Binding and Its Consequences

Christopher J. G. Meacham

Published in *Philosophical Studies*, 149 (2010): 49-71.

Abstract

In "Bayesianism, Infinite Decisions, and Binding", Arntzenius, Elga and Hawthorne (2004) present cases in which agents who cannot bind themselves are driven by standard decision theory to choose sequences of actions with disastrous consequences. They defend standard decision theory by arguing that if a decision rule leads agents to disaster only when they cannot bind themselves, this should not be taken to be a mark against the decision rule. I show that this claim has surprising implications for a number of other debates in decision theory. I then assess the plausibility of this claim, and suggest that it should be rejected.

1 Introduction

In "Bayesianism, Infinite Decisions, and Binding," Arntzenius, Elga and Hawthorne (2004) examine the significance of *binding*—the ability to irrevocably commit oneself to some future plan of action. They show that in a number of cases, decision theoretic agents who can bind themselves will do much better than decision theoretic agents who cannot. Indeed, in some cases, agents who cannot bind themselves will be driven by decision theory to choose sequences of actions that have disastrous consequences, even when the agents know full well that these choices will lead to disaster, and know ahead of time that these are the choices they'll make.

One reaction to these cases is to take them at face value, and to conclude that these results are a mark against standard decision theory. If so, this gives us a *prima facie* reason to look more carefully at alternatives which don't have these consequences, such as the theories of Bratman (1987), Gauthier (1994) and McClennen (1990).

Arntzenius *et al* recommend a different reaction. They suggest that these cases do not give us a reason to be unhappy with standard decision theory. Rather, they argue, the source of these unhappy results is the inability of these agents to bind themselves:

> "The lesson is that under certain circumstances, the following ability can be incredibly helpful: ... the ability to irrevocably bind oneself to future courses of action. ... The lack of such ability is not, we say, a deficiency... It's just that certain situations exploit rational agents who are unable to self-bind."[1]

We can express this sentiment as follows:

The Binding Principle: If a theory of decision making has a counterintuitive result that only arises for agents who cannot bind themselves, this result is not a mark against the theory of decision making in question.[2]

If we adopt the Binding Principle, as Arntzenius *et al* suggest, then the cases they examine pose no threat to standard decision theory. Since it is only agents who cannot self-bind who are driven to disastrous outcomes, we can attribute these unhappy results to their inability to self-bind.

The Binding Principle is interesting for a number of reasons. As we've just seen, it allows standard decision theory to circumvent some otherwise troublesome charges. But it also has consequences for a number of other debates in decision theory. First, it alters the status of the "why ain'cha rich" argument for evidential decision theory. Second, it impacts our assessment of whether decision rules should be self-recommending. Third, it bears on whether decision instability poses a problem for causal decision theory.

Should we adopt the Binding Principle? I'll argue that we should not. I'll suggest that appeals to the Binding Principle mirror earlier appeals to a similar principle regarding mixed acts. And I'll argue that the Binding Principle is problematic for similar reasons.

This paper will proceed as follows. In next section I'll briefly sketch some background. In the third section I'll spell out the implications of adopting the Binding Principle on several debates in decision theory, including the "why ain'cha rich" argument, the question of whether rules should be self-recommending, and decision instability arguments. In the fourth section I'll assess the plausibility of the Binding Principle, and argue that it should be rejected. I conclude in the fifth section by briefly discussing the implications of these verdicts.

[1] Arntzenius, Elga and Hawthorne (2004), p.268–269.

[2] This principle is suggested by the discussion in Arntzenius, Elga and Hawthorne (2004), but not explicitly stated. The authors have confirmed this understanding of their position in correspondence.

Strictly speaking, this principle should include a caveat to bracket certain kinds of decision rules, such as those whose prescriptions explicitly depend on whether or not the agent can bind herself. For example, consider a decision rule that tells you to maximize expected utility if you can bind yourself, and to minimize expected utility if you cannot. Even though the counterintuitive results of the latter prescriptions only arise only for agents who cannot bind themselves, no one would want to claim that these results are not a mark against the decision rule. (Thanks to Adam Elga for this point.)

One way to introduce such a caveat is to restrict the scope of the Binding Principle to decision rules of the standard form—rules whose prescriptions are functions of the agent's current credences, utilities, and the set of available options. This restriction will rule out deviant decision rules of the kind just described, since whether or not one can bind oneself will not supervene on one's current credences, utilities, and the set of available options.

2 Background

2.1 Standard Decision Theory

As I'll understand it, (Bayesian) decision theory can be divided into two parts: a description of the agents to which the theory applies, and a normative claim about how such agents should behave.

The agents to which standard decision theory applies satisfy the following conditions:

A1. The agent's belief state at a time can be represented by a probability function over a space of possibilities. These values, called *credences* or *degrees of belief*, indicate the agent's confidence that the possibility is true, where greater values indicate greater confidence.

A2. The agent's evaluative state at a time can be represented by a function which assigns positive real numbers to elements in the space of possibilities. These assignments, called *utilities*, indicate the extent to which the agent values that possibility obtaining, where higher numbers indicate a higher utility.[3,4]

A3. The agent's potential acts in a decision situation can be represented by a unique set of mutually exclusive propositions $\{a_1, ..., a_n\}$ (where a_i can be thought of as the proposition that the agent performs the ith available act).

Now consider an agent of this kind who has credences cr and utilities u. The *expected utility* (EU) for the agent of an act a is:

$$EU(a) = \sum_{w \in \Omega} cr(w : a) \cdot u_w, \tag{1}$$

where Ω is the space of possibilities, and $cr(w : a)$ is a place holder.[5] By replacing $cr(w : a)$ with different kinds of functions, (1) yields different kinds of expected utility. If we set $cr(w : a)$ equal to the agent's credence in w conditional on a, then we get the *evidential expected utility* of the act. If we set $cr(w : a)$ equal to the agent's credence in w imaged on a, then we get the *causal expected utility* of the act.[6]

The normative part of standard decision theory claims that agents who satisfy A1-A3 ought to satisfy the following constraint:

[3] This understanding of utilities sets up decision theory as an account of prudential rationality. Alternatively, one can understand decision theory as an account of instrumental rationality, and take these utilities to be whatever is valuable according to the standard in question. In either case, decision theory, as understood here, is an account of what acts one ought to perform. It is not an account of how one ought to reason when making decisions, or of what preferences one ought to have (though given certain auxiliary assumptions, it may well bear on these issues).

[4] If we want to allow for well-defined infinite utilities, then we can use the extended reals to represent utilities instead of the reals.

[5] This characterization of (1) assumes a countable number of possibilities. To accommodate uncountably many possibilities, we can extend (1) in the usual way.

[6] I borrow this terminology from Collins (1996). For a discussion of some different ways of cashing out causal expected utility, see Joyce (1999).

Expected Utility Maximization: A condition-satisfying agent should only perform a potential act *a* if the expected utility of this act is at least as large as the expected utility of any alternatives. I.e., agents should perform acts which maximize expected utility.

By plugging different kinds of expected utility into this constraint, we get different kinds of decision theory. If we plug in evidential expected utility, this constraint yields *evidential decision theory.* If we plug in causal expected utility, we get *causal decision theory.*

Before we proceed, a slight revision of standard decision theory is required. Standard decision theory is usually understood to prescribe performing one of the acts with the highest expected utility, in the manner just described. But in some cases, there is no such act. For example, suppose an agent will be given n dollars, where n is a natural number of the agent's choosing. Assume the agent's utilities are linear in dollars. For any natural number n, the expected utility of choosing to get n dollars will increase as n does. Since there is no largest n, there is no act with the highest expected utility.

