Indecisive Decision Theory

Brian Weatherson

2021-10-14

1 Decisiveness

Say a decision theory is **decisive** iff for any decision problem, it says either:

Picture is "Study for 'The Bear Hunt' (for the Alcázar, Madrid)" by Peter Paul Rubens via Cleveland Museum of Art.

1. There is a uniquely best choice, and rationality requires choosing it.; or
2. There is a non-singleton set of choices each of which is tied for being best, and each of which can be permissibly chosen.

A decision theory is **decisive over binary choices** iff it satisfies this condition for all decision problems where there are just two choices. Most decision theories in the literature are decisive, and of those that are not, most of them are at least decisive over binary choices. I'm going to argue that the correct decision theory, whatever it is, is indecisive. It is not, I'll argue, even decisive over binary choices.

The argument turns on a pair of very similar decision problems. Each problem has the following structure. There is a human Player, and a predictor, who I'll call Doctor. Doctor is very good, as good as the demon in Newcomb's problem, at predicting Player's behavior. Doctor will make two decisions. First, they will opt-in (which I'll write as I for In), or opt-out (which I'll write as O for out). If they opt-out, Doctor gets $1, and Player gets $100. (Assume both Doctor and Player prefer more money to less, and indeed that over these small sums there is more or less no declining marginal utility of money.) If they opt-in, this will be publicly announced, an another game will be played. Each of Doctor and Player will (independently) pick a letter: A or B. Doctor will aim to predict Player's choice, and will be rewarded iff that prediction is correct. Here is the payout table for the possible outcomes of this game.

Human Pick	Doctor Pick	Human Reward	Doctor Reward
A	A	$6	$4
A	B	$0	$0
B	A	$3	$0
B	B	$4	$1

If you prefer this in the way we standardly present games in normal form, it looks like this, with the human as Row, doctor as Column, and in each cell the human's payout is listed first. (All payouts are in dollars)

	A	**B**
A	6,4	0,0
B	3,0	4,1

Doctor is good at predictions, and prefers more money to less. So if Doctor predicts Player will choose A, they will opt-in, and also play A, getting $4. But what if they predict Player will choose B. They will get $1 either by opting-out, or by opting-in and choosing B. Since they are indifferent in this case, let's say they will flip a coin to decide which way to go. And Player knows that this is how Doctor will decide, should Doctor predict Player will choose B. Doctor will also flip a coin to decide what prediction to make if they think Player is completely indifferent between the choices, and Player also knows this.

Now I said there were going to be two problems. Here's how they are created. two Players, Amsterdam and Brussels, will play the game. They have identical utility functions over money, and identical prior probability distributions about what Doctor will do conditional on each of their choices. That's to say, they both think the probability that Doctor will be wrong is vanishingly small. But Brussels is busy and has to run to the bank, so they have to write their decision in an envelope that will be revealed, after Doctor chooses A or B, iff Doctor opts-in in the game with Brussels. Amsterdam, on the other hand, gets to see Doctor's decision about whether to opt-in or opt-out, and then (if Doctor opts-in) has to write their decision in an envelope that will be revealed after Doctor chooses A or B. For ease of reference, call Brussels's decision the *early* decision and Amsterdam's decision the *late* decision.

Here is the core philosophical premise in the argument to follow.

The Core Premise If there is is precisely one permissible choice for Amsterdam, and one permissible choice for Brussels, then it must be the same choice. That is, if each of them is obliged to choose a particular letter, it must be the same letter. It can't be that one is obliged to choose A, and the other obliged to choose B.

I'm calling **The Core Premise** a premise, though in the next section I'll offer a few arguments for it, in case you don't think it is obviously correct. (I sort of think it is obviously correct, but I'm really not going to rely on you sharing that intuition.) And I'll argue that any decisive theory (that meets minimal coherence standards) has to violate this constraint. Some decisive theories say Amsterdam should choose A and Brussels should choose B, but a few say the reverse. And a few (otherwise implausible) decisive theories say that Amsterdam and Brussels should do the same thing in this game, but different things in games with the same structure but slightly different payouts. In every case, a decisive theory will make some incoherent pair of recommendations, and so is mistaken.

The Core Premise is a conditional, but any decisive theory that denies that the choices are tied will meet the condition. So from now on in arguing against decisive theories I'll mostly just interpret **The Core Premise** as saying Amsterdam and Brussels must make the same choice. The main thing to check for is that a theory doesn't say the choices are tied. None of the theories I'll look at will say that, but it's an important thing to check.

Before we start there are four pieces of important housekeeping.

First, the definition of decisiveness referred to options being tied. For the definition to be interesting, it can't just be that options are tied if each is rationally permissible. Then a decisive theory would just be one that either says one option is mandatory or many options are permissible. To solve this problem, I'll borrow a technique from Ruth Chang (2002). Some options are **tied** iff either is permissible, but this permissibility is sensitive to sweetening. That is, if options X and Y are tied, then for any positive ε, the agent prefers $X+\varepsilon$ to Y. If either choice is permissible even if X is 'sweetened,' i.e.., replaced in the list of choices by $X+\varepsilon$, we'll say they aren't tied. My thesis then is that the correct decision theory says that sometimes there are multiple permissible options, and each of them would still be permissible if one of them was sweetened.

