The Sure Thing Principle Leads to Instability

J. DMITRI GALLOW †

Orthodox causal decision theory is unstable. Its advice changes as you make up your mind about what you will do. Several have objected to this kind of instability and explored stable alternatives. Here, I'll show that explorers in search of stability must part with a vestige of their homeland. There is no plausible stable decision theory which satisfies Savage's Sure Thing Principle. So those in search of stability must learn to live without it.

1 | INTRODUCTION

Orthodox causal decision theory is *unstable*, in the sense that its recommendations can change as you change your mind about which choice you'll make. Some of us think instability is a problem with causal decision theory. We are interested in exploring alternative theories whose advice does not vary as you change your mind about which option you're most likely to choose.[1] Here, I'll show that these explorers must part with a vestige of their homeland. Any plausible stable successor to causal decision theory will violate a causal variant of Savage (1954)'s Sure Thing Principle. To appreciate what this principle says, suppose I bet you that the world ends tonight, and whether you take the bet doesn't affect whether it ends or not. Of course, if the world ends, no money will change hands. So, on the supposition that the world ends, taking the bet is as rational as not taking it. The principle says, if that's so, then you don't have to worry about what happens when the world ends. Just suppose it doesn't end, and ask yourself whether taking the bet is more rational than not taking it, given this supposition.

Some notation: write '$A \approx B \mid \neg E$' to mean that A is just as rational as B on the supposition that $\neg E$; write '$A \succeq B$' to mean that A is as rational as B; and write '$A \succeq B \mid E$' to mean that A is as rational as B on the supposition that E. Then, the principle says:

Final draft. Forthcoming in *the Philosophical Quarterly*.

† Thanks to David James Barnett, Zachary Goodsell, Brian Weatherson, and two anonymous referees for helpful conversation and feedback on this material.

1. See, for instance, Skyrms, 1986, 1990, Egan, 2007, Arntzenius, 2008, Wedgwood, 2013, Joyce, 2012, 2018, Spencer, 2021*a*,*b*, Barnett, 2022, Gallow, 2020, and Podgorski, 2022.

Sure Thing Principle If whether E is true is causally independent of how you choose and $A \approx B \mid \neg E$, then: $A \succeq B$ iff $A \succeq B \mid E$.[2]

Orthodox causal decision theory satisfies this principle. But I'll show that, given some minimal assumptions, no stable decision theory does.

There is a weaker version of the Sure Thing Principle which says: if whether E is causally independent of how you choose, and A would get you exactly the same outcome as B whenever E is false, then A is as rational as B if and only if A is as rational as B on the supposition that E. That is:

Weak Sure Thing Principle If whether E is true is causally independent of how you choose and A and B would lead to exactly the same outcome whenever E is false, then: $A \succeq B$ iff $A \succeq B \mid E$.

I'll show that no plausible stable decision theory satisfies this principle, either.

The lesson is that we must either learn to live with instability or learn to live without the Sure Thing Principle.

2 | STABILITY

Call a decision between two options, A and B, *self-frustrating* if, when facing the decision, you know 1) if you choose A, then B would make things better than A will; and 2) if you choose B, then A would make things better than B will. And say that a decision between two options, A and B, is *self-reinforcing* if, when making the decision, you know 1) if you choose A, then A will make things better than B would; and 2) if you choose B, then B will make things better than A would.[3]

Self-frustrating and self-reinforcing decisions can arise in a variety of ways. For instance, they can arise when your evidence creates correlations between your choice and causally-independent states of the world. Suppose you've travelled back in time to save the library of Alexandria. You know that one of two Roman generals has orders to start the fire which burns down the library, but you don't know which. But you do know, from your history books, that the library burns. (Let's take for granted that there is a single, unchanging timeline.[4]) So you know that, if you in fact stop the left general, then the right one was given the orders, so that stopping the right one would have prevented the fire. And you know that, if you in fact stop the right

2. Cf. Jeffrey, 1982 and Joyce, 2007.

3. Cf. Hare & Hedden, 2016.

4. See Lewis, 1976.

one, then the left one was given the orders, so that stopping the left one would have prevented the fire. So you are facing a self-frustrating decision.[5]

Decisions like these can also arise when the value you aim to promote depends upon what the actual world is like. Suppose that only the well-being of actual people matter for determining the overall goodness of a possibility. The interests of merely possible people count for nothing.[6] Then, if you are choosing which of two embryos to fertilise, and you know that the person who develops from the fertilised embryo will have a life worth living, you are facing a self-reinforcing decision. If you fertilise the left embryo, it will develop into an actual person whose interests count and whose life is worth living. The interests of the merely possible person who would have developed from the right embryo count for nothing. So things would have been worse, had you fertilised the right embryo instead. And the same can be said the other way around. If you in fact fertilise the right embryo, then only *its* interests will count, and you will therefore have made things better than fertilising the left embryo would have. So your decision is self-reinforcing.

Faced with decisions like these, some decision theories are *unstable*, in the sense that their recommendations change as you change your mind about which choice you'll make. For instance, if you start out thinking that you will stop the right general, then orthodox causal decision theory will say that you are required to stop the left general—until you start to think that you *will* stop the left general, at which point it will say that you are required to stop the right general.[7] In contrast, evidential decision theory is stable.[8] Its verdicts don't depend upon how likely you think you are to choose each available option. And there are also several heterodox versions of causal decision theory which are stable.[9]

Many of us find instability troubling. We think that, if A is a rational choice, then, if you choose A, you will make a rational choice. We think that, if a theory grants permission to choose A, this permission should not be retracted as soon as it is exercised. A permission which is retracted whenever it is exercised is not a genuine permission. But, because of its instability, causal decision theory disagrees. When you think you're

5. This kind of decision is discussed in Egan, 2007.

6. See, for instance, Parsons, 2002, Hare, 2007, and Cohen, 2020.

7. This theory is defended by Stalnaker, 1981, Gibbard & Harper, 1978, Skyrms, 1982, Lewis, 1981, Sobel, 1994, Joyce, 1999, and Armendt, 2019, among others. For discussion of causal decision theory's instability, see Gibbard & Harper, 1978, Hunter & Richter, 1978, Weirich, 1985, Harper, 1986, Skyrms, 1990, Egan, 2007, Arntzenius, 2008, Joyce, 2012, 2018, Armendt, 2019, Bales, 2020, Spencer, 2021*a*,*b*, and Williamson, 2021.