In cases of this kind we can't expect an agent to choose an act with the highest expected utility. There are a couple of ways to try to handle this: we might employ satisficing in cases where there's no highest expected utility act, or we might modify decision theory so that it no longer picks out "best" acts, but instead merely provides a "better than" ordering over them. For the purposes of this paper it will be more convenient to adopt the first approach, so that we can talk about what an agent ought to do, etc., in the usual way.

So in what follows, I'll assume that decision theory makes prescriptions in the following way. If there are acts available which maximize expected utility, then decision theory requires the agent to perform one of these acts. If there are no such acts, then decision theory permits the agent to perform any act within $1/10^n$ of a unit of expected utility of the lowest upper bound (if there is one), or with at least 10^n units of expected utility (if there isn't), for some large n.

2.2 Binding

Arntzenius *et al* speak of agents who can bind themselves to a future course of action. But how should we understand this? What does this ability to bind oneself consist of?

There are two natural ways to think about binding. First, one might think of agents who can bind themselves as agents who have extra "binding" acts available. Each of these "binding acts" corresponds to a different course of action one might pursue, and leads to the same outcome as that course of action would have.

Second, one might think of agents who can bind themselves as agents who can plan to pursue a course of action, and then invariably carry through with this plan. On this conception of binding, agents who can bind themselves are not agents with extra options, they're agents with extra willpower.

More abstractly, consider the possible sequences of decisions an agent might face. Since an agent's earlier decisions can alter the choices she faces later on, this will take the form of a branching tree of decision problems. On the first conception of binding, agents who can bind themselves are agents who are in a particular kind of decision tree. These are decision trees in which the agent always has the option of effectively 'skipping' to the end nodes of the tree—the end nodes that agents following the corresponding plan of action would have ended up at. On the second conception of binding, a binding agent isn't an agent in a particular kind of decision tree, it's an agent with a particular kind of mental fortitude or resolve.[7]

Although both of these notions of binding are interesting, it is the first of these notions that Arntzenius *et al* have in mind. Arntzenius *et al* state that decision theoretic agents who can bind themselves will do better than decision theoretic agents who cannot. But this statement is false given the second notion of binding. Given the second notion of binding, decision theoretic agents who can bind themselves will do the same as decision theoretic agents who cannot. Consider: If such agents plan to act in a way that accords with decision theory, and do as they plan, then they'll be led to the same outcomes as decision theoretic agents who cannot bind themselves. And if decision theoretic agents plan to act in a way that violates decision theory, they can't do as they plan, since they wouldn't be decision theoretic agents if they did.

Given the first notion of binding, on the other hand, it is true that decision theoretic agents who can bind themselves will do better than decision theoretic agents who cannot. So the first notion of binding is what Arntzenius *et al* have in mind. I gave a rough sketch of the first notion of binding above. Now let's spell it out more precisely.

Let a *decision problem* be an ordered triple consisting of an agent's current credences, utilities, and the set of available acts. Let a *comprehensive strategy* be a function which maps every decision problem to one of its available acts. Let a *decision tree* be an ordered triple consisting of an arborescence (a directed, rooted tree in which all arrows point away from the root), a function mapping each node in the tree to a decision problem, and a one-to-one function mapping each of the available acts in the decision problem associated with a node to arrows pointing away from that node.[8]

Let the *binding closure* of a decision tree be the tree that results from adding to each node a set of 'binding acts', one act for each comprehensive strategy. The outcomes of these binding acts are the same as the outcomes that would have come about if someone at that node had followed the corresponding comprehensive strategy. So the binding closure of a tree adds to each node acts which allow one to effectively 'skip' to the end nodes of the tree—the end nodes that agents following the corresponding comprehensive strategies would have ended up at. Call a tree that is the binding closure of some tree a *binding tree*.

To say that an agent can bind herself in the first sense is to say that the agent is in a binding tree. Such an agent always has the option of picking an arbitrary comprehensive

[7] Of course, one might hold that "willpower", "resolve", and the like aren't the right way to describe what binding agents (in the second sense) are like. (One might hold, for example, that the right way to model an agent with an iron will who decides to follow a given plan is to take them to be choosing an act which leads directly to the outcome that following this plan would lead to. If so, then agents with willpower should be understood as binding agents in the first sense, not the second.) The question of how to best understand this second notion of binding in an interesting one. But since this question is orthogonal to the issues we'll be concerned with, I'll put it aside.

[8] This characterization of decision trees builds in information about what the actual outcome of each act will be, regardless of whether that outcome is deterministic or indeterministic. This is a merely a matter of convenience; nothing of importance hangs on this choice.

strategy, and 'skipping' to the end node that this strategy would have led to.[9]

3 Implications of the Binding Principle

3.1 Predictable Disaster

Arntzenius *et al* present a number of cases where the ability to bind oneself seems desirable. Consider the Satan's Apple case:

> Satan cuts up an apple into countably infinite pieces in the following way: he cuts the apple in half, and then cuts the remaining half in half, and then cuts the remaining quarter in half, and so on. In a minute he will offer Eve the first piece, 30 seconds later he will offer her the second piece, 15 seconds after that he will offer her the third piece, and so on. So at the end of two minutes he'll have offered her every piece. Eve likes apples, and the utility she gets from eating a piece is equal to its size ($1/2$ for eating the first piece, 1/4 for eating the second piece, and so on). And Eve knows that she'll be expelled from the Garden of Eden—a consequence with -10 units of utility—*iff* she accepts an infinite number of pieces. What should Eve do?

Assume that Eve acts in accordance with causal or evidential decision theory, that she knows the set-up, and that she takes her decisions about whether to accept each piece to be causally or evidentially independent of her decisions regarding the other pieces.

What will Eve do with respect to the first piece? Eve believes that taking the first piece won't have any bearing on whether she accepts or declines any of the other pieces. If she's going to accept an infinite number of other pieces, then declining the first piece won't save her from getting expelled from Eden, and accepting the first piece will increase her overall utility by $1/2$. If she's only going to accept a finite number of other pieces, then accepting the first piece won't get her expelled from Eden, and eating it will increase her overall utility by 1/2. So at the end of the sequence of offers her overall utility will be greater by $1/2$ if she accepts the first piece no matter what else she does. So Eve will take it.

[9] Two qualifications. First, I said that the 'binding act' corresponding to a comprehensive strategy needs to lead to the same outcome as the corresponding comprehensive strategy. Given a fine-grained notion of outcomes, we can't require these outcomes to be *exactly* the same, since the fact that these outcomes were brought about by different sequences of choices is enough to distinguish them. So we can only require binding acts to lead to outcomes which are the same in the relevant respects to the outcomes of the corresponding strategy. (What counts as "relevant respects"? This is an interesting question. But since it's orthogonal to the issues I'll be concerned with, I'll put it aside.)

Second, the characterization just given only takes into account binding acts which skip to the end nodes of the tree. But one might want to consider ways of binding oneself which still leave some things open. For example, one might want to consider acts which effectively bind you to make certain choices if a particular situation comes up, but which otherwise leave your choices the same. We can extend the characterization of binding given above to include these possibilities. Let a *partial strategy* be a partial function from decision problems to acts. Then require the binding closure of a tree to also add to each node a set of 'partially binding acts', one act for each partial strategy. These 'partially binding acts' will lead to trees which look like the original tree, pruned to eliminate acts which conflict with the prescriptions of the corresponding partial strategy.