Second, there is an important term in the definition of decisiveness that I haven't clarified: **decision problem**. Informally, the argument assumes that in setting out the proble facing Amsterdam and Brussels is indeed a decision problem. More formally, I'm assuming it suffices to specify a decision problem to describe the following four values.

- What choices Player has;
- What possible states of the world there are (where it is understood that the choices of Player make no causal impact on which state is actual)[1];
- What the probability is of being in any state conditional on making each choice; and

[1]My personal preference is to understand states historically. For any proposition relevant to the decision, a state determines its truth value if it is about the past, or its chance at the start of deliberation if it is about the future. And then causal independence comes in from a separate presupposition that there is no backwards causation. But I definitely won't assume this picture of states here.

- What return Player gets for each choice-state pair.

Most recent papers on decision theory do not precisely specify what they count as a decision problem, but they seem to implicitly share this assumption, since they will often describe a vignette that settles nothing beyond these four things as a decision problem. And that's what I did as well! You should understand this as being part of the definition of decisiveness. This implies that there are two ways to reject decisiveness.

First, a theory could say that these four conditions underspecify a real decision problem. In any real situation, decision theory has a decisive verdict, but it rests on information, typically information about Player, not settled by these four values. I'll say a theory that goes this route is *intrapersonally* decisive, but not *interpersonally* decisive.

Second, a theory could say that no matter how much one adds to the specification, there will be cases where the correct decision theory does not issue a verdict. Such a theory is not *intrapersonally* decisive. There is nothing you could add about the person to the specification of a decision problem which decides what they should do, or even which options are tied. I want to ultimately defend such a view, and this paper is a part of the defence. But it's a proper part. Nothing I say here rules out mere interpersonal indecisiveness. That's an argument for another day. Today, we have enough to be getting on with.

Third, I have set up this problem quite explicitly as a game, with another player - Doctor. I don't think this is particularly big deal, though I gather not everyone agrees. In this respect (among others) I'm following William Harper (1986), who recommended treating Newcomb's Problem as a game. It's not clear why it wouldn't be a game. The Newcomb demon makes a choice, and if you assume the demon gets utility 1 from correct predictions and utility 0 from incorrect predictions, they make the choice that maximises their return given their beliefs about what the other (human) player will do. So game theoretic techniques can and should apply. I think any problem with a predictive demon is best thought of as a game where the demonic player gets utility 1 or 0 depending on whether their prediction is right. Here the predictive player, Doctor, has a utility function with slightly more structure. But if we think decision theory should apply in cases where there is a predictor around, it should still apply when that predictor has preferences with slightly more structure.

Fourth, this project was inspired by reading David Pearce's argument against 'Single Solution Concepts' (Pearce 1983). My initial plan was to simply translate his argument into decision theoretic terms and use it as an argument for indecisiveness. I ended up with a somewhat different argument to his, one that draws heavily on Brian Skyrms's work on Stag Hunts (Skyrms 1990, 2001, 2004). But the project started out from an idea of Pearce's. And as you'll see in the conclusion, it will end with a related idea of is.

2 Defending The Core Premise

In this section I'll offer three arguments in defence of **The Core Premise**. The first will be to argue that Amsterdam and Brussels have in a key sense the same choice, the second will argue that violations of **The Core Premise** will violate the Sure Thing Principle, and the third is that violations of **The Core Premise** lead to people being willing to pay to avoid information. The arguments will make frequent use of the following equivalence. A player must choose a letter iff the player prefers choosing that letter. So I'll move freely from saying that a theory says Amsterdam should choose X to saying Amsterdam should prefer X. I don't think this should be controversial, but it's worth noting. Onto the three arguments.

Think about what Brussels is doing when writing in the envelope. They know that the envelope will only be opened if Doctor opts-in. If Doctor does, then what they play will determine their payout. So they should imagine that Doctor has opted-in, and act accordingly. But in that imaginative situation, they will do the same thing as if they knew that Doctor had opted-in. That is, they'll do the same thing Amsterdam will do. There isn't any difference, for purposes of choice, between supposing that Doctor has opted-in, and learning that Doctor has opted-in. And Brussels should suppose that Doctor has opted-in. After all, they are being asked what to contribute to the game if, and only if, Doctor opts-in. So the two choices are effectively the same, and they should get the same verdict. That's the first argument.

Assume that a theory says Amsterdam should do X, but Brussels should do Y, where X and Y are distinct. Now ask the theory, what should Brussels prefer conditional on Doctor opting-in, and conditional on Doctor opting-out. Since Amsterdam should choose X, conditional on Doctor opting-in, Brussels thinks X is better than Y. And conditional on Doctor opting-out, Brussels is indifferent between X and Y, so thinks X is as at least as good as Y. The sure thing principle (or at least the version that matters here) says that if Brussels knows that precisely one of a set of outcomes obtains, and X is at least as good as Y conditional on each member of the set, then X is at least as good as Y overall. But that contradicts the assumption that Brussels should choose Y. So that's the second argument.

