Game Theory as Decision Theory

Brian Weatherson

2023-08-22

Table of contents

Preface

Draft for a book based on my (overly long) paper Gamified Decision Theory.

1 Introduction

Textbook versions of game theory embed a distinctive approach to decision theory. That theory isn't always made explicit, and it isn't always clear how it handles some cases. But we can extract one interesting and plausible theory, which I'll call Gamified Decision Theory (GDT), from these textbooks. There are nine characteristics of GDT (as I'll understand it) that I will focus on. I'll quickly list them here, then the bulk of the paper will consist of a section on each of the nine characteristics.

1. **Idealised**; GDT is a theory of what ideal deciders do.
2. **Expectationist**; the ideal decider prefers getting more expected value to getting less.
3. **Causal**; GDT is a variety of Causal Decision Theory (CDT).
4. **Allows Mixtures**; the ideal decider can perform a probabilistic mixture of any acts they can perform.
5. **Ratificationist**; the ideal decider endorses the decisions they make.
6. **Indecisive**; GDT sometimes says that multiple options are permissible, and they are not equally good.
7. **Dual Mandate**; in a dynamic choice, the ideal decider will follow a plan that's permissible, and take choices at every stage that are permissible.
8. **Substantive Probability**; the ideal decider has rational credences.
9. **Weak Dominance, Once**; the ideal decider will not choose weakly dominated options, but they may choose options that would not survive iterated deletion of weakly dominated strategies.

This is not going to be a work of exegesis, poring over game theory texts to show that they really do endorse all of 1-9. In fact it wouldn't take much work to show that they endorse 1-5, so the work wouldn't be worth doing. And while some books endorse 8 and 9, it would take a lot more investigative work than I'm going to do here to show that anything like a majority of them do. It would be interesting, but not obviously a philosophical question, to see what proportion endorse 6 and 7. But I'm going to set that aside.

What I do want to argue is that you can find some support for all of these in some game theory textbooks, and that combined they produce a plausible decision theory. While the textbooks don't all agree, for simplicity I'm going to focus on one book: Giacomo Bonanno's *Game Theory* (Bonanno 2018). This book has two important virtues: it is philosophically deep, and it is available for free. It isn't hard to find a game theory text with one or other of these virtues, but few have both. So it will be our primary guide in what follows, along with some primary sources (most of which are referenced in that book).

Methodologically, this paper differs from most works in decision theory in two ways. It has been a commonplace since Nozick (1969) to include demons, who are arbitrarily good at predicting a decision, in problems. Some of the cases here will involve two such demons, each of which is arbitrarily good at predicting a decision, and whose errors are probabilistically independent. Second, I'm going to rely less on intuitions about particular cases, and more on intuitions that certain cases should be treated the same way. This makes sense given the history of the field. There is much less consensus about what to do in Newcomb problems than about which problems are Newcomb problems. Judgments, or intuitions if you prefer, about how to classify problems seem more stable and more reliable, and they will be central to this paper.

The conclusion of this paper is that a permissive version of causal ratificationism is correct. Ideal choosers make choices that they do not immediately regret. With two small caveats, that's all there is to decision theory; any ratifiable choice is one an ideal agent might make. The first caveat is to with choices over time; the ideal chooser will make choices such that both their individual choices, and the set of choices they make, are ratifiable. This will exclude some possible choices that are individually ratifiable. Second, the ideal chooser will not choose weakly dominated options. These are fairly minor caveats; the resulting theory is not very different from other forms of permissive causal ratificationism, such as that defended by Melissa Fusco (n.d.).

2 Idealised

Consider the following decision problem. Chooser (our main protagonist) is going to be given a series of multiplication problems, where each of the multiplicands is a four digit number. Chooser doesn't have access to any kind of calculating device, and has no special arithmetic ability. For each question, if Chooser says the right answer, they get $2; if they pass, they get $1; if they say the wrong answer, they get nothing. Table 2.1 has the payout in table form.

Table 2.1: The multiplication game

	Best Guess Correct	**Best Guess Incorrect**
Guess	$2	$0
Pass	$1	$1

Every variety of decision theory defended in philosophy journals in recent years, and every game theory textbook, says that Chooser should simply say the correct answer. After all, Chooser should have probabilistically coherent credences, and every mathematical truth has probability 1, so whatever the correct answer is, saying it is a sure $2.

This is completely terrible advice. Chooser should pass every time, unless the question is something boring like 1000 times 2000. The chance of them getting a question like 5278 times 9713 correct are way less than one in two, so they are better off passing.

This doesn't mean that every philosophical decision theory, and every game theory textbook, is wrong. Those theories are not in the business of giving advice to people like Chooser. They are in the business of saying what it would be ideal, in some sense of ideal, for Chooser to do. And it would be ideal for Chooser to be a reliable computer, so if Chooser were reliable, they would always give the correct answer.

There is a big question about why we should care about what would have if Chooser were ideal. Chooser is not in fact ideal, so who cares what they would do if they were different. One might think that knowing what the ideal is gives Chooser something to aim for. Even if Chooser is not ideal, they can try to be closer to the ideal. The problem is that trying to be more like the ideal will make things worse. The ideal agent will never pass, and even if Chooser doesn't know the answers to the particular questions, they can know this fact. So if they try to be more like the ideal, they will never pass, and things will go badly.[1]

In philosophy we have two very different uses of the term 'idealisation'. One is the kind of idealisation we see in, for example, Ideal Observer theories in ethics. The other is the kind of idealisation we see in, for example, Ideal Gas models in chemistry. It's important to not confuse the two. Think about the volumeless, infinitely dense, molecules in an Ideal Gas model. To say that this is an idealised model is not to say that having volume, taking up space, is an imperfection. The point is not to tell molecules what the perfect size is. ("The only good molecule is a volumeless molecule.") Nor is it to tell them that they should approximate the ideal. ("Smaller the better, fellas.") It's to say that for some predictive and explanatory purposes, molecules behave no differently to how they would behave if they took up no space.[2]

The best way to understand decision theorists, and game theorists, is that they are using idealisations in this latter sense. The ideal choosers of decision theory are not like the Ideal Observers in ethics, but like the Ideal Gases. The point of the theory is to say how things go in a simplified version of the case, and then argue that this is useful for predictive and explanatory purposes because, at least some of the time, the simplifications don't make a difference.