What about the second piece? The same reasoning applies with respect to the second piece, so Eve will accept the second piece as well. Likewise, she will accept every other piece she is offered. But since Eve accepts every piece, she'll get kicked out of the Garden of Eden, and her overall utility will be nine units lower than if she had declined every piece.

This looks like a troubling result for standard decision theory. Given standard decision theory, Eve is rationally required to make a series of choices which she knows will result in disaster: her eviction from the Garden of Eden.

But if we adopt the Binding Principle, as Arntzenius *et al* suggest, then we'll come to a different conclusion. Suppose that each time Satan offers Eve a piece of the apple, he also gives her the option of binding herself to some future course of action. Eve now has an infinite number of actions to choose from: in addition to just accepting or rejecting the piece, she can opt to bind herself to accept or reject some or all of the other pieces she will be offered in the future. And in this case, decision theory will allow her to bind herself to accepting some finite number of pieces, a course of action that leaves her well-fed and safely ensconsed in the Garden.[10]

So Eve will only be driven to disaster if she lacks the ability to bind herself. Given the Binding Principle, it follows that the fact that Eve is driven to disaster if she lacks the ability to bind herself is not a mark against standard decision theory. Rather, it just demonstrates how desirable the ability to bind oneself can be.

3.2 "Why Ain'cha Rich?"

In the standard Newcomb's case, you are presented with the choice of taking the contents of two boxes, or just the contents of the first box. A nearly perfect predictor has attempted to predict your choice. If she thinks you'll take just the first box, she'll put a million dollars in it. If she thinks you'll take both boxes, she'll leave the first box empty. The second box always contains a thousand dollars.[11]

According to the causal decision theorist, you should always take both boxes. That way, you will be a thousand dollars richer, no matter what the predictor has predicted. According to the evidential decision theorist, you should take only the first box. That's because the expected monetary reward for choosing one box is higher than that of choosing two boxes.[12]

If we expect the agents who employ one decision making theory to generally be richer than the agents who employ some other decision making theory, this seems to be a *prima facie* reason to favor the first theory over the second. Both causal and evidential decision theorists agree that, in the Newcomb's case, evidential decision theorists tend to end up wealthier than causal decision theorists. Both expect the evidential decision theorists to get a million dollars when she chooses the first box, and both expect causal decision theorists

[10] Note that this only makes it rationally permissible to make choices that would lead to disaster, not rationally obligatory. (To get the result that it's rationally obligatory, we need to impose some further constraints on Eve's credences.)

[11] See Nozick (1969).

[12] As usual, we're assuming that the agent's utilities are linear in dollars.

to get only a thousand when she chooses both boxes. So the Newcomb's case provides a *prima facie* reason to favor evidential over causal decision theory. As Gibbard and Harper put it, the causal decision theorist faces the question: "if you're so smart, why ain't you rich?"[13]

3.2.1 Response 1: Rewarding Irrationality

The standard response to the "why ain'cha rich?" argument is this:[14]

> In Newcomb's case, the predictor will reliably reward "one-boxing". So those who one-box will reliably end up better off than those who don't. But this doesn't show that one-boxing is rational. It merely shows that "if someone is very good at predicting behavior and rewards predicted irrationality richly, then irrationality will be richly rewarded."[15]

This response shows that the causal decision theorists can provide a consistent explanation for why they don't take the evidential decision theorist's wealth to be an indication of rationality. While the evidential decision theorist takes these rewards to be reason to think the pre-rewarded act is rational, the causal decision theorist takes the rewards to be merely a feature of the background situation that is irrelevant to the rationality of the act.

But as a response to the "why ain'cha rich?" argument, this isn't very satisfying. First, this response doesn't show very much. After all, it's not surprising that causal decision theory will judge the acts it prescribes to be rational, and those it doesn't prescribe irrational. This just demonstrates that causal decision theory is consistent. Second, this response won't cut any ice with the evidential decision theorist, who will maintain that irrational acts *can't* be predictably pre-rewarded—if the act is predictably pre-rewarded, then it will be the rational act.[16] A more satisfying response to the "why ain'cha rich?" argument would do more than just show that the causal decision theorist's position is consistent. It would also undermine the evidential decision theorist's claim that these considerations give us a reason to favor evidential decision theory over causal decision theory.

This is where the second line of response to the "why ain'cha rich?" argument comes in. These responses try to do more than just show that causal decision theory is consistent—they also try to undermine the intuition that these considerations provide a *prima facie* reason to favor evidential decision theory over causal decision theory. Let's look at two such "second-line" responses.

3.2.2 Response 2: Gibbard and Harper

One response, offered by Gibbard and Harper (1985), attempts to show that the "why ain'cha rich?" arguments can be used against both causal and evidential decision theory in order to support apparently crazy theories of decision making. If so, then the evidential

[13] Gibbard and Harper (1985), p.153.

[14] See Gibbard and Harper (1985), Lewis (1981) and Joyce (1999).

[15] Gibbard and Harper (1985).

[16] For more discussion of the evidential decision theorist's stance on this argument, see Lewis (1981).

decision theorist should also doubt that we should take "why ain'cha rich?" considerations into account when evaluating theories of decision making. Inessential details aside, the case they consider is this:

> As in the standard Newcomb's case, suppose you must decide between taking one or two boxes, where a predictor has placed a million dollars in the first box iff he predicts you will take one box, and where the second box always has a thousand dollars. In this case, you'll only be allowed to take the boxes if your decision-making dispositions satisfy certain conditions; but both evidential and causal decision theorists satisfy these conditions.[17] Now suppose that both of the boxes are transparent, so you will see the contents of both boxes before you make your choice. What should you do?

In this case, both causal and evidential decision theory will tell you to take both boxes. And, since the predictor is nearly infallible, both evidential and causal decision theorists will tend to end up with a thousand dollars. By contrast, agents who employ a decision making theory according to which you should only take the first box no matter what you see, will tend to end up with a million dollars. So evidential decision theory seems to be as vulnerable to "why ain'cha rich?" arguments as causal decision theory.

But note that if we adopt the Binding Principle, this response is no longer compelling. An evidential decision theorist who has the ability to self-bind will bind herself to choosing one box before she's shown what's in them. Since the predictor will predict this, she'll put a million into the first box, and the binding evidential decision theorist will end up rich. Since it's only non-binding evidential decision theorists who will end up poor in this case, it follows from the Binding Principle that we shouldn't take this to be a mark against evidential decision theory.

But self-binding causal decision theorists can still end up poor. Consider a version of the Newcomb's case where the predictor makes her prediction before the agent is born. The binding causal decision theorist will be unable to causally influence the prediction, and so she will end up choosing both boxes and getting only a thousand dollars. So even when we restrict our attention to agents who can bind themselves, the "why ain'cha rich" argument against causal decision theory remains.[18]

[17] You're not allowed to take boxes if you have a disposition which would make correct prediction impossible. For example, you're not allowed to take the boxes you choose if your decision making dispositions are: "Take the first box if I see there's nothing in it, and take both boxes if I see there's a million in it." (If your dispositions are such that the predictor can effectively choose which decision you make—you'll take two boxes if you see nothing in the first box, and just the first box if you see the million—we can assume the predictor is stingy, and won't put anything in the first box.)

[18] Suppose we modify the case so that contents of the second box are encoded in the agent's initial credences. Then the binding evidential decision theorist will choose both boxes, and will end up poor, just like the binding causal decision theorist. So doesn't the "why ain'cha rich" argument against evidential decision theory remain as well?