Assume again that a theory says Amsterdam should do X, but Brussels should do Y, where X and Y are distinct. Now imagine a third player, Cardiff. Cardiff isn't busy, like Brussels. But Cardiff hasn't yet found out whether Doctor has opted-in. They are offered the chance to buy ear plugs, so they won't hear the announcement of whether Doctor opts-in or opts-out. They should like to get those, since right now they prefer Y to X, but there's a chance that they'll hear Doctor has opted-in, which will leave them in the same situation as Amsterdam, and hence they'll choose X. And there is nothing they can gain from hearing the announcement. But this is absurd - a player should not pay to avoid relevant information about the game. So that's the third argument.

Now this last argument has one caveat. I didn't calculate how much Player should pay for the ear plugs. That turns out to vary a little depending on just which decisive theory we are looking at. A theory may say that Cardiff should not pay anything for the ear plugs, since they are certain Doctor will opt-out. This third argument isn't particularly effective, I think, against those theories. But it works well, and I feel is the strongest argument, against some other theories. But we'll have to look case by case at just how much a theory would recommend Cardiff pay for the ear plugs.

So that's the defence of **The Core Premise**. What I'll now show is that a wide range of decisive theories violate it, and so we can conclude they are false.

3 Early and Late Choices

To see why theories might violate **The Core Premise**, it's helpful to set out explicitly the choices that Amsterdam and Brussels face. And we'll treat Doctor largely as a non-player character, just as the demon is typically treated in Newcomb's problem. So from now on the columns will not be Doctor's choices, but what Doctor predicts the human player chooses. And we'll assume Doctor maximises their financial return given a correct prediction. It's easy to set out the choice Amsterdam faces; it's just the embedded game with some notational differences.

	PA	**PB**
A	6	0
B	3	4

I've written **PA** and **PB** in the columns to indicated that A or B is Predicted. But in this game that makes little difference, since Doctor will do whatever they predict Amsterdam will do. Things are a little different for Brussels. If Doctor predicts that Brussels has written B, they will flip a coin to decide whether to opt-out, or opt-in. So we can't write Brussels's return in actual dollars, since we don't know how the coin lands. But we can write the return in expected dollars, and we assume that Brussels is after all trying to maximise expected dollars. (We'll come back to this assumption in the next section.) So the table Brussels faces looks like this.

	PA	**PB**
A	6	50
B	3	52

If Doctor predicts B, then Player has a 1 in 2 chance of getting $100, and a 1 in 2 chance of getting the payout from the previous game. So their average payout is $50 if they play A, and $52 if they play B. Hence the values in the right hand column here.

So what **The Core Premise** says is that if each of these games has a uniquely rational choice, it must be the same choice. As we'll see, a lot of theories do not satisfy this constraint.

4 Evidential Decision Theory

Given a perfect predictor, Evidential Decision Theory says that the only payout values that matter are those in the main diagonal, running from northwest to southeast. So Amsterdam should choose A, since they'll expect to get $6 from A and $4 from B. But Brussels should choose B, since they'll expect to get $6 from A and $52 from B. So Evidential Decision Theory violates **The Core Premise**, and hence is mistaken.

When I introduced Cardiff's case, I said we had to check what a particular theory said about what Cardiff would pay for the earplugs. So let's do that for Evidential Decision Theory. If Cardiff thinks they would be told that Doctor has opted-in, they would pay up to $46 to avoid that information, since they think they will get $52 without the information and $6 with it. But maybe they would be told that Doctor opted-out. It turns out the assumption they would be told that is incoherent, and Cardiff knows it. If they are told Doctor has opted-out, Doctor will know they are indifferent between the options. And in that case, Doctor will flip a coin to decide what 'prediction' to make. But if Doctor thinks it is 50/50 what Cardiff will do, they have an expected return of $2 from opting-in and choosing A, but an expected return of $1 from opting-out. So they will opt-in and choose A, contradicting the assumption that Cardiff will be told they opted-out. And Cardiff can do all this reasoning. So Cardiff can predict that if they are told anything, it will be that Doctor has opted-in, putting them in the same position as Amsterdam. But if they are told nothing, they are in the same position as Brussels. And they prefer, by $46, being in the same position as Brussels.

Now this is not really a new objection to Evidential Decision Theory. You can find similar points being made about the strange behaviour of Evidential Decision Theory in dynamic choice settings as far back as Gibbard and Harper (1978). The details of my argument are a bit different, but ultimately they rest on the same foundations. I think those are perfectly solid foundations, but given how long the arguments have been around, clearly not everyone agrees. So I want to note one internal tension within Evidential Decision Theory this case brings up.

As Edward Elliott (2019) notes, within contemporary decision theory there is little overlap between work on what to do when a predictor is around, and work on the nature of risk. All parties to the former dispute take for granted the orthodox view that when there is no

predictor, one should maximise expected utility. But that's very controversial within the debates about risk. There the big question is whether the heterodox risk-weighed utility theory developed by Quiggin (1982) and Buchak (2013) is preferable to orthodoxy.