One nice example of this working is George Akerlof's discussion of the used car market Akerlof (1970). In the twentieth century, it was common for lightly used cars to sell at a massive discount to new cars. There was no good explanation for this, and it was often put down to a brute preference for new cars. What Akerlof showed was that a model where (a) new cars varied substantially in quality, and (b)

[1] This is a special case of Lipsey and Lancaster's Theory of the Second Best (Lipsey and Lancaster 1956). If you don't have control over every parameter, setting the parameters you do control to the ideal values is generally inadvisable.

[2] I'm drawing here on work on the nature of idealisations by Michael Strevens (2008) and by Kevin Davey (2011).

in the used car market, buyers had less information about the car than sellers, you could get a discount similar to what you saw in real life even if the buyers had no special preference for new cars. Rather, buyers had a preference for good cars, and took the fact that this car was for sale to be evidence that it was badly made. It was important for Akerlof's explanatory purposes that he could show that people were being rational, and this required that he have a decision theory that they followed. In fact what he used was something like GDT. We now have excellent evidence that something like his model was correct. As the variation in quality of new cars has declined, and the information available to buyers of used cars has risen, the used car discount has just about vanished. (In fact it went negative during the pandemic, for reasons I don't at all understand.)

I'll end this section with a response to one objection, one caveat, and one surprising bonus to doing idealised decision theory this way.

The objection is that decision theory isn't actually that helpful for prediction and explanation. If all that it says are things like when rain is more probable, more people take umbrellas, that doesn't need a whole academic discipline. The response to that is that in non-cooperative games, the predictions, and explanations, can be somewhat surprising. One nice case of this is the discussion of Gulf of Mexico oil leases in Wilson (1967).[3] But here's a simpler surprising prediction that you need something like GDT to get.[4]

Imagine Row and Column are playing rock-paper-scissors. A bystander, C, says that he really likes seeing rock beat scissors, so he will pay whoever wins by playing rock $1. Assuming that Row and Column have no ability to collude, the effect of this will be to shift the payouts in the game they are playing from left table to right table, where c is the value of the dollar compared to the value of winning the game. This changes the game they are playing from Table 2.2a to Table 2.2b.

The surprising prediction is that this will *decrease* the frequency with which the bystander gets their way. The incentive will not make either party play rock more often, they will still play it one third of the time, but the frequency of scissors will decrease, so the *rock smash* outcome will be less frequent. Moreover, the bigger the incentive, the larger this increase will be[5]. Simple rules like "When behaviour is rewarded, it

[3] I learned about this paper from the excellent discussion of the case in Sutton (2000).

[4] A somewhat similar point is made in the example of the drowning dog on page 216 of Bonanno (2018).

[5] The proof is in Appendix One.

Table 2.2: Two versions of Rock-Paper-Scissors

(a) Original game

	Rock	**Paper**	**Scissors**
Rock	0,0	-1,1	1,-1
Paper	1,-1	0,0	-1,1
Scissors	-1,1	1,-1	0,0

(b) Modified game

	Rock	**Paper**	**Scissors**
Rock	0,0	-1,1	1+c,-1
Paper	1,-1	0,0	-1,1
Scissors	-1,1+c	1,-1	0,0

happens more often" don't always work in strategic settings, and it takes some care to tell when they do work.

The caveat is that there is a reason that this particular idealisation is chosen, at least as the first attempt. There are a lot of stylised facts about people that we could use in a model of behaviour. In game theory we concentrate on the ways in which people are, at least approximately much of the time, somewhat rational. People who prefer vanilla to chocolate really do buy vanilla more than chocolate. We could also choose stylised facts that are not particularly rational. But there is a worry that these will not remain facts, even approximately, when the stakes go up. And for some purposes, what people do in high stakes situations might be really important. If you think that people are more careful in high stakes situations, and that this care translates into more rational action, and it's particularly important to make the right predictions in high stakes cases, it makes sense to focus on idealisations that are also true of perfectly rational people.

There is a tricky complication here that isn't always attended to in theory, though in practice it's less of a problem. Especially in decision theory, we idealise away from computational shortcomings, but not away from informational shortcomings. We take the chooser's information as fixed, and ask how they'll decide. In high stakes cases, people don't just get better calculators, they get more information. If this is why we idealise, why don't we idealise away from ignorance? The reason is that in a lot of cases, either it is impossible to get the evidence the chooser needs, because it is about the future, or it is challenging because there is someone else working just as hard to prevent the chooser getting the information. It's not a coincidence that game theory, and decision theory, are most explanatory when the chooser's ignorance is about the future, or about something that someone is trying to hide from them.

There is one surprising bonus from starting with these rational idealisations. Some-

times one gets a powerful kind of explanation from very carefully working out the ideal theory, and then relaxing one of the components. At a very high level of abstraction, that's what happened with the development of cursed equilibrium (Eyster and Rabin 2005). The explanations one gets these ways are, to my mind, very surprising. The models have people acting as if they have solved very complex equations, but have ignored simple facts, notably that other people may know more than they do. But if the model fits the data, it is worth taking seriously. And while it was logically possible to develop a model like cursed equilibrium without first developing an ideal model and then relaxing it, it seems not surprising that in fact that's how the model was developed.

So our topic is idealised decision theory. In practice, that means the following things. The chooser can distinguish any two possibilities that are relevant to their decision, there is no unawareness in that sense, and they know when two propositions are necessarily equivalent. They can perform any calculation necessary to making their decision at zero cost. They have perfect recall. They don't incur deliberation costs; in particular, thinking about the downsides of an option does not reduce the utility of ultimately taking that option, as it does for many humans. They know what options they can perform, and what options they can't perform. I'll argue in Chapter 5 that it means they can play mixed strategies. Finally, I'll assume it means they have numerical credences and utilities. I'm not sure this should be part of the same idealisation, but it simplifies the discussion, and it is arguable that non-numerical credences and utilities come from the same kind of unawareness that we're assuming away. (Grant, Ani, and Quiggin 2021)

So the problems our choosers face look like this. There are some possible states of the world, and possible choices. The chooser knows the value to them of each state-choice pair. (In Chapter 3 I'll say more about this value.) The states are, and are known to be, causally independent of the choices. But the states might not be probabilistically independent of the choices. Instead, we'll assume that the chooser has a (reasonable) value for $\Pr(s \mid c)$, where s is any one of the states, and c is any one of the choices. The question is what they will do, given all this information.

