has access but lacks stable access to Q cannot know both which option she will choose and that the option she will choose is Q-maximizing. Such an agent either is surprised by which option she chooses, in which case her choice is not guided at all, or she anticipates choosing an option that is not Q-maximizing, in which case her choice is not guided by Q. Stability plugs this gap. An agent who has stable access to Q is capable of being guided by Q because she can know both that she will choose A and that A is Q-maximizing.[31]

2.7 Ratificationism

If an agent is incapable of being guided by a value quantity, then the maximization of that value quantity is not what makes options rationally permissible for the agent to choose. In our view, this is the fact that explains why the rule of c-expected value admits of counterexamples. Agents are not always capable of being guided by c-expected value—that is, agents do not always have stable access to c-expected value. An agent facing *The Fire* or *The Dominating Fire* does not have access, let alone stable access, to c-expected value, since the external time constraints, together with the agent's limited powers of deduction, form an obstacle blocking her from knowing that entering through the right door maximizes c-expected value.[32] An agent facing *The Frustrater* or

[30] By "conditional on A," we have the following in mind. Take the agent's credence function and conditionalize it on A. Then ask whether the agent still is in a position to know that P, relative to her updated credence function. If she is, then she is stably in a position to know that P. If not, not.

[31] *Question:* Does choice-guidance supervene on the agent's mental states? *Answer:* If being in a position to know is a mental state, then yes. Otherwise, no.

[32] *Question*: What value quantity is an agent facing *The Fire* rationally required to maximize? *Answer*: A value quantity that stands to c-expected value as c-expected value stands to actual value; we might call it c-expected$_2$ value. A $cev(A)$-level proposition is of the form $[cev(A) = v]$. Just as the c-expected value of an option is a credence-weighted average of the agent's hypotheses about the actual value of the option (see note 9), the c-expected$_2$ value of an option is a credence-weighted average of the agent's hypotheses about the c-expected value of the option: $cev_2(A) = \sum_v vC([cev(A) = v])$. There is also a value quantity that

The Semi-Frustrater has access but lacks stable access to c-expected value, since there is no option available in either decision problem that maximizes c-expected value conditional on itself.[33] We claim that the rule of c-expected value admits of counterexamples only when agents lack stable access to c-expected value. In other words, we accept the *guiding rule of c-expected value*: that agents who have stable access to c-expected value are rationally required to choose so as to maximize c-expected value.

In spirit, the guiding rule of c-expected value is similar to the ratificationisms defended by Harper, Jeffrey, Sobel, and others.[34] But at the level of detail, the views differ in important ways, and we think that the guiding rule of c-expected value is an improvement upon the more familiar ratificationisms.

The key notion for any ratificationism is that of an option being ratifiable. An option A is *ratifiable* if and only if, conditional on A, A maximizes c-expected value, and *nonratifiable*, otherwise. There are two sorts of ratificationisms. According to *principled* ratificationism, it is never rationally permissible to choose nonratifiable options. Principled ratificationism is implausible: *The Frustrater* and *The Semi-Frustrater* are counterexamples. According to *lexical* ratificationism, the more plausible version of the view, options are lexically ordered by ratifiability: ratifiable options are infinitely more choiceworthy than are nonratifiable options; hence it is rationally permissible to choose nonratifiable options only if none of the available options are ratifiable. Lexical ratificationists can disagree with one another about how to choose among options at the same lexical order. Some lexical ratificationists use e-expected value to choose among options at the same lexical order.[35] They claim that an agent is rationally required to choose a ratifiable option the e-expected value of which is not exceeded by that of any other ratifiable option, unless there are no ratifiable options, in which case the agent is rationally required to

---

stands to c-expected$_2$ value as c-expected value stands to actual value; we might call it c-expected$_3$ value. In general, for any $n > 1$,

$$cev_n(A) = \sum_v vC([cev_{n-1}(A) = v]).$$

[33] *Question*: What value quantity is an agent facing *The Frustrater* rationally required to maximize? *Answer*: The most causally fine-grained expected value quantity to which the agent has stable access. See Spencer and Wells (MS), in which we develop a theory of rational choice in the face of decision instability. Much of decision theory rests on the assumption that there is a single value quantity that any agent facing any decision problem is rationally required to maximize. We reject this claim. We explore the prospects for a unified decision theory that rejects this assumption.

[34] For discussion of ratificationism, see, among others, Eells (1982), Egan (2007), Gustafsson (2011), Hare and Hedden (2016), Harper (1986), Jeffrey (1983), Joyce (2007), Rabinowicz (1988), Sobel (1994), Skyrms (1984), and Weirich (1988, 2004).

[35] Cf. Jeffrey (1983).

choose a nonratifiable option that maximizes e-expected value. Other lexical ratificationists use c-expected value to choose among options at the same lexical order.[36] Note that both forms of lexical ratificationism entail two-boxing, since, in *Classic Newcomb*, taking both boxes is the only ratifiable option.

The fatal flaw in lexical ratificationism is the lexical ordering of options. By treating ratifiable options as infinitely more choiceworthy than nonratifiable options, lexical ratificationists effectively claim that ratifiability is an infinite value. But ratifiability is not a value, let alone an infinite one. An option is not made more choiceworthy by being ratifiable. The counterexamples to lexical ratificationism, like the following, due to Skyrms (1984), exploit precisely this flaw:

> *Three Shells.* The agent has three options: there are three shells, shell J, shell K, and shell L, and the agent can choose any one of them. How much money is contained in each shell depends on a prediction made yesterday by a reliable predictor. If the predictor predicted that the agent would choose shell J, then shell J contains \$1, shell K contains \$0, and shell L contains \$0. If the predictor predicted that the agent would choose shell K, then shell J contains \$0, shell K contains \$9, and shell L contains \$10. If the predictor predicted that the agent would choose shell L, then shell J contains \$0, shell K contains \$10, and shell L contains \$9.

Choosing shell J is the only ratifiable option. Hence, according to any lexical ratificationism, an agent facing *Three Shells* is rationally required to choose shell J, no matter what credence function the agent has. Lexical ratificationism is thereby refuted. An agent facing *Three Shells* is rationally required to choose shell J only if, at the time of decision, she is highly confident that she will. If she is highly confident that she will choose shell J, then she is highly confident that shell J contains \$1 and that the other two shells contain nothing. But if the agent is not highly confident that she will choose shell J, then it is not even rationally permissible for her to choose shell J. In the extreme case, in which the agent is highly confident that she will not choose shell J, the agent regards shell J as the worst of her three options by far. The claim, made by lexical ratificationists, that the agent is nevertheless rationally required to choose shell J is clearly false.

The guiding rule of c-expected value avoids the counterexamples to lexical ratificationism by rectifying the fatal flaw. The guiding rule of c-expected value is not built on the notion of ratifiability. It is built on the notion of having stable access to a value quantity—specifically, having stable access to

[36]See Egan (2007) for discussion.

c-expected value. Whereas ratifiability operates at the ethical level, affecting the choiceworthiness of options, stable access operates at the *metaethical* level, affecting the relevance of value quantities. Stable access is a necessary condition for rational relevance: facts about which options maximize a given value quantity can be relevant to what an agent rationally ought to choose only if the agent has stable access to that value quantity.

The guiding rule of c-expected value thus can handle cases like *Three Shells* correctly. If an agent facing *Three Shells* is highly confident that she will choose shell J, then she has stable access to c-expected value, and the guiding rule of c-expected value correctly entails that she is rationally required to choose shell J. If the agent is not highly confident that she will choose shell J, then she lacks stable access to c-expected value, and the guiding rule of c-expected value is silent. When an agent, ideal or nonideal, lacks stable access to c-expected value, the facts about the c-expected values of options are not relevant to what the agent rationally ought to choose.

Since we accept the guiding rule of c-expected value, we think that there is a sound argument from c-expected value to two-boxing. A competent agent facing *Classic Newcomb* has stable access to c-expected value because (i) she is in a position to know that A_{2B} (uniquely) maximizes c-expected value, and (ii) conditional on A_{2B}, she still is in a position to know that A_{2B} (uniquely) maximizes c-expected value. Hence, by the guiding rule of c-expected value, she is rationally required to take both boxes.

But, as noted above, arguing from c-expected value to two-boxing fails to move the debate forward. What we need is an independent argument for two-boxing.

2.8 The Objective Argument

We think that the best independent argument for two-boxing goes through the guiding rule of actual value.

According to the *rule of actual value*, agents are always rationally required to choose so as to maximize actual value. Everyone rejects the rule of actual value, and for good reason. Counterexamples abound. Rational permission and actual value maximization often come apart. But the rule of c-expected value also fails: rational permission and c-expected value maximization also come apart. An agent is rationally required to choose so as to maximize c-expected value only when she has stable access to c-expected value. We think that the same holds for actual value. We accept the *guiding rule of actual value*: that agents who have stable access to actual value are rationally required to choose so as to maximize actual value.