The Quiggin/Buchak view raises a dilemma for Evidential Decision Theory. If they reject the view and stick with orthodoxy, as most do, they should have an argument against the risk-weighted view. But the strongest such arguments turn on the fact that the risk-weighted view violates the Sure Thing Principle, and leads to people paying to avoid information. Evidential Decision Theorists can't complain about it on those grounds, since their theory does the same thing. Alternatively, they can modify their theory to incorporate the Quiggin/Buchak view. But then they wouldn't have a decisive decision theory, since on that view what to do in a decision problem depends on something not typically specified in the problem, namely the chooser's attitude towards risk. So even if the Evidential Decision Theorist rejects the arguments behind **The Core Premise**, as I suspect most will, they need to either find a new objection to risk-weighted theories of choice, or modify their theory in a way that abandons decisiveness. I think the arguments for **The Core Premise** are sound, but even if they aren't, it seems unlikely that there is a plausible *decisive* theory that can be derived from Evidential Decision Theory.

5 Stag Hunt

It's possible to transform any decision problem involving a predictor into a game. David Lewis (1979) already noted the relationship between Prisoners' Dilemma and Newcomb's Problem. And William Harper (1986) noted that you could turn any problem involving a predictor into a game by assuming the predictor wants to make correct predictions and acts in their own interest.

It's also frequently possible to do the reverse transformation, to turn a game into a decision problem involving a predictor. Start with any one-shot two person game, where each player has the same number of choices in front of them. Then change the payout for Column so that they get 1 if Row and Column make the 'same' choice (for some mapping between Row's and Column's choices), and 0 otherwise. Then just treat the Column player as a known to be accurate predictor, either an agent making predictive choices or a state of the world that tracks the agent's choices in some way. Now Row's choice is just a familiar kind of decision problem.

If you plug various famous games into the recipe from the previous paragraph, you get some familiar examples from modern decision theory. If you start with Prisoners' Dilemma and apply this recipe, you get Newcomb's Problem. If you start with Matching Pennies[2] , you

[2]You can see examples of all these games, and all the game theoretic machinery I use throughout this paper,

get Death in Damascus (Gibbard and Harper 1978). If you start with Battle of the Sexes, you get Asymmetric Death in Damascus (Richter 1984). If you start with Chicken, you get the Psychopath Button (Egan 2007). But there hasn't been quite as much attention paid to what happens if you start with Stag Hunt and run this recipe. The game you get turns out to be very useful for classifying decisive decision theories that choose two boxes in Newcomb's Problem.

Here is an abstract form of a Stag Hunt game, where the options are G/g for Gather or H/h for Hunt. Actually, this is a table for a generic symmetric game; what makes it a Stag Hunt are the four constraints listed below.[3]

	g	**h**
G	x, x	y, z
H	z, y	w, w

- $x > z$
- $w > y$
- $w > x$
- $x + y > z + w$

The first two constraints imply that $\langle G, g \rangle$ and $\langle H, h \rangle$ are both equilibria. This isn't like Prisoners' Dilemma, that only has one equilibrium. But it is like Prisoners' Dilemma in that there is a cooperative solution, in this case $\langle H, h \rangle$, but it isn't always easy to get to it. It isn't easy because there are at least two kinds of reasons to play G.

First, one might play G because one wants to minimise regret. Each play is a guess that the other player will do the same thing. If one plays G and guesses wrong, one loses $w - y$ compared to what one could have received. If one plays H and guesses wrong, one loses $x - z$. And the last constraint entails that $x - z > w - y$. So playing G minimises possible regret.

Second, one might want to maximise expected utility, given uncertainty about what the other player will do. Since one has no reason to think the other player will prefer g to h or vice versa - both are equilibria - maybe one should give each of them equal probability. And then it will turn out that G is the option with highest expected utility. Intuitively, H is

in any standard game theory textbook. My favorite such textbook is Bonanno (2018), which has the two advantages of being philosophically sophisticated and open access. I'm not going to include citations for every bit of textbook game theory I use; that seems about as appropriate as citing an undergrad logic textbook every time I use logic. But if you want more details on anything unfamiliar in this paper, that's where to look.

[3] I've listed the constraints as strict inequalities, but that might be over the top. Sometimes you'll see one or other of these constraints weakened to an inclusive inequality. This difference won't matter for current purposes.

a risky option and *G* is a safe option, and when in doubt, perhaps one should go for the safe option.

What I'll call a *Stag Decision* is basically a Stag Hunt game where the other player is a predictor. So the decision looks like this, where the above four constraints on the values still hold, and **PX** means the predictor predicts **X** will be chosen.

	PG	**PH**
G	*x*	*y*
H	*z*	*w*

These kinds of decisions are important in the history of game theory because they illustrate in the one game the two most prominent theories of equilibrium selection: risk dominance and payoff dominance (Harsanyi and Selten 1988). Risk dominance recommends gathering; payoff dominance recommends hunting. And most contemporary proponents of decisive decision theories in philosophy fall into one of these two camps.