8. See, for instance, Jeffrey, 1965, 2004, and Ahmed, 2021.

9. See, for instance, Egan, 2007, Arntzenius, 2008, Wedgwood, 2013, Spencer, 2021*b*, Barnett, 2022, Gallow, 2020, and Podgorski, 2022, among others.

most likely to stop the right general, it says that stopping the left general is rational. It offers you permission to stop the left general. No sooner is this permission exercised than it is retracted. If you choose to stop the left general, you will learn that you are doing so, and so the choice you make will be (at the time you make it) an irrational one. It's a rational choice, but if you choose it, your choice will be irrational. Many of us find this unnatural.

We likewise think that, if A is an irrational choice, then, if you choose A, you will make an irrational choice. We think that, if a theory forbids you from doing A, this forbiddance should not be retracted as soon as it is violated. But, because of its instability, causal decision theory disagrees. When you think you're most likely to fertilise the right embryo, it says that fertilising the left embryo is irrational. It forbids you from fertilising the left embryo. No sooner is this forbiddance violated than it is retracted. If you choose to fertilise the left embryo, you will learn that you are doing so, and so the choice you make will be (at that time) a rational one. It's an irrational choice, but if you choose it, your choice will be rational. Again, we find this unnatural.

I'm going to suppose that, provided with any well-formed decision, a decision theory will tell you which acts are as rational as which others. So it will provide an ordering over available acts, $\succeq$, where $A \succeq B$ iff A is as rational a choice as B is. We can then define $A \succ B$ (A is more rational than B) and $A \approx B$ (A is just as rational as B) in the usual way.[10] If A is as rational as every other option, then I'll say that it is rational, full stop. (Throughout, I'm going to use 'act' and 'option' interchangeably.) I'll limit my attention to decisions in which the outcome is completely determined by two factors: your choice and a state of the world which is causally independent of your choice. And I'll suppose that you have probabilities for how likely each state is, conditional on you making each choice. For illustration, consider

Newcomb Here is $1000. You can either take it, T, or refuse it, R. Yesterday, I made a prediction about how you'd choose. If I predicted that you'd take it, K_T, then I deposited $1,000,000 in your bank account. If I predicted that you'd refuse, K_R, then I did not. My predictions are pretty reliable. Conditional on any choice you make, your probability that I correctly predicted that choice is 80%.

In this decision, there are two available acts, T and R. And there are two states of the world, K_T and K_R, each of which is causally independent of your choice. Your choice and the state of the world together determine the relevant outcome. And you have probabilistic opinions about how likely each state of the world is, conditional on you choosing each act.

10. $A \succ B$ iff $A \succeq B$ and $\neg B \succeq A$. And $A \approx B$ iff $A \succeq B$ and $B \succeq A$.

We can package this kind of information together into two matrices. For instance, on the left below, we have the matrix $\mathbf{V}(row \wedge col)$, which tells us the *v*alue of choosing each act (specified in the row) in each state (specified in the column). And on the right, we have the matrix $\mathbf{P}(row \mid col)$, which gives your subjective probability that each state (specified in the row) obtains, conditional on you performing each act (specified in the column).

$$\begin{array}{c} \mathbf{V}(row \wedge col) \\ T \\ R \end{array} \begin{array}{c} \begin{array}{cc} K_T & K_R \end{array} \\ \begin{bmatrix} 1000 & 1,001,000 \\ 0 & 1,000,000 \end{bmatrix} \end{array} \qquad \begin{array}{c} \mathbf{P}(row|col) \\ K_T \\ K_R \end{array} \begin{array}{c} \begin{array}{cc} T & R \end{array} \\ \begin{bmatrix} 80\% & 20\% \\ 20\% & 80\% \end{bmatrix} \end{array} \tag{1}$$

(I'm assuming here that your values are linear in dollars.)

I'll call a decision theory *stable* just in case it only needs this kind of information in order to say which options are as rational as which others. I will assume, by the way, that a stable decision theory will provide us with a total pre-order over options. That is, I'll assume that, for any options A, B, and C, a stable decision theory will tell us 1) $A \succeq A$; 2) if $A \succeq B$ and $B \succeq C$, then $A \succeq C$; and 3) either $A \succeq B$ or $B \succeq A$. If the decision theory gives us an irreflexive, intransitive, or non-total ordering, then it doesn't count as a stable decision theory, in my terminology.[11]

Stability A decision theory is *stable* iff it determines a total pre-order over options given only the value of each act in each state, $\mathbf{V}$, and your probability in each state, conditional on you selecting each act, $\mathbf{P}$.

Call your probabilities over the partition of propositions about which of the available acts you'll choose your *act probabilities*. Your act probabilities give your opinions about how you'll choose. If a decision theory is stable, then its verdicts won't change as you change your act probabilities. That's because neither $\mathbf{V}$ nor $\mathbf{P}$ will change as you shift your act probabilities.[12]

Just to fix ideas, let me give one example of a stable decision theory. Notice that you can use your act probabilities, together with the value of each act in each state, $\mathbf{V}$, and the probability of each state, given each act, $\mathbf{P}$, to calculate the expected value of each act.[13] Then, say that your act probabilities are *in equilibrium* iff they meet the

11. You might worry about the assumption of totality because you think that there can be rational incomparabilities, where neither A nor B is at least as rational as the other. If you have this view, then you may interpret '$A \succeq B$' as meaning 'it's not the case that A is less rational than B'. So long as 'A is at least as rational as B' is a pre-order (reflexive and transitive), 'A is not less rational than B' will be a total pre-order.