The guiding rule of actual value entails the uncontroversial claim that rational permission and actual value maximization can come apart when agents lack access to actual value. In *Boxes like Miners*, for example, the agent is rationally required to choose the middle box, even though choosing the right box uniquely maximizes actual value.

The guiding rule of actual value also entails that rational permission and actual value maximization can come apart when an agent has access but lacks stable access to actual value. Not much attention has been paid to the question of whether rational permission and actual value maximization can come apart in such cases, in part because it requires some fancy footwork to devise an example. Here is one:

> *Unstable Boxes like Miners.* There are four boxes, the outside-left box, the middle-left box, the middle-right box, and the outside-right box. The outside boxes are opaque and the middle boxes are transparent. The middle-left box and the middle-right box each contain \$9. One of the outside boxes contains \$0 and the other contains \$10. Which outside box contains which sum depends on a prediction made yesterday by a reliable predictor. If the predictor predicted that the agent would choose either the middle-left box or the outside-left box, the outside-right box contains \$10. If the predictor predicted that the agent would choose either the middle-right box or the outside-right box, the outside-left box contains \$10. The agent knows all this. The agent also believes that she will choose the middle-left box.

Since the agent believes that she will choose the middle-left box and believes that the predictor is extremely reliable, she believes that the outside-right box contains \$10. Moreover, let us suppose that it is true that the outside-right box contains \$10. Presumably, then, if the predictor is reliable enough, the agent *knows* that choosing the outside-right box uniquely maximizes actual value. But her epistemic position is unstable. Conditional on choosing the outside-right box, she ceases to be in a position to know that choosing the outside-right box maximizes actual value. It seems to us clear that an agent facing *Unstable Boxes like Miners* is rationally required to choose either the middle-left box or the middle-right box, and rationally permitted to choose either. Even when an agent knows which option uniquely maximizes actual value, rational permission and actual value maximization can come apart, if the agent's knowledge is unstable. This is a somewhat surprising result.

But, as concerns *Classic Newcomb*, the real substance of the guiding rule of actual value is what it says about the coincidence between rational permission and actual value maximization: namely, that rational permission and actual

value maximization cannot come apart if the agent has stable access to actual value.

The simplest cases in which an agent has stable access to actual value are cases in which the agent knows the actual values of her options. (Imagine an agent choosing among transparent boxes, each containing a sum of money.) *Classic Newcomb* is interesting in part because it is a case in which the agent has stable access to actual value without being in a position to know the actual values of her options. The agent is not in a position to know whether the actual value of A_{2B} is 1,000 or 1,001,000, for example, because she does not know whether the opaque box contains \$0 or \$1,000,000. Nevertheless, she is in a position to know that taking A_{2B} (uniquely) maximizes actual value, and, conditional on A_{2B}, she still is in a position to know that A_{2B} (uniquely) maximizes actual value.

The guiding rule of actual value entails that if an agent is stably in a position to know of an option that it uniquely maximizes actual value, the agent is rationally required to choose the option. This claim is the crucial premise of the *Objective Argument* for two-boxing:

(P1) If an agent is stably in a position to know of an option that it uniquely maximizes actual value, then the agent is rationally required to choose the option.

(P2) An agent facing *Classic Newcomb* is stably in a position to know of taking both boxes that it uniquely maximizes actual value.

(C) Therefore, an agent facing *Classic Newcomb* is rationally required to take both boxes.

The Causal Dominance Argument and the Objective Argument are closely related. If A_i causally dominates A_j, then, unless the agent is otherwise epistemically disabled, the agent is stably in a position to know that the actual value of A_i exceeds the actual value of A_j. Pointing out that taking both boxes causally dominates taking only the opaque box therefore helps to justify the minor premise of the Objective Argument.

The crucial difference between the Causal Dominance Argument and the Objective Argument lies in their respective major premises.[37] The major premise of the Objective Argument amounts to the claim that agents are

[37]Ahmed (2014a, ch. 7) claims that the best argument for two-boxing goes through a principle, akin to Causal Dominance, which he calls CDB: "If you know that a certain available option makes you worse off, given your situation, than you would have been on some identifiable alternative, then that first option is irrational" (p. 202). He then formulates a weaker principle, CDB-sequence: "If you know that a certain available sequence of choices makes you worse off, given your situation, than you would have been on some identifiable

rationally required to be guided by actual value when they are capable of being guided by actual value. Or to put the point in deontological terms (since agents are always objectively required to choose so as to maximize actual value): in the rare cases in which the objective 'ought' provides the agent with guidance, the guidance provided by the rational 'ought' cannot conflict with the guidance provided by the objective 'ought'.[38]

The major premise of the Causal Dominance Argument—namely, Causal Dominance—is refuted by cases like *The Dominating Fire* and *The Semi-Frustrater*. But such cases pose no threat to the major premise of the Objective Argument, since they are not cases in which the agent has stable access to actual value.

2.9 Objective Guidance

The second premise of the Objective Argument is uncontroversial. An agent facing *Classic Newcomb* is stably in a position to know of taking both boxes that it uniquely maximizes actual value. If one-boxers want to resist the Objective Argument, they must reject the guiding rule of actual value.

The guiding rule of actual value is an instance of a schema that all sides should accept. The schema involves two elements: the objective value quantity, which is the value quantity the maximization of which makes options objectively permissible to choose, and the guidance relation, which an agent facing a decision problem bears to value quantities. The schema is an objective guidance constraint on the rational 'ought':

Objective Guidance: If an agent facing a decision problem bears

alternative, then that first sequence is irrational" (p. 211, italics original). He offers a counterexample to CDB-sequence and argues that "accepting CDB and not CDB-sequence looks completely unmotivated" (p. 211). As it turns out, both *Unstable Boxes like Miners* and *The Semi-Frustrater* are counterexamples to CDB. But we do not need anything nearly as strong as CDB to motivate two-boxing. Neither *Unstable Boxes like Miners* nor *The Semi-Frustrater* are counterexamples to the guiding rule of actual value. As for Ahmed's counterexample to CDB-sequence—namely, *Newcomb Insurance*—it matters whether there is a single choice or a sequence of choices, since the value quantities to which the agent has stable access depends on it. If there is a single choice, even a single choice among sequences, we agree with Ahmed's judgments. If there is a sequence of choices, each among non-sequential options, we agree with the recommendations of the rule of c-expected value.

[38]Kotzen (MS) also proposes a connection between the objective 'ought' and the rational 'ought' and suggests that the proposed connection justifies two-boxing. In broad strokes, we agree. But at the level of detail, our suggestion is importantly different from Kotzen's. Kotzen's proposed connection, unlike ours, does not require *stable* access. As such, it is false: *Unstable Boxes like Miners* is a counterexample, as is a variation on *The Semi-Frustrater* in which the agent knows that she will not press the black button.

> the guidance relation to the objective value quantity, then the options that are rationally permissible for the agent to choose must maximize the objective value quantity.

To get from Objective Guidance to the guiding rule of actual value, we need two further claims: that the guidance relation is stable access, and that actual value is the objective value quantity.

One-boxers could disagree with our claim that the guidance relation is stable access, but this will not be of much help in resisting the Objective Argument. After all, the Objective Argument is not wedded to any particular conception of guidance. It can be recast using any conception of guidance on which an agent facing *Classic Newcomb* bears the guidance relation to actual value, and, so far as we can tell, *every* plausible conception of guidance is one on which an agent facing *Classic Newcomb* bears the guidance relation to actual value. To resist the Objective Argument, one-boxers therefore must deny that actual value is the objective value quantity.

In principle, there are three ways that one-boxers could deny that actual value is the objective value quantity: they could deny that there is an objective 'ought'; they could grant that there is an objective 'ought', but deny that there is an objective value quantity; or they could argue that some value quantity besides actual value is the objective value quantity. Only the third way is plausible. The first two ways effectively abandon the consequentialist common ground between one-boxers and two-boxers. One-boxers and two-boxers should agree that, when an agent faces a decision problem, there are facts about which options are objectively permissible for the agent to choose, and that what makes an option objectively permissible for the agent to choose is the maximization of some value quantity, the objective value quantity. The dispute between one-boxers and two-boxers thus reduces to a dispute about what the objective value quantity is. Two-boxing is true if actual value is the objective value quantity. For one-boxing to be true, some other value quantity would have to be the objective value quantity.

It is not clear what one-boxers could take the objective value quantity to be. Heretofore, it has been common ground between one-boxers and two-boxers that actual value is the objective value quantity. The most promising proposal we have yet to see was suggested to us by Arif Ahmed, in personal communication. Ahmed, a committed one-boxer, suggests that one-boxers take *e-actual value* to be the objective value quantity.