In principle, there are three different views that a decisive theory could have about Stag Decisions: always Hunt, always Gather, or sometimes do one and sometimes the other. A decisive theory has to give a particular recommendation on any given Stag Decision, but it could say that the four constraints don't settle what that decision should be. Still, in practice all existing decisive theories fall into one or other of the first two categories.

One approach, endorsed for rather different reasons by Richard Jeffrey (1983) and Frank Arntzenius (2008), says to hunt because it says in decisions with multiple equilibria, one should choose the equilibria with the best payout. This approach will end up agreeing with everything the Evidential Decision Theorist says about the choices facing Amsterdam and Brussels, and should be rejected for the same reason. It treats differently choices that are fundamentally the same, it violates Sure Thing, and it says Cardiff should pay $46 to avoid finding out what Doctor selected. And the same will be true for any decisive theory that says to always Hunt in Stag Decisions.

Another family of approaches says to always Gather in Stag Decisions. For very different reasons, this kind of view is endorsed by Ralph Wedgwood (2013), Dmitri Gallow (2020) and Abelard Podgorski (forthcoming). These three views differ from each other in how they motivate Gathering, and in how they extend the view to other choices, but they all agree that one should Gather in any Stag Decision. And this leads to the reverse problem to that facing the always Hunt view.

Both Amsterdam and Brussels are facing Stag Decisions. But for Amsterdam, choosing A is Hunting and choosing B is Gathering, while for Brussels, choosing A is Gathering and choos-

ing B is Hunting. So any view which says to always Gather will say that Amsterdam should choose B, and Brussels should choose A. Again, this treats differently choices that are fundamentally the same, and violates Sure Thing. But does it mean Cardiff will pay to avoid information? Here things are a little trickier, because Cardiff has four possible choices: Receive information or pay to decline it, and then choose A or B. And the different approaches to Gathering say different things about how to make decisions in four-way choices. So let's set that argument for **The Core Premise** aside - the first two arguments for it still seem like decisive objections to any view that one should always Gather.

What about views that deny that all Stag Decisions should be treated alike? As I've said, I don't think any such view is in the literature, but it's good to think about other views. Let's drop the assumption that we're even looking at a Stag Decision (though it will turn out that we are), and think about what to do in general in cases where there are two strict equilibria. That is, think about what our imaginary decisive decision theory will say about the following case, where we just have the constraints $x > z$ and $w > y$, and again *PX* means the predictor predicts *X*.

	PE	**PF**
E	x	y
F	z	w

Any coherent solution must be invariant under redescriptions of the problem. So if you take a real world example that fits this category, and relabel which option is E and which is F, the recommendation should flip. And if you rescale the utilities by multiplying by a positive constant or adding a constant, the verdict should be unchanged, since utilities are only defined up to positive affine transformation. The only theories that meet these constraints say that a choice has a 'score' $x + my$, where x is the equilibrium payoff, and y is the other possible payoff, and m is a free variable the theory sets which reflects how much it cares about the value of the non-equilibrium payoff. The theory then says to pick the option with the higher score, or to be indifferent otherwise. So it says to strictly prefer E to F iff $x + my > w + mz$ and to be indifferent between the choices if that's an equality not an inequality. Setting m to 0 gives you the view that says one should always Hunt, since one should always pick the equilibrium with the highest equilibrium value. Setting m to 1 gives you the view that you should always Gather, since you should maximise the sum (or, equivalently, the average) of the two payouts you might get with the choice. And both of these views violate **The Core Premise**. But what should we say about views that give m other values?

The first thing to say is that it is very hard to see any good philosophical motivation for values of m other than 0 or 1. Both these values make a certain amount of sense, but the

reasons behind any other value are harder to understand. Still, if coherence required some other value for m, I'm sure someone would come up with a motivation.

The second thing to say is that we have done more already than object just to the theories that set m to 1 or 0. Any theory that has $m < \frac{2}{3}$ will say that Amsterdam should choose A and Brussels should choose B, violating **The Core Premise**. And any view that has $m > \frac{46}{47}$ will say that Amsterdam should choose B and Brussels should choose A, also violating **The Core Premise**. But we don't yet have an objection to theories on which $\frac{2}{3} \leq m \leq \frac{46}{47}$.

To see what's wrong with those theories, keep the structure of the game the same, but change the rewards as follows (all rewards are in dollars).

Human Pick	Doctor Pick	Human Reward	Doctor Reward
None	Opt-out	0	1
A	A	4	4
A	B	0	0
B	A	$2+\frac{1}{m}$	0
B	B	2	1

Then the 'late game' that Amsterdam faces will look like this:

	PA	**PB**
A	4	0
B	$2 + 1/m$	2

Since $2 + m(2 + \frac{1}{m}) > 4 + 0m$ for any value of m satisfying $\frac{2}{3} \leq m \leq \frac{46}{47}$, the theory will say Amsterdam should choose B.[4]

The 'early game' that Brussels faces will look like this:

	PA	**PB**
A	4	0
B	$2 + 1/m$	1

[4]More slowly, we can use the formula to work out the score of each option. The score of A is the value in the top-left, 4, plus m times the value in the top-right, 0. And that's 4, no matter the value of m. The score of B is the value in the bottom-left, 2, plus m times the value in the bottom-left, $2 + \frac{1}{m}$. That is, the score is $2 + (1 + 2m) = 3 + 2m$. Since $m > \frac{1}{2}$, this value is greater than 4, which was the score of A.