3 Expectationist

There is a strange split in contemporary decision theory. On the one hand, there are questions about the way to model attitudes to risk, largely organised around the challenge to orthodoxy from Quiggin (1982) and Buchak (2013). On the other hand, there are questions about what to do in cases where the states are causally but not probabilistically independent of one's actions, with the central case being Newcomb's Problem (Nozick 1969). The strange split is that these two literatures have almost nothing in common.[1]

This split might seem to make sense when one reflects that there is no logical difficulty in endorsing any prominent answer to one set of questions with any prominent answer to the other set. But things get more difficult quickly. For one thing, one answer to questions about risk, what I'll call the expectationist answer, is universally assumed by people working on issues around Newcomb's Problem. For another, the argument forms used in the two debates are similar, and that should affect how the two arguments go.

Say that a normal decision problem is one where the states are probabilistically independent of the choices. A simple example is betting on a coin flip. In talking about normal decision problems I'll normally label the states H, for Heads, or T for Tails. Unless otherwise stated coins are fair, so H and T are equiprobable.

Say that an abnormal decision problem is simply one that isn't normal. A simple example is where the states are predictions of an arbitrarily accurate predictor. I'll normally label such states as PX, where X is a choice the agent may make. In these cases the Predictor is arbitrarily accurate unless otherwise stated, but we will spend some time with more error prone predictors.

[1] There is a survey article from a few years ago - Elliott (2019) - that has summaries of the then state-of-the-art on these two questions. And it makes it very striking how little the literatures on each of them overlap.

The view I call expectationism has two parts. First, it says that in normal decision problems, the rational agent maximises the expected value of something like the value of their action. Second, it says that something like this expected value plays an important role in the theory of abnormal decision problems. These definitions are vague, so there are possible borderline cases. But in practice this doesn't arise, at least in the philosophy literature. Everyone working on abnormal problems is an expectationist. Indeed, most work assumes without even saying it that the first clause of expectationism is correct. Everyone working on normal problems makes it clear which side they fall on, so there is no vagueness there. And every game theory text is expectationist.

I'm going to mostly follow suit. So why am I belabouring this point? One small reason and one large reason. The small reason is that one of the arguments I'll give concerning abnormal cases generalises to an argument for expectationism about normal cases. The other reason is dialectical.

In the debate about normal cases, the method of gathering intuitions about cases, and seeing which theory fits the intuitions best, does not favour expectationism. On the contrary, the Quiggin-Buchak theory does a much better job on that score. There is something incoherent about assuming expectationism is true for normal cases, and then thinking that the right way to theorise about abnormal cases is asking which theory fits intuitions best. If that's the goal of decision theory, we shouldn't be expectationist to start with.

The argument for expectationism is not that it fits the intuitions about cases best, but that it's the only theory that is compatible with various highly plausible principles, such as the Sure Thing Principle. Again, the theorist working on abnormal cases who is an expectationist has a dialectical burden here. They don't have to believe in the Sure Thing Principle, and indeed many expectationists don't (Gallow, n.d.). But they do have to believe in some principle that can be used to make an argument for expectationism. Especially when it comes to evidential decision theorists, I'm not sure what that principle might be. Still, I don't have an argument that there is no such principle, so I'll just note this is a challenge, not any kind of refutation.

Expectationism has a big practical advantage; it lets us treat the payouts in a game table as expected values, not any kind of final value. This is useful because it is very rare that a decision problem results in outcomes that have anything like final value. Often we are thinking about decision problems where the payouts are in dollars, or some other currency. That's to say, we are often considering gambles whose payout

is another gamble. Holding some currency is a bet against inflation; in general, the value of currency is typically highly uncertain.[2] For the expectationist, this is not a serious theoretical difficulty. As long as a dollar, or a euro, or a peso, has an expected value, we can sensibly talk about decision problems with payouts in those currencies. Depending on just how the non-expectationist thinks about compound gambles, they might have a much harder time handling even simple money bets.[3]

[2] See Alcoba (2023) for what happens when people start thinking that bet is a bad one.

[3] Joanna Thoma (2019) develops a subtle critique of some non-expectationist theories starting with something like this point.

4 Causal

It shouldn't be controversial to claim that game theory textbooks are committed a broadly causal version of decision theory.[1] For one thing, they always recommend defecting in Prisoners' Dilemma, even when playing with a twin. As David Lewis showed, this is equivalent to recommending two-boxing in Newcomb's Problem (Lewis 1979). They endorse the causal decision theorist's signature argument form: the deletion of strongly dominated strategies. Indeed, the typical book introduces this before it introduces anything about probability. When they do get around to probabilities, they tend to define the expected value of a choice in a way only a causal decision theorist could endorse. In particular, they define expected values using unconditional, rather than conditional, probabilities.[2] And the probabilities are simply probabilities of states, not probabilities of any kind of counterfactual. Indeed, you can go entire textbooks without even getting a symbol for a counterfactual conditional.

What's more controversial is that they are right to adopt a kind of causal decision theory (CDT).[3] In the recent literature, I think there are four main kinds of objections to CDT. First, it leaves one with too little money in Newcomb's Problem. Second, it gives the wrong result in problems like Frustrator (Spencer and Wells 2019). Third, it gives the wrong result in asymmetric Death in Damascus cases, as in Egan (2007). Fourth, it gives strange results in Ahmed's *Betting on the Past* and *Betting on the Laws*

[1] This point is made by Harper (1988), and many (though not all) of the conclusions I draw in this paper will be similar to ones he drew.

[2] See, for instance, the introduction of them on page 136 of Bonanno (2018). And note that we get 135 pages before the notion of an expectation is introduced; that's how much is done simply with dominance reasoning

[3] Note that I'm using *CDT* here as the name of a family of theories, not a particular theory. So it's not a great name; Causal Decision Theory is not a theory. Different versions of CDT can, and do, differ in what they say about the Stag Hunt cases I'll discuss in Chapter 7. But the label seems entrenched, so I'll use it. In contrast, evidential decision theory, EDT, is a theory; it is a full account of what to do in all cases.

Table 4.1: A Newcomb problem with two demons

(a) Demon-1 predicts Down

	PUp	PDown
Up	1	3
Down	0	2

(b) Demon-1 predicts Up

	PUp	PDown
Up	1001	1003
Down	1000	1002

cases. I'm going to set those problems aside because (a) they require that an agent not always be aware of what actions are possible, and that's inconsistent with the idealisations introduced in Chapter 2, and (b) they raise questions about just what it means for two things to be causally independent that go beyond the scope of this paper.

The intuitions behind the asymmetric Death in Damascus cases are inconsistent with the Exit Principle that I'll discuss in Chapter 7. The Frustrator cases are no problem for a version of CDT that says that idealised agents can always play mixed strategies. Like the game theorists, I will also assume mixed strategies are available, and I'll come back in Chapter 5 to why that assumption should be allowed.