No. The force of the "why ain'cha rich" argument against a decision rule X comes from the fact that cognitively ideal X-decision theorists can expect ahead of time that they will generally end up richer if they choose act a instead of act b, and yet once they're in that decision problem, they'll choose b anyway. In the Newcomb's variant with two transparent boxes, for example, the evidential decision theorist expects dedicated one-boxers to end up

So if we adopt the Binding Principle, the Gibbard and Harper response is ineffective. The case they discuss is not a problem for the evidential decision theorist, but Newcomb's case is still a problem for the causal decision theorist.

3.2.3 Response 3: Arntzenius

Another "second-line" response, offered by Arntzenius (2008), attempts to show that there are cases where we'll expect causal decision theorists to end up richer than evidential decision theorists. If there are such cases, the evidential decision theorist can no longer maintain that evidential decision theorists are generally richer than causal decision theorists, and the "why ain'cha rich" argument for evidential decision theory collapses.

It's not easy to construct such cases. If, for example, we just stipulate that the "evidential decision theory" choice will be punished in some way, then it will no longer be the choice that evidential decision theory recommends (see Lewis (1981)). However, Arntzenius (2008) shows that there are cases in which causal decision theorists will do better than evidential decision theorists. Consider the following case:

> Suppose there will be a 10 game series between the Yankees and the Red Sox. You know that the Yankees have a 90% chance of winning any given game. You'll only be allowed to bet on each game if your decision-making dispositions satisfy certain conditions; both evidential and causal decision theorists satisfy these conditions.[19] If you're allowed to bet, then you can bet on either the Yankees (in which case you earn a dollar if the Yankees win, and lose two dollars if the Red Sox win), or the Red Sox (in which case you lose one dollar if the Yankees win, and earn two dollars if the Red Sox win). Finally, before you place each bet, an infallible predictor will tell you whether you'll win or lose the bet. How should you bet?

richer than two-boxers, but she'll choose both boxes anyway.

This is not what happens in the case just described. If the contents of both boxes are encoded in her initial credences, it's never the case that the evidential decision theorist expects one-boxing to make her rich: she always expects the one-boxer to get nothing and the two-boxer to get a thousand dollars. (And if her credences are accurate, she's right.) So the "why ain'cha rich" argument doesn't apply.

One might try instead to set up an objective version of the "why ain'cha rich" argument against the binding evidential decision theorist using this case. One might stipulate that the predictions in question are made using a chance process that has a 99.9% chance of success, and point out that the expected gain of the binding evidential decision theorist, calculated using the objective chances ("expected$_{chance}$"), is lower than that of a dedicated one-boxer.

But, again, this argument won't work. If the binding evidential decision theorist doesn't know what the chances are, then this argument is merely taking advantage of her ignorance. If the binding evidential decision theorist does know what the chances are, then the initial credences she's been stipulated to have will violate something like the Principal Principle: her credences won't line up with what she thinks the chances are. And it's no surprise that an agent whose credences don't line up with the chances can be expected$_{chance}$ to do poorly.

[19] You're not allowed to bet if you have a disposition which would otherwise make the set-up of the case impossible. For example, you're not allowed to bet if your betting dispositions are: "Bet on the Red Sox if I'm told I'll win my bet, and bet on the Yankees if I'm told I'll lose my bet." Since the predictor can't consistently tell you that you'll win or lose your bet, these dispositions make the set-up of the case impossible.

If no predictor were involved, both causal and evidential decision theory would tell you to bet on the Yankees every time, since the expected utility of betting on the Yankees ($0.9 \cdot 1 + 0.1 \cdot -2 = 0.7$) would be greater than the expected utility of betting on the Red Sox ($0.9 \cdot -1 + 0.1 \cdot 2 = -0.7$). If you bet on the Yankees every time, we'll expect you to win nine times, to lose once, and to be up by $7 by the end of the series.

How should your betting behavior change in light of what the predictor tells you? If you're a causal decision theorist, your behavior won't change at all. The predictor doesn't tell you anything causally relevant to the outcome of the game, so you'll effectively ignore what she says.

If you're an evidential decision theorist, on the other hand, your betting behavior will change. Suppose the predictor tells you that you'll win. Since you get $2 if you win betting on the Red Sox and only $1 if you win betting on the Yankees, you'll bet on the Red Sox. If she tells you that you'll lose, you'll also bet on the Red Sox, since you lose $2 if you lose betting on the Yankees and only $1 if you lose betting on the Red Sox. So you'll bet on the Red Sox no matter what the predictor tells you. And by doing so, we expect you to win once, lose nine times, and be down by $7 by the end of the series.

In this case we expect the causal decision theorists to do better than the evidential decision theorists: we'll expect the causal decision theorists to end up $7 ahead, and we expect the evidential decision theorists to end up down by $7. So the evidential decision theorist can not maintain that evidential decision theorists are generally better off: they're better off in some situations, but worse off in others. This deflates the "why ain'cha rich?" argument for adopting evidential decision theory instead of causal decision theory.

But note that if we adopt the Binding Principle, matters are different. Suppose the evidential decision theorist has the ability to bind herself. Then in the Red Sox and Yankees case she'll bind herself to betting on the Yankees before the predictor informs her of the outcome of the game. So the binding evidential decision theorist will bet on the Yankees every time, and will end up as rich as the causal decision theorist. Since it's only non-binding evidential decision theorists who will end up poor in the Red Sox and Yankees case, it follows from the Binding Principle that we shouldn't take this to be a mark against evidential decision theory. The argument against the causal decision theorist, however, remains. So if we adopt the Binding Principle, Arntzenius' response to the "why ain'cha rich" argument won't work.

3.2.4 Assessing the "Why Ain'cha Rich?" Argument

Without the Binding Principle, the second line of response to the "why ain'cha rich" argument succeeds. The cases that Gibbard and Harper (1985) and Arntzenius (2008) present undermine the evidential decision theorist's claim that "why ain'cha rich" considerations favor evidential decision theory over causal decision theory. One might conclude from this that we should ignore "why ain'cha rich" considerations when assessing decision theories. Alternatively, one might conclude that we should be unhappy with both evidential and causal decision theory. But either way, "why ain'cha rich" considerations will fail to support evidential decision theory over causal decision theory.

The final analysis looks different, however, if we adopt the Binding Principle. With the Binding Principle, the second line of response to the "why ain'cha rich" argument fails. While evidential decision theorists who can't bind themselves may end up losing out in the cases Gibbard and Harper (1985) and Arntzenius (2008) present, this isn't a reason to worry about evidential decision theory. Rather, this just demonstrates the unhappy position of agents who are unable to bind themselves. Causal decision theorists can still consistently deny that the evidential decision theorist's position is rational, of course. But they are unable to diffuse the *prima facie* intuition that "why ain'cha rich" considerations are relevant; an intuition which, given the Binding Principle, tells in favor of evidential decision theory.[20]

3.3 Self-Recommendation

Skyrms (1982) raises the question of when decision rules are *self-recommending*:

> "The question of what decision method to use for a sequence of decision problems is itself a decision problem. If the rules of rational decision are formulated generally enough, they can be applied to such problems. Let us call a sequence of decision problems a *world*, and the problem of which decision theory to adopt for the individual problems in the sequence, the *world decision problem*. For a given world decision problem, a decision rule might recommend adopting a conflicting rule for dealing with the problems of that world. On the other hand, for certain worlds, certain decision rules will be *self-recommending*."[21]