12. I am taking for granted that you'll shift those probabilities in the way recommended by Jeffrey, 1965.

13. That's because $\mathbf{V}$ and $\mathbf{P}$ together determine the *conditional* expected value of each act, A, given any act,

following two requirements. Firstly, if you calculate expected values with those act probabilities, then every act with positive probability will have an expected value at least as great as the expected value of any act with zero probability. And secondly, if you calculate expected values with those act probabilities, then any two acts with positive probability will have the same expected value. For instance, in your self-frustrating decision at the library of Alexandria, the unique equilibrium act probabilities give a 50% probability to stopping the left general and a 50% probability to stopping the right general. And in your self-reinforcing decision with the embryos, there are three equilibrium act probabilities: the one which is 100% sure you'll fertilise the left embryo, the one which is 100% sure you'll fertilise the right embryo, and the one which is 50% sure of left and 50% sure of right. If your act probabilities aren't in equilibrium, then you'll think that you might choose an option which has a lower expected value than some other option you could choose instead. Whereas, if your act probabilities are in equilibrium, then you'll think that every option you might choose maximises expected value.

Then, here's a sample stable decision theory: an option, A, is rational iff there is an equilibrium act probability which gives positive probability to A. Because we can calculate which act probabilities are in equilibrium just from the information in the matrices $\mathbf{V}$ and $\mathbf{P}$, this theory will be stable.[14]

3 | STABILITY AND THE SURE THING PRINCIPLE

In the appendix, I show that, given some minimal assumptions, there is no stable decision theory which satisfies either the Sure Thing Principle or the Weak Sure Thing Principle. In particular, the sample stable theory I provided in §2 does not satisfy either of these principles. In this section, I'll describe the additional assumptions needed for these results. They are each assumptions about what any plausible stable decision theory will look like.

Some terminology: say that an option, A, is *uniformly dominated* by another option, B, iff there's some number x such that 1) in every state, A leads to an outcome worth less than x utiles; and 2) in every state, B leads to an outcome worth more than

---

B, $EV(A \mid B)$. And these values, together with your act probabilities, determine the unconditional expected value of each act, since $EV(A) = \sum_B EV(A \mid B) \cdot P(B)$, by the law of iterated expectations.

14. Theories like this are defended by Skyrms, 1990, Arntzenius, 2008, and Joyce, 2012, though none of these authors quite endorse the version of the theory from the body. Skyrms and Joyce will give difference advice, depending upon what your *initial* act probabilities are. For this reason, their theories won't count as stable. (Though, if you bear them in mind when going through the proofs in the appendix, you'll see that they both still violate the Sure Thing Principle.) Arntzenius will think that some equilibrium act probabilities are better than others; and he'll say that an act is rational iff it is given positive probability in one of the *best* equilibria. Arntzenius's theory will count as stable.

x utiles. And say that an option is *universally uniformly dominated* iff it is uniformly dominated by *every other* option. I'll assume options like this are irrational.

Avoid Universal Uniform Domination (AUUD) A universally uniformly dominated option is irrational.

Being uniformly dominated is a stronger condition than being dominated. The difference has to do with the order of the quantifiers. A dominates B iff, in every state, there's some number x ($\forall\exists$) such that A leads to an outcome worth less than x utiles and B leads to an outcome worth more than x utiles. A *uniformly* dominates B iff there's some number x such that, in every state ($\exists\forall$), A leads to an outcome worth less than x utiles and B leads to an outcome worth more than x utiles. An evidential decision theorist will think that, in Newcomb-like problems, there are dominated options which are rational. But even an evidential decision theorist will accept that uniformly dominated options are irrational. And being *universally* uniformly dominated is a stronger condition still. Some of us have worried about dominance principles in complicated decisions where a dominated option enters into cycles of pairwise rational preference (where, given a pairwise decision between any two adjacent options in the cycle, it would be rational to take the second and irrational to take the first).[15] But these kinds of cases can't arise for a *universally* dominated option—much less a universally uniformly dominated option.

I will also need a second assumption. I think it should be uncontroversial when understood, but it will take a bit of time to explain. It says that, at least for some agents, rationality doesn't distinguish between *known* quantities and *expected* quantities. For instance, if you're neither risk-seeking nor risk-averse, then you should treat a 50% chance of getting \$100 the same as you treat a guaranteed \$50. (I'm still assuming your values are linear in dollars.) Some think that this is rationally required. Others disagree and think that, depending on your attitudes towards risk, it can be rational to prefer \$50 to a 50% chance of \$100.[16] But all should agree that *risk-neutral people*—people who are neither risk-averse nor risk-seeking—shouldn't distinguish between \$50 and a 50% shot at \$100.

By definition, a stable decision theory only needs to be provided with the information in a pair of matrices, $(\mathbf{V}, \mathbf{P})$, where the first matrix, $\mathbf{V}$, gives the value of choosing each act in each state, and the second matrix, $\mathbf{P}$, gives your probability in each state, conditional on you choosing each act. Multiplying these two matrices together gives us your *expectation* of the value of choosing each act, conditional on you

15. See Spencer & Wells, 2019 and Gallow, 2020.

16. See Buchak, 2013.

choosing every other act. For instance, in Newcomb, if we multiply together the two matrices from (1), we'll get the following matrix.

$$(2) \qquad \begin{array}{c} \mathbf{EV}(row|col) \\ T \\ R \end{array} \begin{array}{c} \begin{array}{cc} T & R \end{array} \\ \begin{bmatrix} 201,000 & 801,000 \\ 200,000 & 800,000 \end{bmatrix} \end{array}$$

This matrix tells us that, conditional on you choosing T, the *expected* value of choosing T is $201,000$, whereas the *expected* value of choosing R is $200,000$. And, conditional on choosing R, the *expected* value of choosing T is $801,000$, whereas the *expected* value of choosing R is $800,000$.