We can introduce e-actual value in a way that makes its similarity to actual value clear. Actual value, an explicitly causal notion, can be explicated using e-expected value and knowledge. Let $\mathcal{S} = \{S_1, \ldots, S_n\}$ be the finest partition of causally act-independent states. Let $S_@$ be the member of $\mathcal{S}$ that is true at the actual world, and let $C^{S_@}$ be the agent's credence function conditionalized

on $S_@$. The *actual value* of an option is then equal to the e-expected value of the option relative to $C^{S_@}$. The e-actual value of an option also can be explicated using e-expected value and knowledge. Say that a state, T, is evidentially act-independent just if T is compossible with each $A \in \mathcal{A}$ and, for each $A \in \mathcal{A}$, $C(T) = C(T \mid A)$. Let $\mathcal{T} = \{T_1, \ldots, T_m\}$ be the finest partition of evidentially act-independent states. Let $T_@$ be the member of $\mathcal{T}$ that is true at the actual world, and let $C^{T_@}$ be the agent's credence function conditionalized on $T_@$. The *e-actual value* of an option is then equal to the e-expected value of the option relative to $C^{T_@}$. We can think of it this way, then: the actual value of an option is what the agent expects the value of the option to be, given full causal knowledge (that is, knowledge of which member of the finest partition of causally act-independent states is true), and the e-actual value is what the agent expects the value of the option to be, given full evidential knowledge (that is, knowledge of which member of the finest partition of evidentially act-independent states is true).

There is nothing incoherent about the suggestion that e-actual value is the objective value quantity. But if one-boxing stands and falls with the claim that e-actual value is the objective value quantity, then one-boxers are in serious trouble; for the claim that agents are always objectively required to choose so as to maximize e-actual value is highly unintuitive. To see this, return to two examples from above.

Consider *The Frustrater*, and let us suppose that box A contains \$100, box B contains \$0, and the envelope contains \$40. It seems clear, then, that an agent facing *The Frustrater* is objectively required to take box A, the most lucrative box. But if e-actual value is the objective value quantity, then an agent facing *The Frustrater* is objectively required to take the envelope. The true member of the finest partition of evidentially act-independent states, $T_@$, does not specify how much money is in box A, since any state that specifies how much money is in the boxes is not evidentially act-independent. Hence, the option that uniquely maximizes e-actual value is taking the envelope.

In *Boxes like Miners*, actual value and e-actual value coincide. Taking the right box uniquely maximizes both. But if we imagine that the agent regards herself as slightly intuitive—that the agent's credence that the left box contains \$10 conditional on taking the left box is slightly greater than her unconditional credence that the left box contains \$10—then, although taking the right box still uniquely maximizes actual value, and although it still seems that the agent is objectively required to take the right box, taking the middle box uniquely maximizes e-actual value.

There may be other proposals one-boxers could produce about what the objective value quantity is, and we can evaluate them individually. But we doubt that any will be plausible. To us, it seems obvious that an agent facing

Classic Newcomb is objectively required to take both boxes, so it seems obvious that the objective value quantity, whatever it proves to be, is uniquely maximized by two-boxing in *Classic Newcomb*.

Strictly speaking, the Objective Argument does not require that actual value be the objective value quantity or that stable access be the guidance relation. All that it requires is three claims: (1) that Objective Guidance is true; (2) that two-boxing in *Classic Newcomb* uniquely maximizes the objective value quantity; and (3) that an agent facing *Classic Newcomb* bears the guidance relation to the objective value quantity.

That being said, we believe that the objective value quantity is actual value, and we believe that the guidance relation is stable access. In our view, the guiding rule of actual value is virtually undeniable.

2.10 Explaining the Counterexamples

The foregoing discussion provides us not only with a sound argument for two-boxing, but also with the resources needed to explain why Causal Dominance admits of counterexamples.

Nothing about what an agent rationally ought to do follows from the relations of causal dominance among the options. Of course, both actual value and c-expected value respect causal dominance, so, if A_i causally dominates A_j, the actual value of A_i exceeds the actual value of A_j, and the c-expected value of A_i exceeds the c-expected value of A_j. But nothing about what an agent rationally ought to do follows from the actual values of the options, and nothing about what an agent rationally ought to do follows from the c-expected values of the options. There is no direct connection between dominance or value quantity maximization and rational choice. In order to derive conclusions about what an agent rationally ought to do, we need to know, in addition to the facts about which options maximize which value quantities, the facts about which value quantities the agent has stable access to.

Once we appreciate that stable access mediates the connection between value quantity maximization and rational choice, we can explain the pattern of counterexamples to Causal Dominance that we find. Since both actual value and c-expected value respect causal dominance, and since both the guiding rule of actual value and the guiding rule of c-expected value are true, we should expect counterexamples to Causal Dominance to arise when, but only when, agents lack stable access both to actual value and to c-expected value. This is exactly what we find. There are non-ideal counterexamples to Causal Dominance because a non-ideal agent may lack stable access both to actual value and to c-expected value, despite the fact that one of her options causally dominates another (e.g., *The Dominating Fire*). There are ideal counterexamples

to Causal Dominance because an ideal agent might lack stable access both to actual value and to c-expected value, despite the fact that one of her options causally dominates another (e.g., *The Semi-Frustrater*). There are non-ideal counterexamples to Causal Dominance Selection because a non-ideal agent might lack stable access both to actual value and to c-expected value, despite the fact that one of her options causally dominates all others (e.g., *The Dominating Fire*). There are no ideal counterexamples to Causal Dominance Selection because an ideal agent is guaranteed to have stable access to actual value if one of her options causally dominates all others.[39]

[39]An ideal agent knows that the states $\mathcal{S} = \{S_1, S_2, \ldots, S_n\}$ are causally act-independent. Hence, for each $A \in \mathcal{A}$ and each $S \in \mathcal{S}$, she knows that $(S \leftrightarrow [A \boxdot\!\!\rightarrow S])$. The states are causally act-independent, so there is some $S_j \in \mathcal{S}$ such that, for any $A \in \mathcal{A}$, $av(A) = V(AS_j)$. Hence, if option A_i dominates all other options, the ideal agent knows that A_i uniquely maximizes actual value. Moreover, conditional on A_i, the $V(AS)$'s remain unchanged, and the ideal agent still knows that some $S \in \mathcal{S}$ obtains; hence the ideal agent still is in a position to know that A_i uniquely maximizes actual value.

Chapter 3

Evidence and Rationalization

3.1 Introduction

Suppose that you have to take a test tomorrow. Naturally you'd prefer not to study. But you should study, since you care very much about passing and expect to pass only if you study. Is there anything you can do to change that? Is there any way for you to "rationalize" slacking off?

If you could somehow stop caring about passing the test, then you would be under no rational pressure to study. In general, changing what one cares about is a way of changing what one rationally ought to do. But changing what one cares about is not always an option. It's not as easy as saying "I don't care if I fail." Let's assume that in this case you're not able to kick the desire to pass. At least not within the next 24 hours.

What else might you do? You could drink a bottle of mouthwash. Doing that would make studying irrational, since having drunk the mouthwash you should rather go to the hospital. (You care about passing the test, but much more about keeping your liver.) Then again, drinking the mouthwash would itself be irrational. So it wouldn't really accomplish what you sought in the first place: to avoid studying without thereby doing anything irrational. Nor would it allow you to slack off the rest of the night. By drinking the mouthwash you simply trade one unpleasant obligation (studying) for another (a trip to the emergency room).

If you had some way of forgetting about the test—for example, by taking a memory-erasing pill—then you could rationalize slacking off, since having taken the pill you would no longer have any reason to study. But like drinking the mouthwash, taking the pill would be irrational. For you now expect that it would result in your not studying and failing the test—an undesirable outcome.

What if you could take a different kind of pill, one that would cause you to know everything you need to know to pass the test? In that case you could

rationalize slacking off, since having taken the pill you would no longer need to study. And unlike taking the memory-erasing pill, taking this pill would be rational, since you expect that it would result in your passing the test. So it seems that by taking the pill you can tailor the demands of rationality to your liking, without thereby doing anything irrational.

But taking a knowledge-inducing pill is a suspiciously easy way to rationalize slacking off. There is a simple diagnosis of why that is. The knowledge induced by the pill—consisting of facts about the subject matter of the test—does not play an essential role in the explanation of why, after taking the pill, it is rational for you to slack off. After all, simply *believing* that you took the pill (whether or not you actually took it) is enough to rationalize slacking off.

Let's assume that there is nothing you can do such that simply believing that you did it would rationalize slacking off. In particular, you don't have access to knowledge-inducing pills. Are you therefore stuck with the obligation to study, or is there something else you could do to bypass that obligation?

A natural thought is that you might investigate whether studying really will help you pass the test. For if the investigation led you to sufficiently doubt the connection between studying and passing, then slacking off would become rationally permissible. If, on the other hand, the investigation reinforced the connection, then you would once again find yourself obligated to study.

What kind of investigation might you undertake? Suppose that many students just like you have faced decisions just like yours in the past. There is a trustworthy record book documenting, for each student, whether the student studied and whether the student passed. The book aggregates the data, giving overall pass rates for studying and not studying. Before reading the book, you believe that the pass rate for studying is high while the pass rate for not studying is low. So you expect the book to reflect these estimates. But it could be that the book says that the overall pass rates for studying and not studying are equal. If it does, then after reading the book it would be rational for you to slack off.