That leaves the point that CDT leaves one poorly off in Newcomb's Problem, while other theories, like evidential decision theory (EDT) leave one well off. This isn't a particular mark against CDT, since other theories, like EDT, leave one poorly off in some situations. Here is one such case.[4]

There are two demons, who will predict what Chooser will do. Both of them are arbitrarily good, though not quite perfect, and their errors are independent. Chooser will play either the left or right game in Table 4.1.

If Demon-1 predicts that Chooser will play Down, Demon-1 will offer Chooser Table 4.1a; if Demon-1 predicts that Chooser will play Up, Demon-1 will offer Chooser Table 4.1b. Then Demon-2's prediction will be used for determining whether the payout is from column PU or PD. In almost all cases, if Chooser uses CDT, they will get 1001, while if they use EDT, they will get 2. So in this case, CDT will get more than EDT.

[4] See Wells (2019) for a slightly more complicated case, and Ahmed (2020) for an argument that Wells's argument is unfair to EDT.

This case is not meant as an objection to EDT. It is perfectly fair for the evidential decision theorist to complain that they have simply been the victim of a Demon who intends to punish users of EDT, and reward users of CDT. That seems a perfectly fair complaint. But if the evidential decision theorist makes it, they cannot object when causal decision theorists, such as Lewis (1981), use the same language to describe Newcomb's Problem. The 'objection' that CDT leaves one poorly off in one particular case is equally an objection to everyone, and so it is an objection to no one.

One might worry at this stage that I haven't shown that everyone is vulnerable to this kind of 'objection', just that CDT and EDT are equally vulnerable to it. In particular, so-called 'resolute' decision theories will choose one-box in Newcomb's Problem, and Up in Table 4.1, and so be enriched both times. The so-called 'foundational decision theory' that Levinstein and Soares (2020) endorse also makes that pair of choices. But those theories are vulnerable to much more serious objections, that I'll come to in Chapter 8.

So I conclude that there is no good objection to adopting a broadly causal decision theory, much as the game theorists do. But which version of CDT do they adopt, and are they right to do so? That will take us much more time.

5 Mixtures

Perhaps the biggest difference between the decision theory found in game theory textbooks, and the one found in philosophy journals, concerns the status of mixed strategies. In the textbooks, mixed strategies are brought in almost without comment, or perhaps with a remark about their role in a celebrated theorem by Nash (1951). In philosophy journals, the possibility of mixed strategies is often dismissed almost as quickly.

The philosophers' dismissal is usually accompanied by one or both of these reasons.[1] One is that the chooser might not be capable of carrying out a mixed strategy. They might not, for instance, have any coins in their pocket.[2] The other is that the predictor might punish people for randomising in some way, so the payouts will change. I'm going to argue that both reasons overgenerate. If they are reasons to reject mixed strategies, they are also reasons to reject the claim that agents have perfect knowledge of arithmetic. Since we do assume the latter, in decision theory agents take any bet on a true arithmetic claim at any odds, since all arithmetic truths have probability 1, we should also assume mixed strategies are permitted.

It isn't obvious why choosers should be perfect at arithmetic. True, calculators are a real help, but not everyone has a calculator in their pocket. Even if they do have a smartphone, it's hard to, for instance, solve a travelling salesman problem (Robinson 1949) on a typical smartphone. Those problems involve a lot of arithmetic, but ultimately they are just arithmetic. What we mean by saying choosers are perfect at arithmetic is that we are idealising away from arithmetic shortcomings. Once we're allowed to idealise away from shortcomings, it makes equally good sense to idealise away from inabilities to mix.

[1] These reasons are both offered, briefly, by Nozick (1969), so they have a history in decision theory.

[2] Not a particularly realistic concern when everyone carries a smartphone, but in theory smartphones might not exist.

The thought that predictors might punish randomisation is even less conducive to decision theory as we know it. Compare what happens in this problem. Chooser will be given a sequence of pairs of two digit numbers. They can reply by either saying a number, or saying "Pass". If they say a number, they get $2 if it is the sum of those numbers, and nothing otherwise. If they pass, they get $1. The catch is that if they are detected doing any mental arithmetic between hearing the numbers and saying something, they will be tortured. Decision theory as we know it has nothing to say about this case. Ideally, they simply say the right answer each time, and all the theories in the literature say that's the right thing to do. In practice, that's an absurd strategy. Chooser should utter the word "Pass" as often as they can, before they unintentionally do any mental arithmetic. The point is that as soon as we put constraints on how Chooser comes to act, and not just on what action Chooser performs, decision theory as we know it ceases to apply. And playing a mixed strategy is a way of coming to act. Punishing Chooser for it is like punishing Chooser for doing mental arithmetic, and is equally destructive to decision theory.[3]

Being able to carry out a mixed strategy is of practical value, especially when there are predictors around. It's not good to lose every game of rock-paper-scissors to the nearest predictor. If some mental activity is of practical value, then being able to carry it out is a skill to do with practical rationality. The idealised agents in decision theory have all the skills to do with practical rationality. Hence they can carry out mixed strategies, since carrying them out is a skill to do with practical rationality. So I conclude that if we are idealising, and if that idealisation extends at least as far as arithmetic perfection, it should also extend to being able to carry out mixed strategies.

That's not to say all decision theory should be idealised decision theory. We certainly need theories for real humans. Nor is it to say that decision theory for agents who can't perform mixed strategies is useless. For any set of idealisations, it could in principle be useful to work out what happens when you relax some of them from the model. The thing that is odd about contemporary philosophical decision theory, and the thing I've been stressing in this section, is that there should be some motivation for why one leaves some idealisations in place, and relaxes others. I don't see any theoretical or practical interest in working out decision theory for agents who are logically and mathematically perfect, but can't carry out mixed strategies.

[3] It's important to remember here that we are doing idealised decision theory. My view is that idealised decision theory has nothing to say about cases where someone will be punished for doing mental arithmetic.

Such agents are not a lot like us; since we are not logically and mathematically perfect. And they aren't even particularly close to us; most people are better at carrying out unpredictable mixed strategies than they are at solving the optimisation problems they face in everyday life. That said, it's important to be cautious here. It's often hard to tell in advance which combinations of keeping these idealisations and relaxing those will be useful. Still, I haven't seen much use for the particular combination that most philosophers have landed on, and I'm not sure what use it even could have.