Given this matrix of *expected* values, we could construct another decision exactly like Newcomb, except that expected values have been swapped out for known values. For instance, we could construct a new decision where you must decide between two boxes labelled 'T' and 'R', and where I made a prediction about how you'd choose. But this time, we could suppose that my prediction is certain to be correct. So, conditional on you choosing T, you are *certain* that I predicted you'd choose T; and, conditional on you choosing R, you are *certain* that I predicted that you'd choose R. Moreover, if I predicted that you'd choose T, then I put \$201,000 in T and \$200,000 in R. And, if I predicted you'd choose R, then I put \$801,000 in T and \$800,000 in R. That is, we could construct a decision characterised by the following two matrices.

$$(3) \qquad \begin{array}{c} \mathbf{V}(row \wedge col) \\ T \\ R \end{array} \begin{array}{c} \begin{array}{cc} K_T & K_R \end{array} \\ \begin{bmatrix} 201,000 & 801,000 \\ 200,000 & 800,000 \end{bmatrix} \end{array} \qquad \begin{array}{c} \mathbf{P}(row|col) \\ K_T \\ K_R \end{array} \begin{array}{c} \begin{array}{cc} T & R \end{array} \\ \begin{bmatrix} 100\% & 0 \\ 0 & 100\% \end{bmatrix} \end{array}$$

The second assumption I'll need is that a stable decision theory will not distinguish between a decision characterised by (1) and a decision characterised by (3). The 'perfect correlation' matrix on the right in (3) above is what's known as the 'identity matrix'. I'll denote it with '$\mathbf{I}$'. Then, a concise way of expressing my second assumption is that a stable decision theory will not distinguish between a decision characterised by $(\mathbf{V}, \mathbf{P})$ and a 'perfect correlation' decision characterised by $(\mathbf{V} \cdot \mathbf{P}, \mathbf{I})$. (Remember, the product $\mathbf{V} \cdot \mathbf{P}$ gives us the *expected* value of each act, conditional on you selecting each act.)

Risk Neutrality If you are a risk neutral agent, then a stable decision theory will say that $A \succeq B$ in a decision characterised by $(\mathbf{V}, \mathbf{P})$ iff it says that $A \succeq B$ in a decision characterised by $(\mathbf{V} \cdot \mathbf{P}, \mathbf{I})$.

Risk Neutrality *doesn't* say that it is irrational to be risk-averse or risk-seeking. It simply says something about which acts are as rational as which others if you happen

to be risk-neutral.

In the appendix, I prove the following proposition.

Proposition 1. *There is no stable decision theory which satisfies AUUD, Risk Neutrality, and the Sure Thing Principle.*

If we take AUUD and Risk Neutrality for granted—as I think we should—this shows us that we must choose between stability and the Sure Thing Principle. The sample stable decision theory from §2 satisifies both Risk Neutrality and AUUD. So the proposition teaches us that that theory violates the Sure Thing Principle.

The Weak Sure Thing Principle also leads to instability. There is no plausible stable decision theory which satisfies the Weak Sure Thing Principle. To show this, I'll have to make a different assumption. The assumption is that universally uniformly dominated options aren't just *irrational*—they are, moreover, *ignorable*. That is, if an option is universally uniformly dominated, then you can remove it from the menu of options and decide between the remaining options as though it weren't there.

Ignore Universal Uniform Domination (IUUD) An option is rational only if it would remain rational after a universally uniformly dominated option has been removed from the menu of options.

No option can remain a rational choice after it itself has been removed from the menu of options. So IUUD implies AUUD.

I think any *stable* decision theory should satisfy this principle. But it's worth emphasising that—precisely because of its instability—orthodox causal decision theory does not. To understand why, consider

Three Doors You must decide between three doors, labelled '*P*', '*Q*', and '*X*'. If you choose *X*, you're guaranteed to lose $100. If you choose *P*, you're guaranteed to gain $100. If I predicted that you'd choose either *P* or *Q*, then I left nothing behind *Q*. If, however, I predicted that you'd choose *X*, then I left $101 behind *Q*. You are certain that I correctly predicted your choice. And, at the moment, you think you're very likely to choose *X*.

The value of each act in each state, $\mathbf{V}$, and the conditional probability of each state, given each act, $\mathbf{P}$, are shown in the matrices in (4) below. ('K_A' is the state in which I predicted you'd choose *A*.)

$$(4)\quad \begin{array}{c} \mathbf{V}(row \wedge col) \\ P \\ Q \\ X \end{array} \begin{array}{c} \begin{array}{ccc} K_P & K_Q & K_X \end{array} \\ \begin{bmatrix} 100 & 100 & 100 \\ 0 & 0 & 101 \\ -100 & -100 & -100 \end{bmatrix} \end{array} \qquad \begin{array}{c} \mathbf{P}(row|col) \\ K_P \\ K_Q \\ K_X \end{array} \begin{array}{c} \begin{array}{ccc} P & Q & X \end{array} \\ \begin{bmatrix} 100\% & 0 & 0 \\ 0 & 100\% & 0 \\ 0 & 0 & 100\% \end{bmatrix} \end{array}$$

In this decision, orthodox causal decision theory says that you are required to choose Q. Because you are very confident that you'll end up choosing X, you are very confident that there is more money behind Q than there is behind P. So the expected value of Q is greater than the expected value of P—given your confidence that you'll choose X. (Of course, were you to take causal decision theory's advice to choose Q, and learn that you had, it would then tell you that P is the only rational choice.)