Before saying anything more about the record book investigation, let me make two clarifications. First, it's no surprise that we are sometimes able to rationalize a choice without seeing the rationalization coming. Suppose that you have the option of buying a bet that pays if you develop lung cancer before age 70. At present you believe that you don't have lung cancer, that lung cancer doesn't run in your family, and that you've never smoked and never will. So you shouldn't buy the bet. But now imagine that, before deciding whether to buy the bet, you have the option of getting a CAT scan. Since you believe that you don't have cancer, you expect that the scan will reveal nothing ominous. But, to your dismay, the scan reveals a malignant tumor. Now you have done something—getting the CAT scan—that made it the case

that you should buy the bet. So in a sense you have rationalized buying the bet. But the credit should really go—not to you—but to the world, for setting you straight on your unfortunate condition. Although it was your investigation that ended up rationalizing the choice, you played no part in orchestrating the rationalization.

Second, it's no surprise that we are sometimes able to rationalize a choice without being 100% certain, in advance, that the rationalization is coming. Suppose that you must pick one of two planes to board: plane A or plane B. Each plane has 50 passengers, and you know that of the 100 total passengers exactly one is carrying explosives. You have no idea which plane the bomber is on. But you have the opportunity to perform an experiment before making your choice: first, you choose a plane to investigate; then, one passenger from the chosen plane is selected at random and scanned for explosives; then, you're shown the results of the scan. Suppose that you wisely decide to perform the experiment. Before seeing the results of the experiment, you're 99% confident that the scan will show nothing unusual on the selected passenger. And you know that if that happens, you will have a very slight reason to prefer boarding the plane from which the subject of the experiment was selected. So you're 99% confident that investigating plane A will rationalize boarding plane A. And you're 99% confident that investigating plane B will rationalize boarding plane B. As it happens, you choose to investigate plane A and the scan shows nothing unusual, as expected. So by investigating plane A, you have rationalized boarding plane A, and you have done so in a way that you expected, with 99% confidence, to rationalize boarding plane A. Still, you don't deserve all the credit for the rationalization. You needed a little help from the world. For if the test had come up positive for explosives, then you would have had extremely strong reason to board plane B and your attempt at rationalizing boarding plane A would have failed.

Returning to your decision of whether to study, it's now clear why reading the record book is not a satisfying way of rationalizing slacking off. Although reading the book could rationalize slacking off, you're not confident, let alone certain, that it will. (After all, if you were confident that the record book contained evidence against the connection between studying and passing, then it wouldn't be rational for you to study in the first place.) What you seek when you seek to rationalize slacking off is not just to do something that might rationalize slacking off, but rather to do something that you can foresee with certainty will rationalize slacking off. You want to guarantee the rationalization of your preferred choice.

But so far we have seen no reason to think that you can. Indeed, just thinking abstractly about the matter, the kind of guaranteed rationalization described above is hard to countenance. If rationality was such that one could

manipulate its demands in foreseeable ways, then it would seem too easy to be rational. Rationality—the subjective, practical flavor of rationality under discussion—is supposed to be within our ken, in some important sense, but it is not supposed to be completely under our control. Just as the correct rule of rational action cannot be "do whatever is objectively best," so too it cannot be "do whatever you want." A good theory of practical rationality should strike a balance between these two extremes. But a theory that permitted rationalization would seem to tread too closely to the "do whatever you want" extreme. And in so doing, such a theory would seem to be stripped of any normative force. How could we feel pressure to follow the demands of the theory, when *we* decided what those demands are? Perhaps it is hard to find examples of rationalization because rationalization is impossible.

3.2 Mathematical Background

If rationalization is impossible, then a recently defended theory of practical rationality, known as *evidential decision theory*, stands refuted.[1] For this theory permits rationalization. In §3.4 I will describe an example in which evidential decision theory permits rationalization. In this section and the next, I will introduce the theory, its main rival, and the model in which both are formulated.

The main element of the model is a quadruple $\langle \mathcal{A}, \mathcal{S}, p, u \rangle$, called a *decision problem*, in which $\mathcal{A}$ is a finite partition of propositions $a_1, \ldots, a_m$ representing the *actions* available to a particular agent at a particular time, $\mathcal{S}$ is a finite partition of propositions $s_1, \ldots, s_n$ representing the possible *states* of the world upon which the consequences of the actions depend, p is a probability function representing the agent's *degrees of belief* at the specified time, and u is a utility function representing the agent's *non-instrumental desires* at the time. The actions $a \in \mathcal{A}$ and states $s \in \mathcal{S}$ are chosen so that each conjunction as picks out a unique *consequence*, fully specifying how things stand with respect to everything that matters to the agent. Such consequences form the domain of the agent's utility function.

Various decision theories can be formulated within this model. Given a decision problem $\langle \mathcal{A}, \mathcal{S}, p, u \rangle$, the *simple expected utility* of an action $a_i \in \mathcal{A}$ is identified with a weighted sum, for each state, of the utility of taking the action

[1] The theory originated with Jeffrey (1965) and has been most recently and extensively defended by Ahmed (2014a).

in a world in that state, weighted by the probability of the state obtaining:

$$SEU(a_i) = \sum_{j=1}^{n} p(s_j)u(a_i s_j).$$

Simple decision theory (SDT) enjoins agents to choose among the $a \in \mathcal{A}$ so as to maximize SEU.[2]

The inadequacy of SDT is well known and easy to illustrate. Suppose that Al is deciding whether to smoke. He believes that smoking has a strong tendency to cause cancer and, while he finds some pleasure in smoking, that pleasure is dwarfed by the displeasure he associates with cancer. Given what he believes and desires, Al should not smoke.

But SDT advises that Al smoke. Let $\mathcal{S}$ include the proposition that Al gets cancer and its negation. Let $-\gamma$ be the large amount of disutility associated with cancer, let $\delta > 0$ be the small amount of utility associated with smoking, and set an arbitrary zero where both cancer and smoking are absent.

	cancer	$\neg$*cancer*
smoke	$-\gamma + \delta$	δ
$\neg$*smoke*	$-\gamma$	0

Figure 3-1: The payoff table for the smoking problem.

If x is Al's degree of belief in the proposition that he will get cancer, then his simple expected utilities are related by the following equation:

$$\begin{aligned} SEU(smoke) &= x(-\gamma + \delta) + (1 - x)\delta \\ &= -x\gamma + \delta > -x\gamma = SEU(\neg smoke). \end{aligned}$$

Hence, SDT enjoins Al to smoke, in spite of his belief that smoking causes cancer and his strong desire to avoid cancer.

To handle this kind of problem, Jeffrey (1965) proposed a different definition of expected utility. Whereas simple expected utility weights utilities by unconditional probabilities in states, Jeffrey's definition (EEU) weights

[2]This theory is sometimes associated with Savage (1954). However, Savage's state space is not a space of possible worlds. If we wish to interpret it as such, then plausibly Savage intended his states to form a privileged partition on that space—essentially a partition of dependency hypothesis, a la Lewis (1981). So Savage's theory is best understood as an early version of causal decision theory, for which the smoking problem does not arise. Thanks to Bob Stalnaker on this point.

utilities by conditional probabilities in states, conditional on actions:

$$EEU(a_i) = \sum_{j=1}^{n} p(s_j|a_i)u(a_i s_j).$$

Applying Jeffrey's definition to the smoking problem:

$$\begin{aligned} EEU(smoke) &= p(cancer|smoke)(-\gamma + \delta) + (1 - p(cancer|smoke))\delta \\ &= -p(cancer|smoke)\gamma + \delta. \\ EEU(\neg smoke) &= -p(cancer|\neg smoke)\gamma. \end{aligned}$$

Therefore, not smoking maximizes EEU iff the difference

$$p(cancer|smoke) - p(cancer|\neg smoke)$$

exceeds the fraction δ/γ. In other words, not smoking maximizes EEU iff Al regards smoking as sufficiently strong evidence of cancer—a condition plausibly satisfied by the description of the smoking problem. (The required strength of the evidential connection lowers as γ increases and δ decreases.)

Since Jeffrey's definition uses conditional probabilities of states on actions, and since the differences between these conditional probabilities measure the extent to which the agent regards actions as evidence of states, Jeffrey's brand of expected utility has come to be known as *evidential expected utility*, and the theory enjoining agents to maximize EEU has come to be known as *evidential decision theory* (EDT).

Although EDT gives rational advice in the smoking problem, some believe that it does so only incidentally. For these theorists, Al is irrational to smoke because he believes that smoking *causes* cancer, not because he regards smoking as evidence of cancer. The shift in emphasis makes no difference in the smoking problem because smoking is regarded both as a cause and as evidence of cancer. But it makes a difference elsewhere, such as in the notorious Newcomb problem:[3]

> *Newcomb*: There is a transparent box containing \$1,000 and an opaque box containing either \$1,000,000 (*full*) or nothing (*empty*). Ted has two options: he can take just the opaque box (*onebox*) or he can take both boxes (*twobox*). The content of the opaque box was determined yesterday by a reliable predictor. The opaque box contains \$1,000,000 iff the predictor predicted that Ted would take just the opaque box.