Since in this game Player gets nothing if Doctor opts-out, and there is a 50/50 chance the Doctor will opt-out if they predict B, the returns in the right-hand column are half what they are in the late game. Since $4 + 0m > 1 + m(2 + \frac{1}{m})$ for any value of m satisfying $\frac{2}{3} \leq m \leq \frac{46}{47}$, the theory will say Brussels should choose A.[5]

So any decisive theory will violate **The Core Premise** for some choice pair or other. Hence all decisive theories are mistaken.

5.1 Where To Next?

Decision theory cannot be everything that some of its proponents want it to be. It cannot be a guide that tells us what to do in every situation, even if we allow it to sometimes say that options are tied. So what can decision theory be? A natural answer is that it can tell us which options are rationally permissible, knowing that there will often be a plurality of options that are permissible. I think the way to finding a plausible indecisive theory goes via answering the following three questions.

First, does decision theory start with what the chooser believes, or with what they should believe? If Player is certain that the red box has more money, but they have conclusive evidence that the blue box has more money, which box does decision theory say that they should choose? If decision theory is the theory of which actions "most effectively serve one's desires according to one's beliefs" (Lewis 2020c, 465), then it is the red box. If it is the theory of which choices are rational, then it is the blue box. I'm sympathetic to the arguments that Nomy Arpaly (2002) makes that the theory of rational choice should not pay any special attention to the agent's beliefs. What's rational to choose in a situation is a function of what's rational to believe in that situation, not what one actually believes.[6]

Second, in a given situation, how many different beliefs are rational? The Uniqueness thesis says the answer is one. Permissivism says that Uniquenes is false, and for some propositions in some situations, there are multiple rational attitudes to have. See Kopec and Titelbaum (2016) for a good survey of the issues, Schultheis (2018) for a recent argument for Uniqueness, and Callahan (2021) for a recent argument for Permissivism.[7] I'm on the Permissivist

[5] More slowly, we can use the formula to work out the score of each option. The score of A is the value in the top-left, 4, plus m times the value in the top-right, 0. And that's 4, no matter the value of m. The score of B is the value in the bottom-left, 1, plus m times the value in the bottom-left, $2 + \frac{1}{m}$. That is, the score is $1 + (1 + 2m) = 2 + 2m$. Since $m < 1$, this value is less than 4, which was the score of A.

[6] See also Lewis (2020a) where Lewis sketches a view that would say that each choice is rational in a way, and there need not be anything more to say about which is rational all-things-considered. I take it he means decision theory is the theory of the part of rationality that the red box chooser does well on. A similar point is suggested in Lewis (2020d).

[7] Interestingly, Callahan connects Permissivism to existentialism. I suspect there are deep and unexplored connections between existentialism and decision theory, especially concerning the questions about the prior-

side of this debate.

Now if you think decision theory should be sensitive to rational beliefs rather than actual beliefs, and you think Permissivism is true, you're committed to indecisiveness. You won't even need demons. After all, any situation where any credence in p between x and y is permissible will mean there are multiple bets at distinct odds on p that rationality neither requires taking nor requires passing. I think this is a perfectly sound argument for indecisiveness, but I didn't lean on it here because the premises are considerably less secure than the ones I've appealed to.

But there is a third question that needs answering before we can offer a plausible indecisive theory: what is a mixed strategy? Relatedly, what role do mixed strategies have in the correct decision theory? This is a rather vexed question, and an important one. Almost all recent arguments against causal decision theory seem, to my eyes at least, to turn on attributing a bad theory of mixed strategies to the causal decision theorist. You can see this from the fact that almost all recent papers on decision theory involve problems that, when converted into games, have no pure strategy equilibria, but do have mixed strategy equilibria. We can't offer a full decision theory, even an indecisive one, without resolving these problems, and that means having a theory of mixed strategies. And that's very much a theory for another paper.[8]

Note one thing I haven't said so far, and won't say in what follows. I don't say that the way to find the correct indecisive theory is to come up with a bunch of cases, consult our intuitions about them, and then see which theory can match at least 80% of those intuitions. (Or whatever percentage we are working with this week.) That is a dubious approach in general, but around here it is close to incoherent.

Most contemporary work in decision theory starts with the assumption that when there are no demons around (or anything else vaguely demonic), expected utility maximisation is the correct decision theory. And then theorists will start rolling out fantastic cases involving demons or predictors or lesions or genes or twins or triplets or whatever is in fashion. And they will ask what extension of expected utility theory best tracks intuitions about these cases. But this seems like a very dubious strategy, since intuitions about cases will not lead one to expected utility theory in the first place. Trying to match intuitions about cases like the Allais or Ellsberg paradoxes will lead one to prefer some non-standard theory like the one developed by John Quiggin (1982) or Lara Buchak (2013). It seems very unlikely that the best way to extend a counterintuitive theory like expected utility maximisation is by

ity of strategies or individual choices. But that's for another paper.