X is a universally uniformly dominated option. And, were X removed from the menu of options, you'd be certain that you were not predicted to choose X. So you'd be certain that there's \$100 behind P and nothing behind Q. In that case, orthodox causal decision theory would say that P is the only rational option. So, in Three Doors, orthodox causal decision theory says that Q is more rational than P, and, if a universally uniformly dominated option were removed from the menu, P would be more rational than Q. So it violates the principle IUUD.

My goal here isn't to criticise orthodox causal decision theory for issuing verdicts like this. But I want to emphasise that verdicts like this are precisely what motivate many of us to search for stable revisions to causal decision theory in the first place. Those in search of stability should want to say that, in Three Doors, P is more rational than Q. They should want to endorse the following reasoning: Both P and Q are guaranteed to get you something better than any outcome X could get you. So you shouldn't choose X. And you know for sure that, so long as you don't choose X, there's \$100 behind P and nothing behind Q. So you should choose P. And that's just the kind of reasoning which IUUD encodes.

In the appendix, I prove the following proposition.

Proposition 2. *There is no stable decision theory which satisfies Risk Neutrality, IUUD, and the Weak Sure Thing Principle.*

The sample stable decision theory from §2 satisfies both Risk Neutrality and IUUD. So the proposition teaches us that that theory violates the Weak Sure Thing Principle. More generally, if we take it for granted that any plausible stable decision theory should satisfy Risk Neutrality and IUUD—as I think we should—this shows us that we must choose between stability and the Weak Sure Thing Principle.

Lemma 1. *Any stable decision theory which satisfies Risk Neutrality will also satisfy*

Conditional Expected Values are Sufficient (CEVS) *For risk-neutral agents, whether one act is as rational as another is entirely determined by the expected value of each act, A, conditional on you selecting any act, B, $EV(A \mid B)$.*

Proof. Any stable decision theory's verdicts must be a function of a pair $(\mathbf{V}, \mathbf{P})$ of the value of each choice in each state, and your probability for each state, conditional on each choice. Risk Neutrality tells us that, if you're risk-neutral, and if $\mathbf{V} \cdot \mathbf{P} = \mathbf{V}^* \cdot \mathbf{P}^*$, then a stable decision theory's verdicts about both $(\mathbf{V}, \mathbf{P})$ and $(\mathbf{V}^*, \mathbf{P}^*)$ must be the same as its verdicts about $(\mathbf{V} \cdot \mathbf{P}, \mathbf{I}) = (\mathbf{V}^* \cdot \mathbf{P}^*, \mathbf{I})$, where $\mathbf{I}$ is the identity matrix. So its verdicts about $(\mathbf{V}, \mathbf{P})$ must be the same as its verdicts about $(\mathbf{V}^*, \mathbf{P}^*)$. So a stable decision theory won't distinguish between a decision characterised by $(\mathbf{V}, \mathbf{P})$ and one characterised by $(\mathbf{V}^*, \mathbf{P}^*)$, so long as $\mathbf{V}\cdot\mathbf{P} = \mathbf{V}^*\cdot\mathbf{P}^*$. The product $\mathbf{V}\cdot\mathbf{P}$ is just a matrix of conditional expected values, $\mathbf{EV}(row \mid col)$. So the theory's verdicts must be a function of the conditional expected values alone, and so it must validate CEVS. □

Lemma 2. *Let α lie in the open unit interval $(0, 1)$ and let β be any constant. Assuming the Weak Sure Thing Principle and CEVS, we may take any given column of a conditional expected value matrix and replace each entry x with $\alpha \cdot x + \beta(1 - \alpha)$, without making any difference to rational choice.*

Proof. This proof generalises an argument from Weatherson, ms. Hand me any conditional expected value matrix you like,

(5)
$$\begin{array}{c} \mathbf{EV}(row|col) \\ A_1 \\ A_2 \\ \vdots \\ A_N \end{array} \begin{array}{c} \begin{array}{cccc} A_1 & A_2 & \dots & A_N \end{array} \\ \begin{bmatrix} x_{11} & x_{21} & \dots & x_{N1} \\ x_{12} & x_{22} & \dots & x_{N2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1N} & x_{2N} & \dots & x_{NN} \end{bmatrix} \end{array}$$

any $\alpha \in (0, 1)$, and any $\beta \in \mathbb{R}$. Then, consider a decision characterised by the matrices $\mathbf{V}$ and $\mathbf{P}$ shown below.

(6)
$$\begin{array}{c} \mathbf{V}(row\wedge col) \\ A_1 \\ A_2 \\ \vdots \\ A_N \end{array} \begin{array}{c} \begin{array}{ccccc} K_0 & K_1 & K_2 & \dots & K_N \end{array} \\ \begin{bmatrix} \beta & x_{11} & x_{21} & \dots & x_{N1} \\ \beta & x_{12} & x_{22} & \dots & x_{N2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \beta & x_{1N} & x_{2N} & \dots & x_{NN} \end{bmatrix} \end{array} \qquad \begin{array}{c} \mathbf{P}(row|col) \\ K_0 \\ K_1 \\ K_2 \\ \vdots \\ K_N \end{array} \begin{array}{c} \begin{array}{cccc} A_1 & A_2 & \dots & A_N \end{array} \\ \begin{bmatrix} 1-\alpha & 0 & \cdots & 0 \\ \alpha & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} \end{array}$$

Multiplying these matrices together gives us the conditional expected values in (7).

$$
(7) \qquad \begin{array}{c} \mathbf{EV}(row|col) \\ A_1 \\ A_2 \\ \vdots \\ A_N \end{array} \begin{array}{c} \begin{array}{cccc} A_1 & A_2 & \dots & A_N \end{array} \\ \begin{bmatrix} \alpha \cdot x_{11} + \beta(1-\alpha) & x_{21} & \dots & x_{N1} \\ \alpha \cdot x_{12} + \beta(1-\alpha) & x_{22} & \dots & x_{N2} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha \cdot x_{1N} + \beta(1-\alpha) & x_{2N} & \dots & x_{NN} \end{bmatrix} \end{array}
$$

which is just the matrix from (5), with each entry from column 1, x, replaced with $\alpha \cdot x + \beta(1-\alpha)$.