We can spell out what it is for a rule to be self-recommending in our terms as follows. Recall that a *comprehensive strategy* is a function from decision problems to acts. Let a *perspective* be an ordered pair consisting of a credence and utility function. Given a decision rule ("X-decision theory") we can determine two things. First, we can determine which comprehensive strategies correspond to the choices an X-decision theorist might actually make. Second, we can determine which comprehensive strategies X-decision theory takes to be best from a given perspective. So given evidential decision theory, for example, we can work out which comprehensive strategies describe how an evidential decision theorist might act, and we can work out which comprehensive strategies evidential decision theory takes to be the best from a given perspective—the comprehensive strategies which

[20] It's worth clearing up a potential confusion regarding the role of the Binding Principle. The Binding Principle is being applied here to evaluate whether various *prima facie* counterintuitive results of evidential and causal decision theory—that in certain cases agents who follow their prescriptions will end up poor, even though these agents correctly expect subjects who act in a different manner to end up rich—should be taken as marks against these theories. The Binding Principle is not being applied to the "why ain'cha rich" argument itself, in order to (say) evaluate the merits of this argument. Such an application wouldn't make sense. The Binding Principle only applies when we're evaluating consequences or features of a particular decision rule, and it only makes claims about whether these consequences or features should bear on our evaluation of that rule. It doesn't apply to arguments or considerations independently of a given decision rule, and it doesn't make claims about their general merits.

(Similar remarks apply to the discussions of self-recommendation and decision instability that follow. Thanks to Ted Sider for pointing out this potential confusion.)

[21] Skyrms (1982), p.707.

get assigned the highest evidential expected utility. A decision rule is *self-recommending* from a perspective when the comprehensive strategies that describe the behavior of a rule-following agent are among the strategies that the rule considers best. So a decision rule is self-recommending from a perspective when it takes itself to be a good decision rule to adopt.

Being self-recommending from a wide variety of perspectives is a nice feature for a decision rule to have. A decision making rule that is self-recommending is robustly confident about the acts it prescribes.

That said, we shouldn't expect a decision rule to be self-recommending from every perspective. For instance, we shouldn't expect a decision rule to be self-recommending from the perspective of an agent who believes her future credences will fail to cohere with her current ones in certain ways. Consider an agent who knows that a given coin toss landed heads, and knows that she will forget this information by the time she's in a position to bet on it. Given this, evidential decision theory may assign a higher evidential expected utility to comprehensive strategies that describe an agent who always bets on heads than to strategies that describe how an evidential decision theorist would act. But this doesn't indicate that evidential decision theory lacks confidence in its prescriptions. Rather, evidential decision theory doesn't endorse the comprehensive strategies that describe evidential decision theorists because after a certain point the evidential decision theorists will be making choices with faulty credences.

Likewise, we shouldn't expect a decision rule to be self-recommending from the perspective of an agent who believes her future utilities will differ from her current ones. Consider, for example, an agent who initially only values the happiness of sentient beings, but believes she will come to value only money at some point in the future. Given this, evidential decision theory will assign a higher evidential expected utility to comprehensive strategies which describe an agent who spends her life promoting happiness than to comprehensive strategies that describe the evidential decision theorist, who will soon turn to collecting money at the expense of others. But again, this doesn't indicate that evidential decision theory fails to be confident in its own prescriptions. Rather, evidential decision theory doesn't endorse the comprehensive strategies that describe the evidential decision theorist because after a certain point the evidential decision theorist will be making choices with different utilities.

Finally, we shouldn't expect a decision rule to be self-recommending from the perspective of an agent who thinks she'll come to believe (rightly or wrongly) that she may deviate from the prescriptions of the rule. Consider an alcoholic who enjoys the atmosphere at the local bar, but who believes that she will give into temptation and start drinking if she goes there. According to evidential decision theory, the alcoholic should not go to the bar. But evidential decision theory assigns a higher evidential expected utility to comprehensive strategies which prescribe going to the bar and not drinking than to strategies which prescribe staying home. Again, this doesn't indicate that evidential decision theory lacks confidence in its prescriptions. Rather, these two ways of picking out comprehensive strategies diverge because the first way factors in the possibility of failing to adhere to a strategy, while the second way—assessing how good adhering to a strategy would be—does not.

To avoid these kinds of cases, let's restrict our attention to the perspectives of agents who believe that they will (a) update by conditionalization, (b) have static utilities, and (c) have a negligible credence that they'll deviate from the decision rule. Given these restrictions, are evidential and causal decision theory self-recommending?

No. To see that evidential decision theory can fail to be self-recommending, recall the Gibbard and Harper (1985) variant of Newcomb's case described in the previous section, where you see the contents of both boxes before you make your choice. In this case evidential decision theory will recommend that you take both boxes, regardless of what you see. And since the predictor will predict this, evidential decision theorists will tend to only get a thousand dollars. Now consider the decision rule X-decision theory, which prescribes taking the first box no matter what you see. Since the predictor will predict this, X-decision theorists will tend to get a million dollars. So the evidential expected utility of acting in accordance with X-decision theory will be higher than the evidential expected utility of acting in accordance with evidential decision theory.

To see that causal decision theory can fail to be self-recommending, consider the Satan's Apple case. The causal expected utility of acting like a causal decision theorist will be low—she'll be expelled from Eden. On the other hand, the causal expected utility of acting in accordance with a decision rule that results in her taking only the first 100 pieces will be much higher—she'll remain in Eden and get most of the apple.

So neither evidential nor causal decision theory are self-recommending in all of the cases we'd like. This is a *prima facie* reason to look for some other decision rule.

But if we adopt the Binding Principle, we'll come to a different assessment. If we further restrict our attention to agents who can bind themselves, evidential and causal decision theory will both be self-recommending. An evidential decision theorist who can self-bind in the Gibbard and Harper case will bind herself to choose only the first box before she sees their contents. And the evidential expected utility of these comprehensive strategies will be as high as the evidential expected utility of the comprehensive strategies prescribed by any other decision rule. Likewise, a causal decision theorist who can self-bind in the Satan's Apple case will bind herself to accepting only a finite number of pieces, and the causal expected utility of this kind of comprehensive strategy will be as high as the causal expected utility of the comprehensive strategies prescribed by any other decision rule. So given the Binding Principle, both evidential and causal decision theory are appropriately self-recommending.

3.4 Decision Instability

One of the worries that has been raised for causal decision theory is that it leads to cases of *decision instability*. We have a case of decision instability if, for every available act a, the expected utility of a conditional on a is lower than the expected utility of some other act b conditional on a.[22] In such cases there's a sense in which you'll be displeased with your choice no matter what, since as soon as you choose an act you'll come to believe that

[22] These are sometimes called cases of "pure" decision instability, with "impure" cases being ones in which the above condition only holds for some of the available acts (c.f. Richter (1986)).

some other act is better.[23]

A classic example of decision instability is the Death in Damascus case presented by Gibbard and Harper (1985):

> "Suppose the man knows the following. Death works from an appointment book which states time and place; a person dies if, and only if, the book correctly states in what city he will be at the stated time. The book is made up weeks in advance on the basis of a highly reliable prediction. An appointment on the next day has been inscribed for him. Suppose, on this basis, the man would take his being in Damascus the next day as strong evidence that his appointment with Death is in Damascus, and would take his being in Aleppo the next day as strong evidence that his appointment is in Aleppo."[24]
>
> If the man's only choices are to go to Aleppo or to go to Damascus, what should he do?

According to causal decision theory, the man's decision is unstable. Conditional on the assumption that he'll go to Aleppo, the causal expected utility of going to Damascus will be higher, since he'll expect Death to be at Aleppo. Conditional on the assumption that he'll go to Damascus, the causal expected utility of going to Aleppo will be higher, since he'll expect Death to be at Damascus. So no matter what the man decides to do, as soon as he makes his choice he'll come to believe that the other act is better.