[8]For what it's worth, I think that theory must include the following two factors. First, playing a mixed strategy is just what Lewis (2020b) calls using a tie-breaking procedure. Second, the output of such a tie-breaking procedure is in principle unpredictable by anything that doesn't time travel.

consulting intuitions about puzzle cases. It is much better to ask what principles we want our theory to endorse, and work towards a theory that satisfies those principles. And that is the methodology I have adopted here.

I've relied heavily in this paper on two such principles: The Sure Thing principle, and the principle that information has non-negative value. I'll end by describing one more principle, and noting two questions it raises. The principle is that a decider should be a probabilist, and that they should maximise expected utility. More precisely, it says that if the states are $\{S_1, \dots, S_m\}$, and the choices are $\{O_1, \dots, O_n\}$, then O_i is a permissible choice just in case there is some probability function *Pr* such that

$$\sum_{k=1}^{m} V(S_k \wedge O_i) Pr(S_k) \geq \sum_{k=1}^{m} V(S_k \wedge O_j) Pr(S_k)$$

for all $j \in \{1, \dots, n\}$. Even if the subjective probability of the state is affected by the choice one makes, there should be some probability function that the chooser ends up with, and their choice should make sense by the lights of that probability function. Note that if we assume that the chooser can select any mixed strategy from among their choices, there is guaranteed to be at least one strategy that satisfies this requirement, even if one thinks the states are choices of a demon who can predict one's strategy.[9]

So that seems to me like a minimal constraint on choices. As Pearce (1984) shows, it is equivalent to the requirement that one not make a choice that is strictly dominated by some other choice, or by some mixture of other choices. (This result is hardly obvious, but it turns out to be a reasonably straightforward consequence of the existence of Nash equilibria for all finite zero-sum games.) That's hardly an uncontroversial principle, but it is also one I'm happy to adopt. If you're still on board, there are two more questions that we need to answer before we finish our decision theory.

Are all further constraints on rational decisions representable as constraints on the *Pr* in this principle? There surely are some further constraints on rational decisions. If you're offered a bet at even money on whether I will become Canadian President next week, the only rational thing to do is to decline it. And that's true even though there is a *Pr* such that taking the bet maximises expected utility. But that *Pr* is completely irrational given your evidence. So *Do something that maximises expected utility given some probability* is too liberal a rule; we need to say something about the *Pr*. Do we need to say more than that? My answer is no, though I'm not even going to start defending that here.[10]

[9] If the demon can predict what one will do on a given occasion while playing a mixed strategy, this guarantee may fail. But assuming what I said in the last footnote about mixed strategies, that would mean we're in the realm of backwards causation, and the states are not causally independent of the actions.

[10] Note that if you say no to this question, and you think that probabilities have to be real-valued, then you're

The Canadian Presidency examples suggests that there are constraints on *Pr* that are external to decision theory. You shouldn't take that bet because you shouldn't have probability above 0.5 that I'll become Canadian President next week. The order of explanation runs from the (ir)rationality of the credal state to the (ir)rationality of the decision. Our fifth and final question is, are there any cases where the order of explanation goes the other way? Arntzenius (2008) argued that one should have credences such that the highest value equilibrium was also the choice that maximised expected utility. That's an example of a constraint on *Pr* where the order of explanation runs from decisions to beliefs. I argued against that principle, but not because of a systematic reason to think that the order of explanation can't run that way. Instead I argued that this particular principle was dynamically incoherent. That leaves open the general question of whether any such principles, where constraints on decisions explain constraints on belief, are right.

The long term goal of the project behind this paper is to argue that there are no such principles. The only constraints on rational decision are that one should maximise expected utility given some *Pr*, and this *Pr* should satisfy independently motivated epistemic requirements. Now I haven't come close to arguing for that here, and it's a very strong claim. Given everything else I've said, it basically amounts to the claim that the theory of equilibrium selection has no role to play in normative decision theory. It may have a central role to play in descriptive decision theory, in explaining why people end up at a certain equilibrium. But it can't justify that equilibrium, since any equilibrium could be rationally justified.[11]

But all of this is for future work. The aim of this paper has been to open up the possibility of an indecisive, i.e., permissive, decision theory. Decisive decision theories have to take a stand on Stag Decisions, and there is no coherent way for them to do that. So no decisive theory is correct, and the correct decision theory is indecisive.

References

Arntzenius, Frank. 2008. "No Regrets; or, Edith Piaf Revamps Decision Theory." *Erkenntnis* 68 (2): 277–97. https://doi.org/10.1007/s10670-007-9084-8.

Arpaly, Nomy. 2002. "Moral Worth." *Journal of Philosophy* 99 (5): 223–45. https://doi.org/10.2307/3655647.