Take any two options A_i, A_j. If K_0, A_i and A_j are guaranteed to get you the same outcome, namely β. So, in this decision, the Weak Sure Thing Principle says that $A_i \geq A_j$ iff $A_i \geq A_j \mid \neg K_0$. Conditional on $\neg K_0$, your values will be unchanged, but your probabilities will be given by

$$
(8) \qquad \begin{array}{c} \mathbf{P}(row|col) \\ K_0 \\ K_1 \\ K_2 \\ \vdots \\ K_N \end{array} \begin{array}{c} \begin{array}{cccc} A_1 & A_2 & \dots & A_N \end{array} \\ \begin{bmatrix} 0 & 0 & \cdots & 0 \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} \end{array}
$$

And multiplying the values from the left-hand side of (6) by the probabilities from (8) will give us back the original matrix of conditional expected values from (5). So, assuming CEVS and the Weak Sure Thing Principle, a decision theory must say precisely the same thing when it's handed the conditional expected values in (7) as it does when it's handed the conditional expected values in (5). Of course, there's nothing special about the first column here. A similar argument shows that we can pick any column and uniformly replace each entry x in that column with $\alpha \cdot x + \beta(1-\alpha)$ without making any difference to rational choice. □

Corollary 1. *Let a be any positive constant and let b be any constant. Assuming the Weak Sure Thing Principle and CEVS, we may take any given column of a conditional expected value matrix and replace each entry x in that column with $a \cdot x + b$ without making any difference to rational choice.*

Proof. We will establish the corollary in two steps. First, by showing that we can multiply each column by the same positive constant $a > 0$ without making any difference to rational choice. And next, showing that we can add any constant b to any column without making any difference to rational choice. Since 'not making any difference to rational choice' is a transitive relation, the corollary follows.

First, take lemma 2 and fix $\beta = 0$. Then, we have that we can multiply the entries in any column by the same constant $\alpha \in (0, 1)$ without making any difference to rational choice. Since 'not making any difference to rational choice' is a symmetric relation, this shows that we can just as well *divide* the entries in any column by the same constant $\alpha \in (0, 1)$ without making any difference to rational choice. And this shows that we can multiply the entries in

any column by *any* positive constant $a > 0$ without making a difference to rational choice, since every constant $a > 1$ is equal to $1/\alpha$, for some $\alpha \in (0, 1)$.

Next, hand me any constant you like, b. Then, use lemma 2 and fix $\alpha = 1/2$ and $\beta = b$. This tells us that uniformly replacing each entry x in a column with $(x + b)/2$ makes no difference to rational choice. And the first part of the corollary, which we've already established, tells us that multiplying each of those entries by 2 makes no difference to rational choice. Putting these results together, we have that we can replace each entry x in a column with $x + b$, for any constant b you like, without making any difference to rational choice. □

Proposition 1. *There is no stable decision theory which satisfies AUUD, Risk Neutrality, and the Sure Thing Principle.*

Proof. I'll assume that we have such a decision theory and derive a contradiction. Assume that you are risk-neutral, and consider a decision characterised by the following values and probabilities.

(9)
$$\begin{array}{c} \mathbf{V}(row \wedge col) \\ A \\ B \end{array} \begin{array}{c} \begin{array}{ccc} K_0 & K_A & K_B \end{array} \\ \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \end{array} \qquad \begin{array}{c} \mathbf{P}(row|col) \\ K_0 \\ K_A \\ K_B \end{array} \begin{array}{c} \begin{array}{cc} A & B \end{array} \\ \begin{bmatrix} \delta & \delta \\ 1-\delta & 0 \\ 0 & 1-\delta \end{bmatrix} \end{array}$$

By AUUD, A is more rational than B, conditional on K_0, $A \succ B \mid K_0$. We will now show that, conditional on $\neg K_0$, A is just as rational as B, $A \approx B \mid \neg K_0$. Then, we will conclude, from the Sure Thing Principle, that, unconditionally, A is more rational than B, $A \succ B$.

Conditional on $\neg K_0$, your probability for each state, conditional on each act, will be given by (10).

(10)
$$\begin{array}{c} \mathbf{P}(row|col) \\ K_0 \\ K_A \\ K_B \end{array} \begin{array}{c} \begin{array}{cc} A & B \end{array} \\ \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \end{array}$$

Multiplying the values on the left-hand side of (9) by the probabilities from (10) gives the matrix of conditional expected values in (11).

(11)
$$\begin{array}{c} \mathbf{EV}(row|col) \\ A \\ B \end{array} \begin{array}{c} \begin{array}{cc} A & B \end{array} \\ \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \end{array}$$

By lemma 1, any stable decision theory which satisfies Risk Neutrality satisfies CEVS. So the verdict of our hypothesised decision theory must be determined by the conditional expected values in (11). But this matrix is perfectly symmetric with respect to A and B. So it must be that, conditional on $\neg K_0$, A is as rational as B iff B is as rational as A, $A \succeq B \mid \neg K_0 \leftrightarrow B \succeq A \mid \neg K_0$. Since $\succeq$ is a total order, this biconditional implies that, conditional on $\neg K_0$, A is just as rational as B, $A \approx B \mid \neg K_0$.