[3]Attributed to physicist William Newcomb, *Newcomb* was popularized by Nozick (1969).

	full	*empty*
twobox	1,001,000	1,000
onebox	1,000	0

Figure 3-2: The payoff table for *Newcomb*.

Supposing for simplicity that Ted's utilities are linear and increasing in dollars, the evidential expected utilities of his options in *Newcomb* are:

$$\begin{aligned} EEU(twobox) &= p(full|twobox)(1,001,000) + (1 - p(full|twobox))(1,000) \\ &= p(full|twobox)(1,000,000) + 1,000. \\ EEU(onebox) &= p(full|onebox)(1,000,000). \end{aligned}$$

Therefore, one-boxing maximizes EEU iff the difference

$$p(full|onebox) - p(full|twobox)$$

exceeds the fraction $1/1,000$. In other words, EDT advises one-boxing so long as Ted regards one-boxing as at least a little evidence that the opaque box contains \$1,000,000—a condition satisfied by the description of *Newcomb*.

Those who reject EDT's diagnosis of the smoking problem also find fault in its treatment of *Newcomb*. They maintain that Ted should take both boxes, since he knows that his actions have no causal effect on the content of the opaque box, and since—no matter what the opaque box contains—two-boxing nets \$1,000 more than one-boxing.[4]

Many of those who reject EDT accept an alternative decision theory. As formulated by Lewis (1981), the alternative theory is essentially a return to SDT, with one caveat. Whereas in SDT the set of states can be any partition of logical space, Lewis requires that the states be *(causal) dependency hypotheses*. A dependency hypothesis, for an agent at a time, is a proposition fully specifying how things the agent cares about do or do not depend causally on the agent's present actions.[5] It is a hypothesis about the causal structure of the world, as it pertains to the decision. Lewis proves that, necessarily, the dependency hypotheses for an agent are causally independent of the actions between which the agent is deciding. So, for example, the proposition that

[4]See Spencer and Wells (Forthcoming) for a more detailed defense of two-boxing.

[5]In order to account for decision problems in which the objective chance of an action yielding a particular consequence is neither 0 nor 1, we would need to alter the framework slightly, removing the stipulation that each ac entails a unique consequence and requiring that the c specify objective conditional chances of consequences on actions. However, the decision problems discussed in this paper require no such alteration, so we will work with the simpler albeit less general framework sketched above.

the opaque box contains $1,000,000 is a dependency hypothesis for Ted in *Newcomb*, whereas getting cancer is not a dependency hypothesis for Al in the smoking problem.

Replacing the state space of SDT with a partition of dependency hypotheses $\mathcal{C} = \{c_1, \ldots, c_n\}$, we can characterize a third kind of expected utility:

$$CEU(a_i) = \sum_{j=1}^{n} p(c_j)u(a_i c_j).$$

Since the concept of a dependency hypothesis is causal by definition, this kind of expected utility has come to be known as *causal expected utility*, and the theory enjoining agents to maximize CEU has come to be known as *causal decision theory* (CDT).

CDT directly opposes EDT in *Newcomb*. If x is Ted's degree of belief in the proposition that the opaque box contains $1,000,000, then his causal expected utilities are related by the following equation:

$$\begin{aligned} CEU(twobox) &= x(1,001,000) + (1-x)1,000 \\ &= 1,000,000x + 1,000 > 1,000,000x = CEU(onebox). \end{aligned}$$

Hence, CDT enjoins Ted to take both boxes.

3.3 Evidence Gathering

Each of the decision problems considered so far is non-sequential, in the sense that there is just one set of options and the choice between the members of that set occurs at a single time. A sequential decision problem, on the other hand, has at least two sets of options corresponding to two different times at which a choice must be made.

Sequential decisions are common. Often when we make a decision it is just one move in a chain of subsequent decisions. One particularly common kind of sequential decision problem involves evidence gathering. Often we decide to ask a question, make an observation, look up something on the internet or perform an experiment before proceeding further with our lives. We saw an example of this in §3.1, with the decision of whether to read the pass-fail data before deciding whether to study.

The problem that I will present in the next section—to illustrate the possibility of rationalization under EDT—is a sequential problem involving evidence gathering. It will take a little work to see exactly how to apply EDT and CDT to this kind of problem. The purpose of this section is to extend the framework

of §3.2 so that we can more easily apply the theories presented in that section to the problem of the next section.

Start with a simple two-stage sequential decision problem where the options include gathering more information before acting.

> *Simple Problem.* There are two opaque boxes before you: A and B. One contains \$100; the other is empty. You're 75% confident that A contains \$100. You may take either box, but not both. Alternatively, you may look inside A and then take a box.

	$fullA$	$fullB$
A	100	0
B	0	100

Figure 3-3: The payoff table for the *Simple Problem.*

In the *Simple Problem* your first set of options includes looking in A, taking A straightaway, or taking B straightaway. We already know how to calculate the expected utilities of the latter two options and it is clear that the expected utility of taking A straightaway (75) exceeds the expected utility of taking B straightaway (25). The question is whether the expected utility of taking A straightaway also exceeds that of looking in A before choosing a box.

Let us assume that, if you look in A, you will act rationally thereafter, in the sense that you will update your degrees of belief by conditionalizing on the truth about what is in the box, and that after updating you will choose the option that maximizes expected utility relative to your updated degrees of belief. So, for example, if you see that A contains \$100, your new expected utility for taking A will be 100, and you will take A. And if you see that A contains nothing, your new expected utility for taking B will be 100, and you will take B. So in either case, you will choose an action with expected utility 100, and you can be certain of this. So the expected utility of looking in A is itself 100, i.e. greater than that of taking A straightaway. Hence, the uniquely rational choice is to look in A before choosing a box.

To generalize the informal reasoning above, we take as given a partition of propositions $\mathcal{E} = \{e_1, \ldots, e_m\}$ representing the possible pieces of evidence that you might learn by making a particular observation. We then define the expected utility of using a particular piece of evidence $e_k \in \mathcal{E}$ to inform your decision (call this act use_{e_k}) as the expected utility of the action that maximizes expected utility relative to your updated-on-e_k degrees of belief. This definition yields causalist and evidentialist formulae:

$$CEU(use_{e_k}) = \max_i \sum_{j=1}^{n} p(c_j|e_k)u(a_i c_j). \tag{3.1}$$

$$EEU(use_{e_k}) = \max_i \sum_{j=1}^{n} p(s_j|e_k a_i) u(a_i s_j). \tag{3.2}$$

Next we define the expected utility of gathering and using the evidence gathered (call this action *look*) as a weighted sum, for each possible piece of evidence, of the expected utility of using that piece of evidence, weighted by the probability that the piece of evidence is true (i.e. that it will be the piece of evidence gathered). This definition also yields causalist and evidentialist formulae:

$$CEU(look) = \sum_{k=1}^{m} p(e_k) CEU(use_{e_k}). \tag{3.3}$$

$$EEU(look) = \sum_{k=1}^{m} p(e_k|look) EEU(use_{e_k}). \tag{3.4}$$

Applying these formulae to the *Simple Problem*, it is straightforward to confirm that looking in A maximizes both CEU and EEU.

There is an alternative version of *Newcomb* that has garnered some attention in the literature.[6] This alternative version may seem to supply a case in which EDT permits rationalization. In fact, it does not. But it is instructive to see why it does not. Here is the problem:

> *Viewcomb*: Everything is the same as in *Newcomb*, only now Ted has the option of looking inside the opaque box before making his decision.

According to EDT, Ted should one-box straightaway. For suppose that Ted looks in the box and sees that it is full. Then the act that maximizes EEU relative to his updated degrees of belief is two-boxing, and its EEU is 1,001,000. Suppose on the other hand that Ted sees that the box is empty. Then the act that maximizes EEU relative to his updated degrees of belief is again two-boxing, although its EEU in this case is 1,000. Hence, by equation (3.2),

$$EEU(use_{e_{full}}) = 1,001,000; \text{ and,}$$
$$EEU(use_{e_{empty}}) = 1,000.$$

[6] See, for example, Gibbard and Harper (1978), Adams and Rosenkrantz (1980), Skyrms (1990), Arntzenius (2008), Meacham (2010), Ahmed (2014a) and Hedden (2015b).