Bonanno, Giacomo. 2018. "Game Theory." Davis, CA: CreateSpace Independent Publishing Platform. 2018. http://faculty.econ.ucdavis.edu/faculty/bonanno/GT_Book.html.

committed to weak dominance not having a role to play in decision theory. So this is a non-trivial question.

[11]But note here that what I'm calling an equilibrium is just a coherent set of beliefs that is grounded in the evidence. It doesn't include the requirement, typical in game-theory, that the chooser has true beliefs about some aspect of the world around them.

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

Callahan, Laura Frances. 2021. "Epistemic Existentialism." *Episteme*. https://doi.org/10.1017/epi.2019.25.

Chang, Ruth. 2002. "The Possibility of Parity." *Ethics* 112 (4): 659–88. https://doi.org/10.1086/339673.

Egan, Andy. 2007. "Some Counterexamples to Causal Decision Theory." *Philosophical Review* 116 (1): 93–114. https://doi.org/10.1215/00318108-2006-023.

Elliott, Edward. 2019. "Normative Decision Theory." *Analysis* 79 (4): 755–72. https://doi.org/10.1093/analys/anz059.

Gallow, J. Dmitri. 2020. "The Causal Decision Theorist's Gudie to Managing the News." *The Journal of Philosophy* 117 (3): 117–49.

Gibbard, Allan, and William Harper. 1978. "Counterfactuals and Two Kinds of Expected Utility." In *Foundations and Applications of Decision Theory*, edited by C. A. Hooker, J. J. Leach, and E. F. McClennen, 125–62. Dordrecht: Reidel.

Harper, William. 1986. "Mixed Strategies and Ratifiability in Causal Decision Theory." *Erkenntnis* 24 (1): 25–36. https://doi.org/10.1007/BF00183199.

Harsanyi, John C., and Reinhard Selten. 1988. *A General Theory of Equilibrium Selection in Games*. Cambridge, MA: MIT Press.

Jeffrey, Richard. 1983. "Bayesianism with a Human Face." In *Testing Scientific Theories*, edited by J. Earman (ed.). Minneapolis: University of Minnesota Press.

Kopec, Matthew, and Michael G. Titelbaum. 2016. "The Uniqueness Thesis." *Philosophy Compass* 11 (4): 189–200. https://doi.org/10.1111/phc3.12318.

Lewis, David. 1979. "Prisoners' Dilemma Is a Newcomb Problem." *Philosophy and Public Affairs* 8 (3): 235–40.

———. 2020a. "Letter to D. H. Mellor, 14 October 1981." In *Philosophical Letters of David K. Lewis*, edited by Helen Beebee and A. R. J. Fisher, 2:432–34. Oxford: Oxford University Press.

———. 2020b. "Letter to Gregory Kavka, 10 July 1979." In *Philosophical Letters of David K. Lewis*, edited by Helen Beebee and A. R. J. Fisher, 2:423–24. Oxford: Oxford University Press.

———. 2020c. "Letter to Huw Price, 17 May 1988." In *Philosophical Letters of David k. Lewis*, edited by Helen Beebee and A. R. J. Fisher, 2:464–66. Oxford: Oxford University Press.

———. 2020d. “Letter to William j. Talbott, 22 June 1984.” In *Philosophical Letters of David k. Lewis*, edited by Helen Beebee and A. R. J. Fisher, 2:448–49. Oxford: Oxford University Press.

Pearce, David G. 1983. “A Problem with Single Valued Solution Concepts.” 1983. https://sites.google.com/a/nyu.edu/davidpearce/.

———. 1984. “Rationalizable Strategic Behavior and the Problem of Perfection.” *Econometrica* 52 (4): 1029–50. https://doi.org/10.2307/1911197.

Podgorski, Aberlard. forthcoming. “Tournament Decision Theory.” *Noûs*, forthcoming. https://doi.org/10.1111/nous.12353.

Quiggin, John. 1982. “A Theory of Anticipated Utility.” *Journal of Economic Behavior & Organization* 3 (4): 323–43. https://doi.org/10.1016/0167-2681(82)90008-7.

Richter, Reed. 1984. “Rationality Revisited.” *Australasian Journal of Philosophy* 62 (4): 393–404. https://doi.org/10.1080/00048408412341601.

Schultheis, Ginger. 2018. “Living on the Edge: Against Epistemic Permissivism.” *Mind* 127 (507): 863–79. https://doi.org/10.1093/mind/fzw065.

Skyrms, Brian. 1990. *The Dynamics of Rational Deliberation*. Cambridge, MA: Harvard University Press.

———. 2001. “The Stag Hunt.” *Proceedings and Addresses of the American Philosophical Association* 75 (2): 31–41. https://doi.org/10.2307/3218711.

———. 2004. *The Stag Hunt and the Evolution of Social Structure*. Cambridge: Cambridge University Press.

Wedgwood, Ralph. 2013. “A Priori Bootstrapping.” In *The a Priori in Philosophy*, edited by Albert Casullo and Joshua C. Thurow, 225–46. Oxford: Oxford University Press.