So, in the unconditional decision characterised by the matrices from (9), we have that $A \succ B \mid K_0$ and $A \approx B \mid \neg K_0$. So, by the Sure Thing Principle, $A \succ B$. Multiplying together the matrices in (9) gives the conditional expected values in (12).

$$(12) \qquad \begin{array}{c} \mathbf{EV}(row|col) \\ A \\ B \end{array} \begin{array}{c} \begin{array}{cc} A & B \end{array} \\ \begin{bmatrix} \delta & 1 \\ 1-\delta & 0 \end{bmatrix} \end{array}$$

By CEVS, then, we have that, presented with the conditional expected values from (12), $A \succ B$.

The Strong Sure Thing Principle implies the Weak Sure Thing Principle. So, by corollary 1, we can subtract δ from every entry in the first column of (12), and then multiply every entry in the first column by $1/(1-2\delta)$ without making any difference to rational choice. This will give us the the matrix of conditional expected values from (11). So, by corollary 1, we must say that A is more rational than B, $A \succ B$, when presented with (11). But as we've already seen, CEVS and totality require us to say that A is just as rational as B, $A \approx B$, when presented with (11). So, presented with (11), we have that $A \succ B$ and $A \approx B$. Contradiction. □

Proposition 2. *There is no stable decision theory which satisfies Risk Neutrality, the Weak Sure Thing Principle, and IUUD.*

Proof. I'll assume that we have such a decision theory and derive a contradiction. Suppose that you are risk-neutral, and consider a decision characterised by the following values and probabilities.

$$(13) \qquad \begin{array}{c} \mathbf{V}(row \wedge col) \\ A \\ B \\ C \end{array} \begin{array}{c} \begin{array}{ccc} K_A & K_B & K_C \end{array} \\ \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \end{array} \qquad \begin{array}{c} \mathbf{P}(row|col) \\ K_A \\ K_B \\ K_C \end{array} \begin{array}{c} \begin{array}{ccc} A & B & C \end{array} \\ \begin{bmatrix} 2/6 & 1/6 & 3/6 \\ 3/6 & 2/6 & 1/6 \\ 1/6 & 3/6 & 2/6 \end{bmatrix} \end{array}$$

In this decision, A, B, and C are perfectly symmetric. So any stable decision theory will say that $A \approx B \approx C$. If K_C, then A and B lead to exactly the same outcome. So the Weak Sure Thing Principle implies that $A \approx B \mid \neg K_C$. Conditional on $\neg K_C$, your decision will be characterised by the matrices in (14).

$$(14) \qquad \begin{array}{c} \mathbf{V}(row \wedge col) \\ A \\ B \\ C \end{array} \begin{array}{c} \begin{array}{ccc} K_A & K_B & K_C \end{array} \\ \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \end{array} \qquad \begin{array}{c} \mathbf{P}(row|col) \\ K_A \\ K_B \\ K_C \end{array} \begin{array}{c} \begin{array}{ccc} A & B & C \end{array} \\ \begin{bmatrix} 2/5 & 1/3 & 3/4 \\ 3/5 & 2/3 & 1/4 \\ 0 & 0 & 0 \end{bmatrix} \end{array}$$

So, faced with the decision in (14), our hypothesised decision theory must say that $A \approx B$. By lemma 1, any stable decision theory which satisfies Risk Neutrality satisfies CEVS. So the verdict of our hypothesised decision theory must say the same thing about the decision in (14)

as the decision in (15) (since they have the same conditional expected values).

(15)
$$\begin{array}{c} \mathbf{V}(row \wedge col) \\ A \\ B \\ C \end{array} \begin{array}{c} \begin{array}{ccc} K_A & K_B & K_C \end{array} \\ \begin{bmatrix} 2/5 & 1/3 & 3/4 \\ 3/5 & 2/3 & 1/4 \\ 0 & 0 & 0 \end{bmatrix} \end{array} \qquad \begin{array}{c} \mathbf{P}(row|col) \\ K_A \\ K_B \\ K_C \end{array} \begin{array}{c} \begin{array}{ccc} A & B & C \end{array} \\ \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \end{array}$$

So, presented with the decision in (15), our hypothesised stable decision theory must say that $A \approx B$. In this decision, C is universally uniformly dominated. So IUUD tells us that we must have $A \approx B$ after the option C has been removed:

(16)
$$\begin{array}{c} \mathbf{V}(row \wedge col) \\ A \\ B \end{array} \begin{array}{c} \begin{array}{ccc} K_A & K_B & K_C \end{array} \\ \begin{bmatrix} 2/5 & 1/3 & 3/4 \\ 3/5 & 2/3 & 1/4 \end{bmatrix} \end{array} \qquad \begin{array}{c} \mathbf{P}(row|col) \\ K_A \\ K_B \\ K_C \end{array} \begin{array}{c} \begin{array}{cc} A & B \end{array} \\ \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \end{array}$$

By CEVS, the verdict of our hypothesised decision theory must say the same thing about the decision in (16) as it does about the decision in (17) (since they have the same conditional expected values).