Plugging these values into equation (3.4), we have:

$$\begin{aligned} EEU(look) &= p(full|look)EEU(use_{e_{full}}) + p(empty|look)EEU(use_{e_{empty}}) \\ &= p(full|look)(1,001,000) + p(empty|look)(1,000) \end{aligned}$$

Notice that, for Ted, *look* entails two-boxing, since he is certain that he will two-box no matter what he learns by looking. Plausibly, then, $p(full|look) = p(full|twobox)$ and $p(empty|look) = 1 - p(full|twobox)$. Supposing for concreteness that the predictor is believed to be 60% reliable, we have:

$$\begin{aligned} EEU(look) &= (.4)(1,001,000) + (.6)(1,000) \\ &= 401,000. \end{aligned}$$

Note also that

$$\mathtt{max}_i\ EEU(a_i) = EEU(onebox) = (.6)(1,000,000) = 600,000.$$

Hence,

$$EEU(look) = 401,000 < 600,000 = \mathtt{max}_i\ EEU(a_i).$$

Hence, EDT recommends that Ted one-box straightaway.

Note that in *Viewcomb* Ted can change EDT's recommendation as he pleases. Although at the outset EDT recommends that Ted one-box, it is within Ted's power to look in the opaque box, and he can be certain, in advance, that if he looks in the box, EDT will thereafter recommend that he two-box. We thus *seem* to have a case in which EDT countenances a rationalization of the kind discussed in §3.1.

But we do not. The reason is that EDT does not permit looking in the box. From an evidentialist perspective, looking in the box (to avoid one-boxing) is just like drinking the mouthwash or taking the memory-erasing pill (to avoid studying). In each case, one is able to change the demands of rationality in a foreseeable way, but only by first doing something irrational. There is nothing odd about such irrational rationalizations. The odd rationalization is that which is itself rational. Our question is whether there is a theory of rationality that *permits* rationalization.

3.4 The Switch Problem

Although EDT does not permit rationalization in *Viewcomb*, there exists a decision problem in which it does. Here is the problem:

The Switch Problem: There are two opaque boxes, A and B. One contains \$100. The other is empty. Sue may take A ($TakeA$) or B ($TakeB$) but not both. Additionally, there are two colored switches, one red and one green, blocked from Sue's view. Each switch is either on or off. Before choosing a box, Sue may look at the red switch ($LookR$) or the green switch ($LookG$) but not both. The statuses of the switches and the contents of the boxes were determined in advance, by a predictor, in the following way.

If the predictor predicted that Sue would take A ($PredA$), she tossed a fair coin, put \$100 in A (InA) if it landed heads (H), and tossed another coin. If the second coin landed heads, she flipped both switches on (RG). If it landed tails, she flipped just the green switch on ($\neg RG$). Alternatively, if the first coin landed tails (T), she put \$100 in B (InB), and tossed another coin. If the second coin landed heads, she flipped just the green switch on. If it landed tails, she flipped both switches off ($\neg R\neg G$).

If the predictor predicted that Sue would take B ($PredB$), she tossed a coin, put \$100 in B if it landed heads, and tossed another coin. If the second coin landed heads, she flipped both switches on. If it landed tails, she flipped just the red switch on ($R\neg G$). Alternatively, if the first coin landed tails, the predictor put \$100 in A, and tossed a second coin. If it landed heads, she flipped just the red switch on. If it landed tails, she flipped both switches off.

Sue is fully aware of the foregoing details, which are summarized in figures 3-4 and 3-5.

	$PredA$				$PredB$			
	InA		InB				InA	
	RG	$\neg RG$	$\neg RG$	$\neg R\neg G$	RG	$R\neg G$	$R\neg G$	$\neg R\neg G$
$TakeA$	100	100	0	0	0	0	100	100
$TakeB$	0	0	100	100	100	100	0	0

Figure 3-4: The payoff table for *The Switch Problem*.

Here is how to interpret figure 3-4. For each cell containing a number, the number in the cell represents the payoff of choosing the option that is directly left of the cell, at a world in which each proposition directly above the cell is true. So, for example, the number 100 in the top-left cell represents the payoff of taking box A at a world in which each proposition directly above the cell is true, i.e. a world in which the predictor predicted that A would be taken, put \$100 in A, and flipped both switches on.

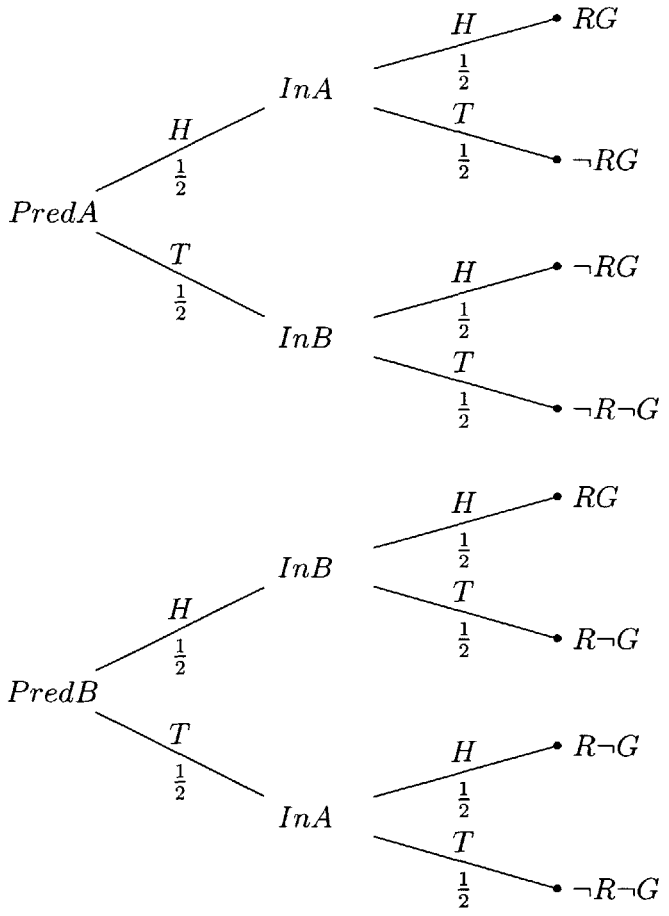

Figure 3-5: Probability trees representing the probabilistic relations in *The Switch Problem.*

In *The Switch Problem*, EDT permits Sue to manipulate rationality in foreseeably contradictory ways. By this I mean the following:

> **Claim 1**. Sue is certain that if she looks at the red switch, EDT will require that she take box A.
> **Claim 2**. Sue is certain that if she looks at the green switch, EDT will require that she take box B.
> **Claim 3**. EDT permits Sue to look at either switch.

To prove these claims I will make two simplifying assumptions, each of which may be relaxed without loss. First, I will assume that Sue believes

the predictor to be 100% reliable.[7] Second, I will assume that Sue's utilities are linear and increasing in dollars. I will also carry over the 'transparency' assumption from the discussion of evidence gathering in §3.3. That is, I will assume that Sue is certain that if she looks at a switch, she will act rationally thereafter, in the sense that she will update her degrees of belief by conditionalizing on the truth about the switch, and that after updating she will choose the option that maximizes expected utility relative to her updated degrees of belief.[8] Let p be Sue's probability function before she decides whether to look at a switch. For any proposition e, let p_e be Sue's credence function after conditionalizing on e: $p_e(.) = p(.|e)$.

To prove **Claim 1**, it suffices to prove the embedded conditional from premises of which Sue is certain. Suppose that Sue looks at the red switch. Then, either she learns R or she learns $\neg R$.

Suppose that she learns R. Then her new probability function is p_R. Note that there is only one possibility in which the red switch is on and the predictor predicted that Sue would take box A; and, in that possibility, \$100 is in box A. Moreover, conditional on her taking box A, Sue is certain that the predictor predicted that she take A. Hence, $p_R(InA|TakeA) = 1$. Note also that there are three equiprobable possibilities in which the red switch is on and the predictor predicted that Sue would take box B; and, in two of them, \$100 is in box B. Hence, $p_R(InB|TakeB) = 2/3$. Hence, after learning R, Sue's EEU of taking box A is 100 while her EEU of taking box B is 100(2/3). Hence, EDT requires that Sue take box A.

Suppose, on the other hand, that Sue learns $\neg R$. Then her new credence function is $p_{\neg R}$. Note that there are three equiprobable possibilities in which the red switch is off and the predictor predicted that Sue would take box A; and, in one of them, \$100 is in A. Hence, $p_{\neg R}(InA|TakeA) = 1/3$. Note also that there is only one possibility in which the red switch is off and the predictor predicted that Sue would take box B; and, in that possibility, \$100 is in A. Hence, $p_{\neg R}(InB|TakeB) = 0$. Hence, after learning $\neg R$, Sue's EEU of taking box A is 100(1/3) while her EEU of taking box B is 0. Hence, EDT requires that Sue take box A.

Hence, if Sue looks at the red switch, then, no matter what she learns, EDT will require that she take box A. The only premise is that Sue will conditionalize on the truth about whether the switch is on. By the transparency

[7]This assumption may lead to Sue assigning zero probability to some of her available actions, in which case the evidential expected utilities of those actions would be undefined. To avoid this, we may assume instead that the probability that the predictor is mistaken is non-zero but negligible. Thanks to Bob Stalnaker on this point.