Given evidential decision theory, on the other hand, decision instability cannot arise.[25] The evidential expected utility of going to Damascus conditional on going to Aleppo is not well-defined, so the means of comparing the two acts that decision instability requires is unavailable.[26]

Is this kind of case a problem for causal decision theory? A number of people have thought so, and have proposed revisions of causal decision theory in order to avoid this kind of instability.[27] That said, it's not clear that the decision instability that appears in cases like Death in Damascus is sufficient to justify these revisions. First, as Gibbard and Harper note, it's not clear that this result is counterintuitive. Arguably, decision instability is exactly what we should expect in this kind of case:

> "Any reason the doomed man has for thinking he will go to Aleppo is a reason for thinking he would live longer if he stayed in Damascus, and any reason he has for thinking he will stay in Damascus is a reason for thinking he would live longer if he stayed in Aleppo. Thinking he will do one is a reason for

[23] Assuming that you update by conditionalization, and that the evidence you get from performing an act is just that you've performed the act.

[24] Gibbard and Harper (1985), p.154–155.

[25] Though there are variants of canonical evidential decision theory, such as the "ratificationism" of Jeffrey (1983) which are also subject to decision instability. Our conclusions regarding decision instability and causal decision theory apply *mutatis mutandis* to these variants.

[26] We might adopt primitive conditional probabilities to get around this obstacle. But it still seems unlikely that decision instability will arise for evidential decision theory, for reasons given by Gibbard and Harper (1985).

[27] For example, see Sobel (1983), Eells (1985), and Weirich (1985).

> doing the other. That there can be cases of unstable [causal expected utility]-maximization seems strange, but the strangeness lies in the cases, not in [causal expected utility]-maximization: instability of rational decision seems to be a genuine feature of such cases."[28]

Second, the decision instability in this case doesn't result in any strange behavior: the man will still go to either Aleppo or Damascus. So the fact that his decision is unstable doesn't seem particularly troubling.[29]

But more troubling worries for causal decision theory arise when we consider multiple-act versions of the cases where decision instability arises. For example, consider the following variant of the Death in Damascus case, suggested by Richter (1986):

> Add to the Death and Damascus case the following details. The man is halfway between Damascus and Aleppo. He must decide whether to take a step toward Damascus, or a step toward Aleppo. After he's taken his first step, he has to decide whether to take his second step toward Damascus or Aleppo, and so on. It will take him an hour to reach either city, and there are 5 hours left before nightfall. He knows that if he is not in a city by nightfall, Death's jackal companions will come and eat him, and this will be a much more unpleasant way to die than meeting Death in the city. Finally, he's thirsty, tired and hungry, and he would like to get to a city as soon as possible. What should the man do?

If the man acts in accordance with causal decision theory, he will act as follows. He will start going toward one city until his credence that Death will be there starts to increase. Then he'll start taking steps back toward the other city until his credence shifts again. And so on.[30] He'll continue to dither about unhappily in the desert until he has no more time

[28] Gibbard and Harper (1985), p.156.

[29] Egan (2007) presents some other cases in which decision instability arises, and argues that in these cases causal decision theory delivers the wrong verdicts. Arntzenius (2008) suggests that the proponent of causal decision theory can reasonably deny that the verdicts in question are counterintuitive. But in any case, Egan's concern is not with decision instability *per se*, but with the fact that he thinks causal decision theory's prescriptions are counterintuitive.

Weirich (1985) and Richter (1986) raise a different kind of worry: they argue that in the kinds of cases in which decision instability arises for causal decision theory, it should generally be rationally permissible to know what you're going to do before you do it. But one cannot both satisfy causal expected utility-maximization in cases of decision instability and know what you're going to do before you do it. So if we grant that knowing what you're going to do before you do it is rationally permissible in cases like the Death in Damascus case, causal decision theory is in trouble.

[30] One might worry about whether his credence *should* increase. After all, there's no reason to think he can't predict how he's likely to behave ahead of time. And why should his credence that he will end up at Aleppo increase as he takes more steps toward Aleppo if he knows he's going to change his mind and turn around? Of course, similar reasoning can be applied if we claim that his credence that he'll end up at Aleppo should remain the same. If his credence in ending up at Aleppo will remain the same as he takes steps toward Aleppo, then we'd expect him to end up at Aleppo, in which case it seems his credence that he'd end up at Aleppo should have been increasing after all.

In any case, we can side-step the issue by stipulating that, after every step, the man has a chance of involuntarily left to spare if he's to make it to one of the cities before nightfall. At that point it will become a choice between going to the closer city or be eaten by jackals, and so he'll head toward the nearest city.

This seems like an unhappy result for causal decision theory. If he adheres to causal decision theory, the man can know ahead of time that he'll spend 4 extra hours dithering about unhappily in the desert instead of relaxing and sipping margaritas at a local tavern, and yet he'll do it anyway.

But if we adopt the Binding Principle, we should view these cases in a different light. If the man has the option of binding himself, then he won't dither about in the desert—he'll choose one of the two cities, and bind himself to taking steps in that direction the rest of the way. So agents who can bind themselves won't suffer from the kind of problem posed in the Dithering in Damascus case. Thus, given the Binding Principle, we shouldn't take these cases to indicate problems with causal decision theory. Rather, we should take these cases to be further demonstrations of the fact that agents who are unable to bind themselves can end up in unhappy situations.

4 Assessing The Binding Principle

Adopting the Binding Principle is an appealing way of maintaining the *status quo* with respect to standard decision theory. But there are reasons to doubt it. To make these doubts salient, let's first look at a similar issue that arises in discussions about decision instability.

4.1 Decision Instability and Mixed Acts

Causal decision theory has been criticized for giving rise to decision instability. One response to this worry has been to note that the instability tends to go away if we allow for *mixed acts*.[31] A mixed act is typically thought of as a decision to base one's act on the result of some chance event. For example, one might flip a fair coin in order to decide between performing act *A* or act *B*. More precisely, call the acts typically considered in decision theory—pull a lever, accept a bet, etc.—*pure acts*. We can characterize a mixed act as a probability distribution over these pure acts, where these probabilities are independent of the possible outcomes of the pure acts in question. In the example just given, the mixed act assigns a probability of 50% to *A* and a probability of 50% to *B*.

Responses to decision instability that employ mixed acts have generally been met with skepticism.[32] One such response is to claim that agents like us do have access to mixed acts, and thus won't face cases of decision instability. But this claim is implausible: we don't generally have access to the probability-generating devices required, and it's hard to

taking another step in a random direction (and after that random step, a chance of taking yet another step in a random direction, and so on). With this addition, the man's credence that he'll end up at Aleppo *should* increase after he takes a step toward Aleppo, based solely on these chances.

[31] See Harper (1986).

[32] See Arntzenius (2008) for criticisms of this kind.

see how we could perform mixed acts without them. Another response is to restrict the application of decision theory to agents who have access to mixed acts, and *a fortiriori* restrict decision theory to cases where decision instability won't arise. But this robs standard decision theory of much of its interest. We want to know how agents like ourselves should act, and if standard decision theory doesn't provide this guidance, we'll have to look elsewhere.

More generally, it seems like shifting to cases in which the agents have access to mixed acts is changing the topic. After all, the mixed acts version of a decision problem is a different decision problem than the original. In the original Death in Damascus case, for example, the agent has only two choices: go to Damascus or go to Aleppo. But if we take the agent to also have mixed acts available then the agent has an infinite number of choices: the two pure acts of the original case, and the mixed acts consisting of every probability distribution over those pure acts.