(17)
$$\begin{array}{c} \mathbf{V}(row \wedge col) \\ A \\ B \end{array} \begin{array}{c} \begin{array}{cc} K_A & K_B \end{array} \\ \begin{bmatrix} 2/5 & 1/3 \\ 3/5 & 2/3 \end{bmatrix} \end{array} \qquad \begin{array}{c} \mathbf{P}(row|col) \\ K_A \\ K_B \end{array} \begin{array}{c} \begin{array}{cc} A & B \end{array} \\ \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \end{array}$$

So, presented with the decision in (17), our hypothesised stable decision theory must say that $A \approx B$. But, in this decision, A is universally uniformly dominated. So IUUD implies that A cannot be rational, which implies that we must have $A \prec B$. So, presented with the decision in (17), we have that $A \approx B$ and $A \prec B$. Contradiction. □

REFERENCES

Ahmed, Arif. 2021. *Evidential Decision Theory*. Cambridge: Cambridge University Press. [3]

Armendt, Brad. 2019. "Causal Decision Theory and Decision Instability." In *The Journal of Philosophy*, **116**: 263–277. [3]

Arntzenius, Frank. 2008. "No regrets, or: Edith Piaf revamps decision theory." In *Erkenntnis*, **68**: 277–297. [1], [3], [6]

Bales, Adam. 2020. "Intentions and instability: a defense of causal decision theory." In *Philosophical Studies*, **177** (3): 793–804. [3]

Barnett, David James. 2022. "Graded Ratifiability." In *The Journal of Philosophy*, **119** (8): 57–88. [1], [3]

Buchak, Lara. 2013. *Risk and Rationality*. Oxford: Oxford University Press. [7]

Cohen, Daniel. 2020. "An Actualist Explanation of the Procreation Asymmetry." In *Utilitas*, **32** (1): 70–89. [3]

Egan, Andy. 2007. "Some Counterexamples to Causal Decision Theory." In *The Philosophical Review*, **116** (1): 93–114. [1], [3]

Gallow, J. Dmitri. 2020. "The Causal Decision Theorist's Guide to Managing the News." In *The Journal of Philosophy*, **117** (3): 117–149. [1], [3], [7]

Gibbard, Allan & Harper, William L. 1978. "Counterfactuals and Two Kinds of Expected Utility." In *Foundations and Applications of Decision Theory*, edited by A. Hooker, J.J. Leach, & E.F. McClennan, Dordrecht: D. Reidel, 125–162. [3]

Hare, Caspar & Hedden, Brian. 2016. "Self-Reinforcing and Self-Frustrating Decisions." In *Noûs*, **50** (3): 604–628. [2]

Hare, Caspare. 2007. "Voices from Another World: Must We Respect the Interests of People Who Do Not, and Will Never, Exist?" In *Ethics*, **117** (3): 498–523. [3]

Harper, William. 1986. "Mixed Strategies and Ratifiability in Causal Decision Theory." In *Erkenntnis*, **24**: 25–36. [3]

Hunter, Daniel & Richter, Reed. 1978. "Counterfactuals and Newcomb's Paradox." In *Synthese*, **39** (2): 249–261. [3]

Jeffrey, Richard. 1965. *The Logic of Decision*. New York: McGraw-Hill. [3], [5]

Jeffrey, Richard. 1982. "The Sure Thing Principle." In *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, **2**: 719–730. [2]

Jeffrey, Richard. 2004. *Subjective Probability: the Real Thing*. Cambridge: Cambridge University Press. [3]

Joyce, James M. 1999. *The Foundations of Causal Decision Theory*. Cambridge: Cambridge University Press. [3]

Joyce, James M. 2007. "Are Newcomb problems really decisions?" In *Synthese*, **156**: 537–562. [2]

Joyce, James M. 2012. "Regret and instability in causal decision theory." In *Synthese*, **187** (1): 123–145. [1], [3], [6]

Joyce, James M. 2018. "Deliberation and Stability in Newcomb Problems and Pseudo-Newcomb Problems." In *Newcomb's Problem*, edited by Arif Ahmed, Oxford: Oxford University Press. [1], [3]

Lewis, David K. 1976. "The Paradoxes of Time Travel." In *American Philosophical Quarterly*, **13** (2): 145–152. [2]

Lewis, David K. 1981. "Causal Decision Theory." In *Australasian Journal of Philosophy*, **59** (1): 5–30. [3]

Parsons, Josh. 2002. "Axiological Actualism." In *Australasian Journal of Philosophy*, **80** (2): 137–147. [3]

Podgorski, Abelard. 2022. "Tournament Decision Theory." In *Noûs*, **56** (1): 176–203. [1], [3]

Savage, Leonard J. 1954. *The Foundations of Statistics*. Wiley Publications in Statistics. [1]

Skyrms, Brian. 1982. "Causal Decision Theory." In *Journal of Philosophy*, **79** (11): 695–711. [3]

Skyrms, Brian. 1986. "Deliberational Equilibria." In *Topoi*, **5** (1): 59–67. [1]

Skyrms, Brian. 1990. *The Dynamics of Rational Deliberation*. Cambridge, MA: Harvard University Press. [1], [3], [6]

Sobel, Jordan Howard. 1994. *Taking Chances: Essays on Rational Choice*. Cambridge: Cambridge University Press. [3]

Spencer, Jack. 2021*a*. "An Argument Against Causal Decision Theory." In *Analysis*, **81** (1): 52–61. [1], [3]

Spencer, Jack. 2021*b*. "Rational Monism and Rational Pluralism." In *Philosophical Studies*, **178**: 1769–1800. [1], [3]

Spencer, Jack & Wells, Ian. 2019. "Why Take Both Boxes?" In *Philosophy and Phenomenological Research*, **99** (1): 27–48. [7]

Stalnaker, Robert C. 1981. "Letter to David Lewis." In *Ifs*, edited by William Harper, Robert Stalnaker, & Glenn Pearce, Dordrecht: D. Reidel Publishing Company, 151–152. [3]

Weatherson, Brian. ms. "Indecisive Decision Theory." [i]

Wedgwood, Ralph. 2013. "Gandalf's solution to the Newcomb Problem." In *Synthese*, **190** (14): 2643–2675. [1], [3]

Weirich, Paul. 1985. "Decision Instability." In *Australasian Journal of Philosophy*, **63** (4): 465–478. [3]

Williamson, Timothy Luke. 2021. "Causal Decision Theory is Safe From Psychopaths." In *Erkenntnis*, **86**: 665–685. [3]