So from now on I'll assume (a) if two strategies are available, so is any mixed strategy built on them, and (b) if Chooser plays a mixed strategy, Demon can possibly predict that they play the mixed strategy, but not the output of it.

6 Ratificationist

Solution concepts in game theory tend to be equilibria. And by an equilibria, everyone is happy with their moves knowing what all the moves of all the players are. (Or, at least, they are as happy as they can be.) Put in decision theoretic terms, that means that all solutions are ratifiable; Chooser is happy with their choice once it is made.

Ratificationism used to be a more popular view among decision theorists. Richard Jeffrey (1983) added a ratifiability constraint to a broadly evidential decision theory. And ratifiability was endorsed by causal theorists such as Weirich (1985) and Harper (1986). It fell out of popularity, though it has been recently endorsed by Fusco (n.d.). The loss of popularity was for two reasons.

One was the existence of cases where there is (allegedly) no ratifiable option. Table 6.1 is one such case.

Table 6.1: A case with no pure ratifiable options.

	PUp	**PDown**
Up	3	5
Down	4	3

If Chooser plays Up, they would prefer to play Down. If Chooser plays Down, they would prefer to play Up. Things get worse if we add an option that is ratifiable, but unfortunate, as in Table 6.2.

Table 6.2: A case with only a bad pure ratifiable option.

	PUp	**PDown**	**PX**
Up	3	5	0
Down	4	3	0

X	0	0	0

The only ratifiable option is X, but surely it is worse than Up or Down. One might avoid this example by saying that there is a weak dominance constraint on rational choices, as well as a ratifiability constraint. That won't solve the problem, but it will turn it into a problem like Table 6.1, where there is no good solution. But that won't help us much, as was pointed out by Skyrms (1984), since in Table 6.3 there is no weakly dominant option, but X is surely still a bad play.

Table 6.3: Skyrms's counterexample to ratificationism.

	PUp	**PDown**	**PX**
Up	3	5	0
Down	4	3	0
X	0	0	ε

A better option is to insist, as Harper (1986) did, and as I argued in previous section, that if Chooser is rational, they can play a mixed strategy. In all three of these games, the mixed strategy of (0.5 U, 0.5 D) will be ratifiable, as long as Chooser forms the belief (upon choosing to play this), that Demon will play the mixed strategy (1/3 U, 2/3 D). And that's a sensible thing for Demon to play, since it is the only strategy that is ratifiable for Demon if Demon thinks Chooser can tell what they are going to do. And given Chooser's knowledge of Demon's goals, Chooser can tell what Demon is going to do once they choose.

So if mixed strategies are allowed, none of the problems for ratifiability persist. And since mixed strategies should be allowed, since Chooser is an ideal practical actor, and not being able to play mixed strategies is an imperfection.

Moreover, ratifiability is an intuitive constraint. There is something very odd about saying that such-and-such is a rational thing to do, but whoever does it will regret it the moment they act. So I'll follow the game theory textbooks in saying ratifiability should be part of the correct decision theory.

This does not mean that we need to have an explicit ratifiability clause in our theory. It could be, and arguably should be, that ratifiability is a consequence of the theory, not an explicit stipulation.

Could we defend ratifiability without appeal to mixed strategies? It's not a completely impossible task, but nor is it an appealing one.

Table 6.1 poses no serious problem. Without mixed strategies, the case is simply a dilemma. And we know that there are dilemmas in decision theory. Here's one familiar example. A sinner faces Judgment Day. Because of his sins, it is clear things will end badly for him. But he has done some good in his life, and that counts for something. The judge thinks he should get some days in the Good Place before being off to the Bad Place. But the judge can't decide how many. So the judge says to the sinner to pick a natural number n, and the sinner will spend n days in the Good Place, and then goodbye. This clearly is a dilemma; for any large n, saying $n!$ would be considerably better.[1] Ahmed (2012) says that it is an objection to a theory that it allows dilemmas in cases with finitely many options; dilemmas should only arise in infinite cases. But he doesn't really argue for this, and I can't see what an argument would be. Once you've allowed dilemmas of any kind, the door is open to all of them.

Nor does Table 6.2 pose a problem, since as I said, the ratifiability theorist could add a weak dominance constraint and turn Second table into another dilemma.

The problem is Table 6.3. There the ratifiability theorist who does not allow mixed strategies has to say that the case is an odd kind of Newcomb Problem, where the rational agent will predictably do badly. But it's a very odd Newcomb Problem; by choosing X the chooser didn't even make themselves better off. Indeed, they guaranteed the lowest payout in the game. I don't have a knock-down argument here, and maybe there is more to be said. This is where I think the argument for ratificationism really needs mixed strategies.

[1] Note that this is true even if days in heaven have diminishing marginal utility, so the dilemma can arise even if we work within bounded utility theory. This is not just the kind of problem, as discussed by Goodsell (n.d.), that arises in decision theory with unbounded utilities.

7 Indecisive

Game theory is full of *solution concepts*; ideas for how to solve a game. That is, they are methods for determining the possible outcomes of a game played by rational players. Compared to philosophical decision theory, there are two big things to know about these solution concepts. One is that there are many of them. It isn't like having a single theory to rule all cases. More complex theories tend to give more intuitive results on more cases. But the complexity is a cost, and in any case no theory gets all the intuitions about all the cases. The other thing is that these will often say that there are multiple possible outcomes for a game, and that knowing the players are rational doesn't suffice to know what they will do. It's this latter feature of game theory that I'll argue here decision theory should imitate.

Say that a theory is *indecisive* if for at least one problem it says there are at least two options such that both are rationally permissible, and the options are not equally good. And say, following Ruth Chang (2002), that two options are equally good if improving either of them by a small amount epsilon would make that one better, i.e., would make it the only permissible choice. So an indecisive theory says that sometimes, multiple choices are permissible, and stay permissible after one or other is sweetened by a small improvement. The vast majority of decision theories on the market are decisive. That's because they first assign a numerical value to each option, and say to choose with the highest value. This allows multiple options iff multiple choices have the same numerical value. But sweetenings increase the value, so they destroy equality and hence the permissibility of each choice.

Perhaps the most intuitive case for indecisiveness involves what I'll call Stag Hunt decisions.[1] Here is an example of a Stag Hunt decision.

[1]For much more on the philosophical importance of Stag Hunts, see Skyrms (2004).

Table 7.1: An example of a Stag Hunt.