Moreover, the addition of these choices substantially alters the nature of the case. Not only are we assuming that the agent has an adjustable chancy device and the willpower to commit herself to certain acts given certain outcomes, we're also assuming that the outcomes of the chance device are independent of the outcomes of the pure acts in question. This is a substantial assumption. In order to maintain this independence in the Death and Damascus case, we have to assume that Death, astounding predictor that she is, isn't able to predict the outcomes of the agent's chance device. This makes it clear that the agent in the mixed acts version of the case is in a very different situation from the agent in the original case. In the original case, the agent has virtually no chance of staying alive, since Death is a fantastically good predictor. In the mixed acts variant the agent is in a much better position: since Death can't predict the outcome of her chance device, if she uses it to randomly determine which city she'll go to, her chance of surviving will be 50%.

A different way of using mixed acts to ward off decision instability problems is this:

The Mixed Acts Principle: If a theory of decision making has counterintuitive results that only arise for agents without access to mixed acts, these results are not a reason to reject the theory of decision making in question.

This response grants that agents like us might face cases of decision instability on causal decision theory. But since it's only agents who don't have access to mixed acts who will face such cases, these cases should not be taken to be marks against causal decision theory. Rather, the moral is that having access to mixed acts is desirable. Agents without such access might run into uncomfortable situations, like cases of decision instability. But this doesn't indicate a flaw in the decision rule in question; rather, it's just a demonstration of why it's nice to have access to mixed acts.

It's true that having access to mixed acts is desirable. In the Death in Damascus case, for example, the agent who has access to mixed acts has a much better chance of staying alive. But this isn't enough to justify the Mixed Acts Principle. Being omniscient is desirable too, but that doesn't mean we can ignore any counterintuitive consequences a decision rule has for agents who are not.

I take the Mixed Acts Principle to be implausible. The mixed acts version of a case is simply a different case from the original. If a decision rule delivers a counterintuitive result in the original case, then that seems to be a mark against the theory. And whether this consequence also appears when we consider the mixed acts version of the case is irrelevant to our evaluation of the original case.

The Mixed Acts Principle may derive some apparent plausibility from a related claim:

Ought Implies Can (Mixed Acts): If no decision rule can avoid a given counterintuitive results without employing mixed acts, then it's not a demerit of a particular decision rule that it can't avoid these results without employing mixed acts.

This claim is plausible. But unlike the Mixed Acts Principle, it does not support the mixed acts response to the decision instability worries since there are decision rules which won't lead to decision instability, even when mixed acts are not available. As we've seen, evidential decision theory is one such rule, and there are many others. So this claim doesn't allow the causal decision theorist to ignore cases of decision instability that arise for agents who don't have access to mixed acts.

Causal decision theory might still be the most plausible decision rule, of course. One's assessment of a theory depends on lots of factors, of which this is only one. But none of that changes the fact that decision instability seems to be a demerit of the theory. And when we assess the pros and cons of the theory, it should be taken as such.

4.2 The Binding Principle

We've seen some reasons to be skeptical of the Mixed Acts Principle. These same considerations apply to the Binding Principle.

The binding version of a decision problem, like the mixed acts version, is a different decision problem than the original: the agent is presented with a number of additional acts to choose from. And like the Mixed Acts Principle, the Binding Principle is *prima facie* implausible. If a decision rule delivers a counterintuitive result in a given case, then that seems to be a mark against the theory. And whether this counterintuitive result also appears when we consider some other case—a case where the agent has binding acts available, say—is irrelevant to our evaluation of this case.

To put it another way, if we accuse standard decision theory of allowing agents who can't bind themselves to choose acts which lead to disaster, the proponent of the Binding Principle will respond: "Well, if they had access to self-binding they'd be alright." But this is just to say: "Well, if they were in some other situation instead of this one then they'd be alright." This is true, of course, but how does it make the fact that agents are led to disaster in this situation any less problematic?

There is a claim that resembles the Binding Principle that may lend it plausibility:

Ought Implies Can (Binding): If no decision rule can avoid a given counterintuitive result without employing binding, then it's not a demerit of a particular decision rule that it can't avoid these results without employing binding.

This claim is plausible. But unlike the Binding Principle, it does not lend itself to the defense of standard decision theory, since there *are* decision rules which won't lead agents who can't bind themselves to disaster.

For instance, we can construct a rule which effectively advises an agent to choose the acts that she would have wanted to bind herself to. Let $iu(\cdot)$ represent the agents initial utilities, let $ic(\cdot)$ represent the agents initial credences, and let CS be the proposition that the agent will act in accordance with a particular comprehensive strategy. Let the "cohesive expected utility" of a comprehensive strategy be:

$$CoEU(CS) = \sum_i ic(w_i : CS) \cdot iu(w_i). \tag{2}$$

As before, we can get causal and evidential versions of this theory by substituting the appropriate replacement in for $ic(w_i : CS)$. Now consider a decision rule which requires agents to satisfy the following constraint:

Cohesive Expected Utility Maximization: A condition-satisfying agent should perform the act picked out by a comprehensive strategy which maximizes cohesive expected utility.

This "cohesive decision theory" prescribes the acts that the agent would have bound herself to had she the ability to do so. And non-binding agents who adopt cohesive decision theory will not be driven to disaster in the kinds of cases Arntzenius *et al* describe.

Cohesive decision theory is one rule which avoids the counterintuitive consequences that Arntzenius *et al* describe, but there are many others. Similar rules have been proposed by Bratman (1987), Gauthier (1994) and McClennen (1990). And other rules with similar features are not hard to find.[33]

Indeed, we can even find rules of this kind which have the standard form—rules whose prescriptions are functions of the agent's current credences, current utilities and the available acts. For example, starting with (2), we can replace the agent's initial utilities with her current ones, and replace the agent's initial credences with the expectation of what the agent currently believes her initial credences might have been. Call the resulting expression the *cohesive expected utility*$_2$ of a comprehensive strategy:

$$CoEU_2(CS) = \sum_i cr(ic_i) \sum_j ic_i(w_j : CS) \cdot u(w_j). \tag{3}$$

We can plug cohesive expected utility$_2$ into the constraint described above to get *cohesive decision theory*$_2$. And, as before, non-binding agents who adopt cohesive decision theory$_2$ will not be driven to disaster in the kinds of cases Arntzenius *et al* describe.[34]

[33] Note that some of these rules will yield stronger prescriptions in the Satan's Apple case than others. The cohesive decision theories described here, for example, yield the result that it is permissible to stop taking pieces at a some point. So these rules do not require agents to perform acts which will lead to disaster. Rules like those of Bratman (1987), Gauthier (1994) and McClennen (1990) yield the stronger result that it is *obligatory* to stop taking pieces at some point, at least in situations in which the agent has planned or committed themself to stopping at that point ahead of time. So (under certain conditions) these rules will *forbid* agents from performing acts which will lead them to disaster.

[34] I'm using these cohesive decision theories as counterexamples to the claim that every decision rule will lead agents who can't bind themselves to disaster. But these theories are of independent interest. As such, it's worth noting a few things about them.

[1] Cohesive decision theories don't require agents to have willpower, plans or foresight. Nor do they require agents to have committed themselves to future courses of action at some earlier time. In this respect, these cohesive decision theories differ from the proposals offered by Bratman (1987), Gauthier (1994) and McClennen (1990). Like standard decision theory, cohesive decision theories are simply rules which prescribe acts to agents in decision problems. They require no more of agents than standard decision theory does.