[8]This assumption may also lead to Sue assigning zero probability to some of her available actions. As before, we can avoid this by assuming instead that irrational actions get negligible positive probability.

assumption, Sue is certain of this. Hence, Sue is certain of the conclusion. **Claim 1** follows.

To prove **Claim 2**, Suppose that Sue looks at the green switch. Then, either she learns G or she learns $\neg G$. Suppose that she learns G. Then her new probability function is p_G. Note that there is only one possibility in which the green switch is on and the predictor predicted that Sue would take box B; and, in that possibility, \$100 is in box B. Moreover, conditional on her taking box B, Sue is certain that the predictor predicted that she take B. Hence, $p_G(InB|TakeB) = 1$. Note also that there are three equiprobable possibilities in which the green switch is on and the predictor predicted that Sue would take box A; and, in two of them, \$100 is in box A. Hence, $p_R(InA|TakeA) = 2/3$. Hence, after learning G, Sue's EEU of taking box B is 100 while her EEU of taking box A is 100(2/3). Hence, EDT requires that Sue take box B.

Suppose, on the other hand, that Sue learns $\neg G$. Then her new probability function is $p_{\neg G}$. Note that there are three equiprobable possibilities in which the green switch is off and the predictor predicted that Sue would take box B; and, in one of them, \$100 is in B. Hence, $p_{\neg G}(InB|TakeB) = 1/3$. Note also that there is only one possibility in which the green switch is off and the predictor predicted that Sue would take box A; and, in that possibility, \$100 is in B. Hence, $p_{\neg G}(InA|TakeA) = 0$. Hence, after learning $\neg G$, Sue's EEU of taking box B is 100(1/3) while her EEU of taking box A is 0. Hence, EDT requires that Sue take box B.

Hence, if Sue looks at the green switch, then, no matter what she learns, EDT will require that she take box B. **Claim 2** follows by the transparency assumption.

To prove **Claim 3**, first consider the option of looking at the red switch. The associated evidence partition is $\{R, \neg R\}$. We saw above that the EEU of the option that maximizes EEU relative to p_R (namely, $TakeA$) is 100, and the EEU of the option that maximizes EEU relative to $p_{\neg R}$ (namely, $TakeA$) is 100(1/3). Moreover, **Claim 1**, together with the transparency assumption, entails that $p(TakeA|LookR) = 1$. Since $p(PredictA|TakeA) = 1$ and $p(R|PredictA) = 1/4$, it follows that $p(R|LookR) = 1/4$. Hence, by equation (3.4),

$$EEU(LookR) = 100(1/4) + 100(1/3)(3/4) = 50.$$

Next consider the option of looking at the green switch. The associated evidence partition is $\{G, \neg G\}$. We saw above that the EEU of the option that maximizes EEU relative to p_G (namely, $TakeB$) is 100, and the EEU of the option that maximizes EEU relative to $p_{\neg G}$ (namely, $TakeB$) is 100(1/3). Moreover, **Claim 2**, together with the transparency assumption, entails that

$p(TakeB|LookG) = 1$. Since $p(PredictB|TakeB) = 1$ and $p(G|PredictB) = 1/4$, it follows that $p(G|LookG) = 1/4$. Hence, the EEU of looking at the green switch is also 50.

It is simple to confirm that $EEU(TakeA) = EEU(TakeB) = 50$ as well. After all, $p(InA|TakeA) = p(InB|TakeB) = 1/2$. Hence, before looking at a switch, the EEUs of each of Sue's four options are equal. Hence, EDT initially permits Sue to take any option, including either evidence gathering option. **Claim 3** follows. Hence, EDT permits rationalization in *The Switch Problem.*

CDT handles *The Switch Problem* much more sanely. Like EDT, CDT initially permits each of Sue's four options. The question is whether the demands of CDT change after Sue looks at a switch. Whether they do depends on Sue's beliefs about which box she will ultimately decide to take. We have not yet said anything about those beliefs, so, from a causalist perspective, our problem is not yet fully specified. Let us fully specify it. Let us stipulate that Sue is maximally unsure about what she will do, so that $p(TakeA) = p(TakeB) = 1/2$. We can now show that the CEU of taking box A will always equal the CEU of taking box B, no matter what Sue learns after looking at a switch.

Suppose that Sue looks at the red switch and learns R. Then her updated causal expected utilities are as follows:

$$CEU_R(TakeA) = p_R(InA)(100).$$
$$CEU_R(TakeB) = p_R(InB)(100).$$

Since Sue's degrees of belief are split evenly over her options, they are also split evenly over the two possible predictions. Hence, the four possibilities in which the red switch is on are equiprobable. Half of those possibilities are ones in which box A contains \$100, and the other half are ones in which box B contains \$100. Hence, $p_R(InA) = p_R(InB) = 1/2$. Hence,

$$CEU_R(TakeA) = 50 = CEU_R(TakeB).$$

By the symmetry of the problem, parallel reasoning shows that

$$CEU_{\neg R}(TakeA) = 50 = CEU_{\neg R}(TakeB),$$

as well. Indeed, the same holds true for the causal expected utilities of Sue's options after looking at the green switch.

What happens if we relax the assumption that Sue is maximally unsure of what she will do? Then CDT's recommendations may change. For instance, if Sue is antecedently *certain* that she will take box A, and she sees that the

red switch is on, then CDT requires that she take box A.[9] But if, on the other hand, Sue sees that the red switch is off, CDT requires that she take box B.[10] So, although CDT sometimes changes its requirements, Sue cannot be antecedently certain about the way in which the theory will change its recommendations, since she cannot be antecedently certain that—for example—the red switch is on. As we saw in §3.1, there is nothing odd about rationalizations that aren't anticipated with certainty. So there is nothing odd about CDT's treatment of *The Switch Problem*.

3.5 Managing Rationality

The Switch Problem is similar to *Viewcomb* in that the agent can predict with certainty that if she makes a particular observation, the demands of evidential rationality will change in a particular way. In *Viewcomb*, the agent can predict with certainty that if she looks in the opaque box, EDT will no longer permit one-boxing. In *The Switch Problem*, the agent can predict with certainty that if she looks at the red switch, EDT will no longer permit taking box B (and that if she looks at the green switch, EDT will no longer permit taking A).

However, in *Viewcomb*, EDT does *not* permit gathering evidence in such a way as to predictably manipulate the demands of rationality.[11] This fact may be seen as a sort of saving face for the theory. Although it is physically possible for an agent to manipulate EDT's demands in *Viewcomb* (by looking in the

[9] *Proof:* Suppose that Sue learns R. By hypothesis, $p(TakeA) = 1$. Hence, in this case, the causal expected utility of taking box A equals its evidential expected utility, which, we have seen, is 100. The causal expected utility of taking box B is $p_R(InB)(100)$, or equivalently $p_R(InB|TakeA)$. We have seen that $p_R(InA|TakeA) = 1$. Hence, $p_R(InB|TakeA) = 0$. Hence, in this case, after learning R, Sue's causal expected utility of taking box A exceeds that of taking box B by 100.

[10] *Proof:* Suppose that Sue learns $\neg R$. By hypothesis, $p(TakeA) = 1$. Hence, in this case, the causal expected utility of taking box A equals its evidential expected utility, which, we have seen, is 100/3. The causal expected utility of taking box B is $p_{\neg R}(InB)(100)$, or equivalently $p_{\neg R}(InB|TakeA)(100)$. We have seen that $p_{\neg R}(InA|TakeA) = 1/3$. Hence, $p_{\neg R}(InB|TakeA) = 2/3$. Hence, in this case, after learning $\neg R$, Sue's causal expected utility of taking box B exceeds that of taking box A, 200/3 to 100/3.

[11] There is a famous theorem due to I. J. Good (1967) according to which it is always rational to gather more evidence before making a decision, provided that the cost of so doing is negligible. *Viewcomb* is a counterexample to the version of Good's theorem wherein 'rational' is given an evidential interpretation. EDT's violation of Good's theorem is often used as an argument against the theory, but Maher (1990) has shown that CDT also violates Good's theorem on occasion. Hence, violations of Good's theorem do not, on their own, cut any ice in the debate between EDT and CDT. Nevertheless, this paper suggests that there is a problem for EDT surrounding its treatment of evidence gathering that arises not when the theory *prohibits* the collection of cost-free evidence, but rather when it *permits* such collection.

box), it is rationally impermissible, according to the theory. If we personify EDT as an advisor, it is as if the advisor is saying: "You should one-box. Of course, if you look in the box, then, no matter what you see, I will tell you to two-box. But I don't want to tell you that, since I think you should one-box. So you shouldn't look in the box." This advice may seem odd but it is at least self-reinforcing. The advisor acknowledges that her advice will change from what it is at the outset, but she does not support that change.