	PUp	**PDown**
Up	6	0
Down	5	2

Note three things about this game. First, both Up and Down are ratifiable. Second, Up has a higher expected return than Down. Third, Up has a higher possible regret than Down. If Chooser plays Up and Demon is wrong, Chooser gets 2 less than they might have otherwise. (They get 0 but could have got 2.) If Chooser plays Down and Demon is wrong, Chooser only gets 1 less than they might have otherwise. (They get 5 but could have got 6.)

There is considerable disagreement about what this means for Chooser. EDT says that Chooser should play Up, as does the ratifiable variant of EDT in Jeffrey (1983), and some causal decision theorists such as Arntzenius (2008) and Gustafsson (2011). On the other hand, several other theorists who endorse two-boxing in Newcomb's Problem, like Wedgwood (2013), Gallow (2020), Podgorski (2022), and Barnett (2022), endorse playing Down on the ground of regret miminisation. I think both Up and Down are permissible.[2] I also think this is the intuitively right verdict, though I place no weight on that intuition. In general, I think in any problem that has the three features described in the last paragraph (two equilibria, one better according to EDT, the other with lower possible regret), either option is permissible. Since lightly sweetening either Up or Down in this problem doesn't change either feature, that is why my theory is indecisive.

My argument for indecisiveness will turn on a case that all seven of the views mentioned in the last paragraph agree on, namely Table 7.2.

Table 7.2: An example of a coordination game.

	PUp	**PDown**
Up	4	0
Down	0	3

[2]The view I'm going to develop is hence similar to the 'permissive CDT' defended by Fusco (n.d.).

Table 7.3: The abstract form of an exit problem.

(a) Exit Parameters

Exit Payout	e
Pr(Exit | PUp)	x
Pr(Exit | PDown)	y

(b) Round 2 game

	PUp	**PDown**
Up	a	b
Down	c	d

All of them agree that Up is the uniquely rational play in this example, and I think intuition agrees with them. I'll argue, however, that Down is permissible. The argument turns on a variation that embeds cite table in a more complicated problem. This problem involves two demons, each of whom are arbitrarily good at predicting Chooser. The (first version of) the problem involves the following sequence.

1. Both Demon-1 and Demon-2 predict Chooser, but do not reveal their prediction.
2. If Demon-1 predicts Chooser plays Up, they Exit with probability 0.5, and Chooser gets 0. If Demon-1 predicts Chooser plays Down, they do not Exit. (That is, they Exit with probability 0.) If they Exit, the problem ends, and Chooser is told this. Otherwise, we go to the next step.
3. Chooser chooses Up or Down.
4. Demon-2's prediction is chosen, and that determines whether we are in state PU or state PD.
5. Chooser's payouts are given by cite above table.

I'll call these Exit Problems, and Table 7.3 gives the general form of such a problem. Our problem has this abstract of Table 7.3 with $b = c = e = y = 0$, $x = 0.5$, $a = 4$, $d = 3$.

Now consider a simple variant of the above 5 step problem. The same things happen, but steps 2 and 3 are reversed. That is, Chooser decides on Up or Down after Demons make their predictions, but before they are told whether Demon-1 decided to Exit. Still, their choice will only matter if Demon-1 decided not to Exit, since their choices do not make a difference if Demon-1 Exits. Call this variant the Early Choice version, and the original the Late Choice variant. I don't have any clear intuitions about what to do in most Exit Problems, save for this constraint on choices.

- **Exit Principle**: In any Exit Problem, the same choices are permissible in the Early Choice and Late Choice variants.

The reason comes from thinking about what Chooser is doing in the Early Choice variant. They are making a decision about what to do if Demon-1 doesn't Exit. The way to make that decision is just to assume that Demon-1 doesn't Exit, and then decide what to do. It just is the same choice as they face in the Late Choice variant, except now they make it in the context of a conditional. So they should decide it the same way.

To put the point in game-theoretic terms, there is no difference between extensive form and normal form reasoning when a decider has only one possible choice to make. And there is a natural argument for this claim. It starts with the idea that for any p, the following two questions have the same answers.

1. If p happens, what do you want to do?
2. So, p happened. What do you want to do?

When one is asked to choose a strategy for a tree that has only one possible decision in it, the question one is being asked is that if we get to the point in the tree where one has to decide, what will you do. And the 'Late Question' is that that point in the tree has been reached; now what will you do? So the questions fit the schema, and should get the same answers. One could see this as a consequence of applying something like the Ramsey test to conditional questions (Ramsey 1990). Denying Exit Principle means treating these two very similar sounding questions differently, and that's implausible.

One could also argue, I think correctly, that anyone who violates Exit Principle will violate a plausible version of the Sure Thing Principle.[3] Such an argument seems sound to me, but the Sure Thing Principle is controversial, and I prefer to put more weight on the argument from how conditional reasoning works in the previous paragraph. (Indeed, I think using the Exit Principle to motivate a version of the Sure Thing Principle is more plausible than the reverse argument.)

[3]To be sure, it's not entirely clear how to even state the Sure Thing Principle in the framework of causal ratificationism. Ratificationism does not output a preference ordering over options; it just says which options are and are not choice-worthy. And exactly how to translate principles like Sure Thing that are usually stated in terms of preference to ones in terms of choiceworthiness isn't always clear. One consequence of this is that I don't want to lean on Sure Thing as a premise. Another is that ratificationism isn't really subject to the objections that Gallow (n.d.) makes to theories that endorse Sure Thing, since the version of Sure Thing he uses is stated in terms of preferences. (Officially, ratificationism is 'unstable' in his sense because it doesn't output a preference ordering over unchosen options; that doesn't seem like a weakness to me.)

Any plausible theory that says that only Up is rationally playable in problems like Table 7.2 cite above table will violate Exit Principle. Think about what they will say Table 7.4.

Table 7.4: The Early Choice decision.

	PUp	**PMixed**	**PDown**
Up	2	3	0
Down	0	3	3

In this problem, PUp means that both demons predict Up, PDown means that they both predict Down, and PMixed means that one predicts one, and one the other. This possibility is arbitrarily improbable, and the two strategies have the same expected return given M in any case, so we can ignore it. So really this game comes to Table 7.5.

Table 7.5: The Early Choice decision simplified.

	PUp	**PDown**
Up	2	0
Down	0	3

Now presumably if one prefers Up in above table, it is because one prefers Up in any game like Table 7.6 Table below where $x > y > 0$.

Table 7.6: General coordination game.