[2] Cohesive decision theories do take what you've learned into account when prescribing acts. Cohesive decision theories prescribe the acts selected by the comprehensive strategies which maximize cohesive expected utility. And comprehensive strategies are functions from decision problems—ordered triples consisting of the agent's credences, utilities and the set of available acts—to a subset of the available acts. So even though cohesive decision theories are insensitive to your current credences when evaluating comprehensive strategies, the comprehensive strategies themselves are sensitive to your current credences when recommending acts. As a result, what you've learned does end up getting taken into account by cohesive decision theories, once we get down to the level of which acts you ought to perform.

[3] Several people have offered the following complaint about cohesive decision theory: "Why should the agent choose acts that seem reasonable according to her initial credences? She should choose acts that seem reasonable according to her current credences, not her initial ones." Although this is a natural worry, it's difficult to cash it out in a compelling way.

One might be asking why an agent should do what's reasonable according to her initial credences instead of what's reasonable according to her current credences. But if we're assuming that what is reasonable is what standard decision theory prescribes—expected utility maximization—then this is question begging. And if we're assuming that what is reasonable is what the cohesive decision theory in question prescribes, then an agent who satisfies cohesive decision theory *is* doing what's reasonable according to her current credences.

Alternatively, one might be asking why one ought to believe that the cohesive decision theory in question is the right decision rule. But the answer to this question is straightforward: we're justified in thinking that a version of cohesive decision theory is the right rule to the extent to which it provides the intuitively correct prescriptions. And, as we've seen, there are several ways in which cohesive decision theories are arguably more appealing than standard decision theory.

[4] The content of these theories hangs on how we understand the *ic* function (2) and (3) employ. One possibility is to take *ic* at face value, as the agent's first credence function. This option is uncomfortable, however, since the use of the subject's first credence function (as opposed to her last one, say) seems arbitrary. A second possibility, and a more attractive one, is to take *ic* to be something like the agent's "ur-priors"—the credences the agent ought to have if she had no evidence whatsoever. (Objective Bayesians will hold that all agents have the same ur-prior, while subjective Bayesians will hold that different agents can have different ur-prior functions.) A third and related possibility, suggested by Dennis Whitcomb, is to take *ic* to be something like the initial credences of an ideal subject in the agent's situation. This would allow us to think of cohesive decision theories as a kind of "ideal observer theory" of prudential rationality. (If we're considering (2), we might also take *iu* to be the initial utilities of this ideal subject. We could call the resulting theory the "WW(baby)JD?"-theory.)

[5] One can think of cohesive decision theory as an attempt to allow for a kind of coordination between one's actions at different times. One might want to allow for a similar kind of coordination between the actions of different agents. Here's a natural way to formulate such a rule, given a commitment to something like objective Bayesianism. Let a *global comprehensive strategy* (GCS) pair every possible agent with a comprehensive strategy. Let *oup* stand for the objective ur-prior function. (The 'objective' part allows us to avoid the awkward task of selecting an agent to get priors from.) Let the "global cohesive expected utility" of a global cohesive strategy be:

$$GCoEU(GCS) = \sum_i cr(oup_i) \sum_j oup_i(w_j : GCS) \cdot u(w_j). \tag{4}$$

We can then characterize *global cohesive decision theory* as the rule which prescribes performing the act picked out for you by the global comprehensive strategy which maximizes global cohesive expected utility.

5 Conclusion

The Binding Principle that Arntzenius *et al* employ has a number of interesting consequences. It bolsters evidential decision theory by undercutting the second line of response to the "why ain'cha rich" argument. It bolsters causal decision theory by undercutting the decision instability arguments. It allows standard decision theory to escape the blame for driving agents to disastrous outcomes in the Satan's Apple case. And it provides standard decision theory with an excuse for failing to be self-recommending in certain ways. All told, the Binding Principle offers an appealing way of maintaining the front-runner status of standard decision theory.

Unfortunately, the Binding Principle is implausible. The Binding Principle is analogous to the Mixed Acts Principle, and is problematic for the same reasons. As a result, the various marks against standard decision theory described above still apply.

This makes rules like (2) and (3) more attractive by comparison. These rules escape "why ain'cha rich" and decision instability arguments, don't drive agents to disaster in scenarios like the Satan's Apple case, and are appropriately self-recommending.[35] For similar reasons, the proposals of Bratman (1987), Gauthier (1994) and McClennen (1990) are more attractive in this light.

Of course, some form of standard decision theory may still be the most plausible decision rule. One's assessment of a rule depends on lots of factors besides those considered here. And rules like (2) and (3) have counterintuitive consequences of their own.[36] But none of this changes the fact that the problems described above are demerits of standard decision theory. And when we assess the pros and cons of standard decision theory, they should be taken as such.[37]

References

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

Arntzenius, Frank, Adam Elga and John Hawthorne. 2004. "Bayesianism, Infinite Decisions, and Binding." *Mind* 113:251–283.

[35] Though we need to proceed with caution regarding this last claim. The characterization of being self-recommending we've been working with presupposes decision rules of the standard form—rules whose prescriptions are functions of the agent's current credences, utilities, and the available acts (see section 3.3). But (2) is not a rule of this kind (though (3) is). Likewise, rules like those of Bratman (1987), Gauthier (1994) and McClennen (1990) are not of this form.

[36] For example, the evidential versions of these rules will prescribe the one-boxing response to the Gibbard and Harper case discussed in section 3.2.2. And these rules will recommend that the alcoholic discussed in section 3.3 go to the bar, even though she believes she will probably start drinking if she does.

[37] I would like thank Frank Arntzenius, Philip Bricker, Maya Eddon, Adam Elga, David Etlin, Barry Lam, Ted Sider, Dennis Whitcomb, participants of the Second Formal Epistemology Festival, and participants of the Bellingham Summer Philosophy Conference, for helpful comments and discussion.

Bratman, Michael. 1987. *Intentions, Plans and Practical Reason*. Harvard University Press.

Collins, John. 1996. "Supposition and Choice: Why 'Causal Decision Theory' is a Misnomer." Presented at the CUNY Graduate Center Philosophy Colloquium.

Eells, Ellery. 1985. "Weirich on Decision Instability." *Australasian Journal of Philosophy* 63:473–478.

Egan, Andy. 2007. "Some Counterexamples to Causal Decision Theory." *Philosophical Review* 116:93–114.

Gauthier, David. 1994. "Assure and Threaten." *Ethics* 104:690–716.

Gibbard, Allan and William Harper. 1985. Counterfactuals and Two Kinds of Expected Utility. In *Paradoxes of Rationality and Cooperation: Prisoner's Dilemma and Newcomb's Problem*, ed. R Campbell and L Sowden. University of British Columbia Press.

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

Jeffrey, Richard. 1983. *The Logic of Decision*. University of Chicago Press.

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

Lewis, David. 1981. "Why ain'cha rich?" *Nous* 15:377–380.

McClennen, Edward. 1990. *Rationality and Dynamic Choice*. Cambridge University Press.

Nozick, Robert. 1969. Newcomb's Problem and Two Principles of Choice. In *Essays in Honor of Carl G. Hempel*, ed. Nicholas Rescher. Reidel: Dordretcht.

Richter, Reed. 1986. "Further Comments on Decision Instability." *Australasian Journal of Philosophy* 64:345–349.

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

Sobel, Howard. 1983. "Expected Utilities, and Rational Actions and Choices." *Theoria* 49:159–183.

Weirich, Paul. 1985. "Decision Instability." *Australasian Journal of Philosophy* 63:465–472.