The Switch Problem is different. There, EDT *permits* gathering evidence in such a way as to predictably manipulate the demands of rationality. It is as if the EDT-advisor is saying: "You are permitted to take box B. Of course, if you look at the red switch, then, no matter what you see, I will tell you that you are *not* permitted to take box B—that you should rather take box A. But I have no problem telling you that. So go ahead, look at the red switch. Or don't. I'll tell you whatever you want to hear." Not only is this advice odd, it rings of self-doubt. Not only does the advisor acknowledge that her advice will change, she apparently endorses that change at the outset.

To make things vivid, imagine that Sue, who only speaks English, has a Chinese-speaking duplicate, Lu. Suppose that Sue and Lu face *The Switch Problem* together, and they have shared interests: they will split whatever earnings they make from whichever box they collectively decide to take. Moreover, they are both convinced by evidential reasoning when it comes to making choices. At the outset Sue and Lu are indifferent as to which box to take, and they are also indifferent as to which switch to look at. As it happens, Sue looks at the red switch and Lu looks at the green switch. After looking at their respective switches Sue and Lu reconvene to decide which box to take. Sue now knows the status of the red switch (say, that it's on) and Lu knows the status of the green switch (say, that it's off), but neither knows the statuses of both switches, and they are unable to communicate with one another. Sue reaches for box A but just as she is about to take it, Lu grabs her arm and gestures for them to take box B. Sue shakes her head and points at box A. They start fighting.

But should they fight? Sue and Lu both want the same thing. And they both agree on how their beliefs and desires should combine to guide their decision. Of course, they have different beliefs: Sue believes that the red switch is on but has no idea whether the green switch is on, whereas Lu believes that the green switch is off but has no idea whether the red switch is on. There is nothing odd about two people with different beliefs rationally disagreeing about what to do, even when their interests align. What is odd is that Sue knew, in advance, that after looking at the red switch she would prefer that they take box A, and that after Lu looked at the green switch, Lu would prefer that they take box B. And Lu knew this as well. So they knew

that they would fight, and they knew which sides they would be taking in the fight. In that case, why not start fighting right then and there, before looking? Either way, fighting over which box to take at any stage in the problem seems perverse. The location of the money was wholly determined by the toss of a fair coin, so, intuitively, taking either box is permissible—regardless of what is known about the switches.

David Lewis (1981) once chided evidential decision theory for commending "an irrational policy of managing the news." To this we should add that the theory commends an irrational policy of managing its own requirements.

3.6 Conclusion

We began with a case in which one option—slacking off—was antecedently irrational, and we asked whether it was possible to "rationalize" that option. We found that it was not. It is worth noting that the case with which we ended takes a slightly different form. In *The Switch Problem*, both box-taking options are antecedently permissible, yet it is possible to render one or the other option uniquely rational.

Two questions arise. First, is the type of rationalization permitted under EDT in *The Switch Problem* any less toxic than the type of rationalization identified in §3.1? In other words, does the fact that the initial expected utilities of the options in *The Switch Problem* are equal make EDT's treatment of the case any less problematic? I see no reason to think that it does, though I think that this question is worth pursuing further.

Second, is it possible to design a case in which EDT permits the rationalization of an initially impermissible option? If it was, then the type of rationalization in such a case would seem exactly analogous to the studying problem with which we began.

Consider a variation on *The Switch Problem* in which there is a small prize—say, one cent—for looking at a switch. The addition of the prize tips the initial expected utilities in such a way that taking a box straightaway is now impermissible. So we have designed a case in which EDT permits the rationalization of an initially impermissible option. Still, the case is such that the initial expected utilities of the two box-taking options are equal—a feature not shared by the studying case. I leave it as an open question whether the two cases can be made exactly the same.

Bibliography

Adams, E. and Rosenkrantz, R. 1980. "Applying the Jeffrey decision model to rational betting and information acquisition." *Theory and Decision* 12:1–20.

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

—. 2014b. "Dicing with Death." *Analysis* 74:587–94.

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.

Bostrom, N. 2001. "The Meta-Newcomb Problem." *Analysis* 61:309–10.

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

—. 2015. "Costs of Abandoning the Sure-Thing Principle." *Canadian Journal of Philosophy* 45:827–40.

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

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

—. 2015. "Revisiting Risk and Rationality: A Reply to Pettigrew and Briggs." *Canadian Journal of Philosophy* 45:841–62.

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

Conee, E. and Feldman, R. 2004. *Evidentialism.* Oxford University Press.

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

Egan, A. 2007. "Some Counterexamples to Causal Decision Theory." *Philosophical Review* 116:94–114.

Feldman, F. 2006. "Actual Utility, the Objection from Impracticality, and the Move to Expected Utility." *Philosophical Studies* 129:49–79.

Fudenberg, D. and Tirole, J. 1991. *Game Theory.* MIT 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*, 125–62. Reidel.

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

Gustafsson, J. 2011. "A Note in Defense of Ratificationism." *Erkenntnis* 75:147–150.

Hájek, A. 2016. "Deliberation Welcomes Prediction." *Episteme* 507–28.

Hare, C. and Hedden, B. 2016. "Self-Reinforcing and Self-Frustrating Decisions." *Noûs* 50:604–28.

Harper, W. 1986. "Mixed Strategies and Ratifiability in Causal Decision Theory." *Erkenntnis* 24:25–36.

Hedden, B. 2015a. "Options and Diachronic Tragedy." *Philosophy and Phenomenological Research* 90:423–45.

—. 2015b. *Reasons without Persons.* Cambridge University Press.

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

—. 1983. *The Logic of Decision.* 2nd ed. University of Chicago Press.

Joyce, J. 1998. "A Nonpragmatic Vindication of Probablism." *Philosophy of Science* 65:575–603.

—. 1999. *The Foundations of Causal Decision Theory.* Cambridge University Press.

—. 2002. "Levi on Causal Decision Theory and the Possibility of Predicting One's Own Actions." *Philosophical Studies* 110:69–102.

—. 2007. "Are Newcomb Problems Really Decisions?" *Synthese* 156:537–62.

Kagan, S. MS. "The Paradox of Methods."

Kotzen, M. MS. "Three Principles of Inference and Deliberation."

Kyburg, H. 1980. "Acts and Conditional Probabilities." *Theory and Decision* 12:149–71.

Levi, I. 1975. "Newcomb's Many Problems." *Theory and Decision* 6:161–75.

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

—. 1997. *The Covenant of Reason: Rationality and the Commitments of Thought.* Cambridge University Press.

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

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

Maher, P. 1990. "Symptomatic Acts and the Value of Evidence in Causal Decision Theory." *Philosophy of Science* 57:479–98.

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

Myerson, R. 1991. *Game Theory: Analysis of Conflict.* Harvard University Press.

Nozick, R. 1969. "Newcomb's Problem and Two Principles of Choice." In N. Rescher (ed.), *Essays in Honor of Carl G. Hempel*, 114–46. Reidel.

Pettigrew, R. 2015. "Risk, Rationality, and Expected Utility Theory." *Canadian Journal of Philosophy* 45:798–826.

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

—. 2010. "A Resource-Bounded Agent Addresses the Newcomb Problem." *Synthese* 176:57–82.

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

Rabinowicz, W. 1988. "Ratifiability and Stability." In P. Gärdenfors and N. Sahlin (eds.), *Decision, Probability, and Utility*, 406–25. Cambridge University Press.

—. 2002. "Does Practical Deliberation Crowd Out Self-Prediction?" *Erkenntnis* 57:91–122.

Ramsey, F. 1990 [1926]. "Truth and Probability." In D. H. Mellor (ed.), *Philosophical Papers*. Cambridge University Press.

Savage, L. 1954. *The Foundations of Statistics*. Wiley Publications in Statistics.

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

—. 1984. *Pragmatics and Empiricism*. Yale University Press.

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

Sobel, J. H. 1994. *Taking Chances: Essays on Rational Choice*. Cambridge University Press.

Spencer, J. and Wells, I. Forthcoming. "Why Take Both Boxes?" *Philosophy and Phenomenological Research* .

—. MS. "The Metaethical Foundations of Decision Theory."

Spohn, W. 1977. "Where Luce and Krantz Do Really Generalize Savage's Decision Model." *Erkenntnis* 11:113–34.

—. 2012. "Reversing 30 Years of Discussion: Why Causal Decision Theorists Should One-Box." *Synthese* 187:95–122.

Stalnaker, R. 1981. "Letter to David Lewis." In Stalnaker R. Harper, W. L. and G. Pearce (eds.), *Ifs: Conditionals, Belief, Decision, Chance and Time*, 151–53. Reidel.

Weirich, P. 1988. "Hierarchical Maximization of Two Kinds of Expected Utility." *Philosophy of Science* 55:560–82.

—. 2004. *Realistic Decision Theory: Rules for Nonideal Agents in Nonideal Circumstances*. Oxford University Press.

Wells, I. Forthcoming. "Equal Opportunity and Newcomb's Problem." *Mind* .

Williamson, T. 2000. *Knowledge and its Limits*. Oxford University Press.