	PUp	**PDown**
Up	x	0
Down	0	y

How could it be otherwise? Given expectationism, it's not like there is anything special about the numbers 4 and 3. But anyone who endorses this policy will play Down Table 7.5 and so, presumably, in Table 7.4. And that means they will violate Exit Principle.

Table 7.8: An exit problem with Frustrating Button in round 2.

(a) Exit Parameters

Exit Payout	-50
Pr(Exit | PUp)	0.8
Pr(Exit | PDown)	0

(b) Frustrating Button

	PUp	**PDown**
Up	10	10
Down	15	0

The only view that is consistent with Exit Principle in cases like Table 7.6 is that both Up and Down are permissible. And since in any such case, improving Up or Down be a tiny amount wouldn't materially change the case, they must both be permissible after small sweetenings. So, given Exit Principle, the only viable theories are indecisive.

Exit Principle also offers a response to some intuitions that have led people to question CDT in recent years. Table 7.7 is an example that Jack Spencer (2023) used to model the kind of case that's at issue. As he notes, it is similar to the psychopath button case (Egan 2007), the asymmetric Death in Damascus case (Richter 1984), and other puzzles for CDT.

Table 7.7: Frustrating Button (from Spencer (2023)).

	PUp	**PDown**
Up	10	10
Down	15	0

Apparently the common intuition here is that that Up is the uniquely rational play. Note though that if we embed Frustrating Button in an exit problem, as in Table 7.8, the intuitions shift.

The Early Version of Table 7.8 is Table 7.9.

Table 7.9: Early Version of Table 7.8.

	PUp	**PDown**
Up	-38	10
Down	-37	0

And if there is an intuition here, it is that it's better to choose Down rather than Up.[4] This violates Exit Principle, and it seems incoherent to say that one would choose Down in this game, when Down just means playing Down in round 2, and if one were to reach round 2, one would prefer Up.

Exit Principle can also be used to argue against the non-expectationist theory offered by Lara Buchak (2013), but that argument is more complicated, and I'll leave it to Appendix Two.

[4]The theory offered in Spencer (2021b) agrees with intuition here.

8 Dual Mandate

Say a decision tree is a series of steps with the following characteristics.

- At every step, Chooser either receives some information, or makes a choice.
- Chooser knows before the first step what possible choices will be available at each step, given the prior steps, or what possible pieces of information could be received.
- No matter what happens, the tree ends after finitely many steps. (Though it may end after more or fewer steps depending on what happens).
- Chooser knows before the first step what payout they will receive given each possible sequence of choices and information.
- Before the first step, chooser has a probability for each possible piece of information they could receive, given the prior steps in the tree.

That's incredibly abstract, but it excludes some possibilities. It excludes cases where Chooser learns along the way that they have hitherto unknown abilities. It excludes cases where Chooser gains the capacity to think new relevant thoughts along the way, say by meeting a new person and gaining the capacity to have singular thoughts about them.[1] Still, it does cover a lot of cases.

Say a strategy for a decision tree is a plan for what to do in every possible choice situation. Following the game theory textbooks, I really do mean *every* here. A strategy should say what to do in cases that are ruled out by Chooser's prior choices. A strategy for playing chess as White might say to start with e4, but also include plans for what to do if you inexplicably start Na3. There are both mathematical and philosophical reasons for having such an expansive conception of strategies, but going into why is beyond the scope of this paper.

[1] Following Stalnaker (2008), I think it excludes the Sleeping Beauty case, since there Beauty gains the capacity to have singular thoughts about a time, the 'now' when she awakes, that she did not previously have.

In philosophy, there are two common approaches to decision trees. The so-called resolute approach[2] says that one should simply treat the problem as like the kind of one-shot decisions we have discussed so far, except now one is choosing a strategy. Whatever one's theory of choice is, one should simply apply it to the question of which strategy is best. The so-called sophisticated approach says that one should make the current choices that make best sense given one's views about what one's future self will do.

The orthodoxy in game theory, going back to at least Reinhard Selten (1965), is that both views are correct. When faced with a decision tree, Chooser should follow the advice of the sophisticated theorists, and (given they are ideally rational) do what would be best on the assumption that future choices will be rational. But in doing so, they should instantiate (part of) a strategy that could be rationally chosen by the resolute chooser. I call this the Dual Mandate approach, and I am going to defend it.

Start with why it is bad to just have a resolute approach.[3] Game theorists usually reject this approach because it means sometimes making a decision that one knows will have worse consequences than an available alternative. I'll go over an example of this, though I should note it is rather violent. This is unavoidable; it is only in these violent cases that we can be sure the Chooser is really making things worse, and not acting for a strategic or reputational goal.

Chooser is the Prime Minister of a small country, and they are threatened by a large neighbour. Unfortunately, neighbour is thinking of carpet bombing Chooser's capital, in retaliation for some perceived slight. Chooser has no air defences that would prevent a great destruction, and no allies who will rally to help. Fortunately, Chooser has a mighty weapon, a Doomsday device, that could destroy neighbour. Chooser has obviously threatened to use this, but neighbour suspects it is a bluff. This is for a good reason; the doomsday device would also destroy Chooser's own country. Neighbour is known to employ a Demon who is at least 99% accurate in predicting what military plans Chooser will take. So Chooser can do Nothing (N), or use the Doomsday device (D), should neighbour attack. Chooser would obviously prefer no attack, and would certainly not use the device preemptively. So here is the table.

[2] Most notably defended by McClennan (1990).

[3] The so-called Foundational Decision Theory of Levinstein and Soares (2020) agrees with the resolute approach in the special case where the only information Chooser will receive are the results of predictions, and is subject to the criticisms I'll make of resolute theories.

Table 8.1: Deciding whether to retaliate.

	PN	**PD**
N	-1	0
D	-50	-50

In the top left, neighbour bombs Chooser's capital, thinking correctly that Chooser will not retaliate. In the top right and lower right, neighbour is sufficiently scared of the doomsday device that they do nothing. But in the bottom left, neighbour attacks, and Chooser retaliates, creating a disaster for everyone, something 50 times worse than even the horrors of the carpet bombing.

Still, if Chooser is picking a strategy before anything starts, the strategy with the highest expected return is to plan to retaliate. This has an expected return of -0.5; since one time in a hundred it returns -50, and otherwise it returns 0. The resolute theorist says that's what Chooser should do, even if they see the bombers coming, and they realise their bluff has failed. This seems absurd to me, and it is the kind of result that drives game theorists to the dual mandate, but resolute theorists are familiar with the point that their theory says that sometimes one should carry out a plan now known to be pointless. So instead of resting on this case, as decisive as it seems to many, I'll run through two more arguments against a purely resolute theory.

Change the example so that Chooser has two advisors who are talking to him as the bombers come in. One of them says that the Demon is 99% reliable. The other says that the Demon is 97% reliable. Whether Chooser launches the doomsday device should, according to the resolute theorist, depend on which advisor Chooser believes. This is just absurd. A debate about the general accuracy of a demon can't possibly be what these grave military decisions are based on.

Change the example again, and make it a bit more realistic. Chooser has the same two advisors, with the same views. Chooser thinks the one who says the Demon is 99% reliable is 60% likely to be right, and the other 40% likely. So Chooser forms the plan to retaliate, because right now that's the strategy with highest expected return. But now, to everyone's surprise, neighbour attacks. The resolute theorist will say that Chooser should stick to their (overpowered) guns. But think about how the choice of plans looks to Chooser now. The actions of neighbour are evidence about the reliability of the demon. And a simple application of Bayes' Rule says that

Chooser should now think the advisor who thought the demon was 97% reliable is 2/3 likely to be right. That is, given Chooser's current evidence, retaliating wasn't even the utility maximising strategy to start with. Yet it is what the resolute theorist, or at least the resolute theorist who is not also sophisticated, would have Chooser do. This is, again, absurd, and enough reason to give up on such a theory.

What about the other direction? Is it sensible to have a sophisticated theory that is not resolute? There does seem to be something puzzling about such theories. They are "diachronically exploitable" in the sense described by Spencer (2021a). Let's start with one example. Extend the theory offered by Gallow (2020) to make it a pure sophisticated dynamic theory. That is, in a decision tree, the chooser values future choices by the expected value of the choice they'll make, and if that choice is guaranteed to end the decision tree, they use Gallow's theory. Chooser is now offered the following two-step option. At step 1 they can choose to receive 1 or play the game in Table 8.2.

Table 8.2: A challenge for pure sophisticated decision.

	PU	**PD**
U	2	2
D	5	0

If Chooser gets to step 2, they'll play D, since it is the best option according to Gallow's theory. So at step 1 they'll choose the 1 rather than playing this game. But that's absurd; they know they could have done better by simply playing the game and choosing U.

What the Dual Mandate says is that the last step of reasoning here is sound; it is a fair criticism of an agent to say that their strategy doesn't make sense even if every step makes sense taken on its own. Since this does seem like a fair criticism, it is reasonable to adopt the Dual Mandate.

If one has a decisive theory, then a huge number of decision trees will be dilemmas, since it is unlikely that the optimal strategy matches the series of optimal choices. This is not a reason to reject the Dual Mandate; it's another reason to reject decisiveness.

You might worry that the argument based around Table 8.2 is not really an objection to theories that reject the Dual Mandate, but just to the combination of that rejection

Table 8.3: Ahmed Insurance (from Spencer (2021a)).

(a) First game

	$\mathbf{PU_1}$	$\mathbf{PD_1}$
$\mathbf{U_1}$	50	-50
$\mathbf{D_1}$	60	-40

(b) Second game

	Correct	**Incorrect**
$\mathbf{U_2}$	25	-75
$\mathbf{D_2}$	-25	75

and the endorsement of Gallow's particular theory of decision. That worry is half right. This result is a problem for Gallow's theory. But that doesn't mean it isn't also an argument for the Dual Mandate. The point of the Dual Mandate is not to criticise individual decisions, like taking the 1 in this game. It's to criticise theories that endorse those decisions. It's true that once we find the right theory of synchronic choice, the Dual Mandate will be unnecessary, since it will be automatically satisfied. But the Dual Mandate plays an essential role in selecting that theory.

Jack Spencer (2021a) has an example which he thinks tells against the Dual Mandate, or what he calls the requirement that Chooser not be diachronically exploitable.[4] The agent will play first the left and then the right game, and their payouts (shown in dollars) will be summed over the game. They won't be told between the games what they got from the first game.[5]

Note that in Table 8.3b, the states are not the usual ones about Demon's predictions. Rather, they are that the Demon made the Correct, or Incorrect, prediction in Table 8.3a. There are eight strategies in this game, but since the Demon doesn't care about what happens at non-chosen nodes, we won't care either, and just focus on the four combinations of moves Chooser might make, and how they interact with Demon's prediction. If we do that, we get the following table (also given by Spencer, and also with payouts in dollars).

Table 8.4: Strategic form of Ahmed Insurance.

	$\mathbf{PU_1}$	$\mathbf{PD_1}$

[4] Spencer's non-exploitability isn't quite the same thing as the Dual Mandate, but it's close enough for these purposes. Spencer rejects non-exploitability, but endorses a weaker constraint he calls the Guaranteed Principle. I don't see any reason to distinguish between these constraints, in part because of the argument that follows in the text.

[5] Assume Chooser is reasonably risk-neutral over dollars over this range of outcomes.

$\mathbf{U_1U_2}$	75	-125
$\mathbf{U_1D_2}$	25	25
$\mathbf{D_1U_2}$	-15	-15
$\mathbf{D_1D_2}$	135	-65

Spencer argues that even though D_1U_2 is dominated by U_1D_2 it might be rational to play it. After all, it is rational to bet on Demon being correct in Table 8.3b, since Demon is arbitrarily good. And if one knows one is going to do that, one may as well take the sure extra \$10 that playing U_1 rather than D_1 gives. So diachronic exploitability is consistent with rationality.

The reasoning of the previous paragraph fails because neither CDT, nor any other sensible decision theory, recommends taking two boxes in Newcomb Problems embedded in strategic interactions. This would be like thinking that CDT recommended always defecting in Iterated Prisoners' Dilemma, even it was chancy whether the iterations came to an end after each round, so backward induction reasoning was unavailable. If Chooser has convinced themselves that they will play U_2, and we'll come back to whether they should believe that, then the choice in Table 8.3a comes down to this.

Table 8.5: First game in Ahmed Insurance, if D_2 will be played.

	$\mathbf{PU_1}$	$\mathbf{PD_1}$
$\mathbf{U_1}$	75	-125
$\mathbf{D_1}$	-15	-15

This game has two pure strategy equilibria, and on its own I think (because of the arguments in Chapter 7) that either play is acceptable. In context though, either play is clearly unacceptable. Given that one chooses either U_1 or D_1, the only reasonable thing to believe is that Demon has almost certainly predicted this, so it makes to play U_2, since Demon is almost certainly correct. So one ends up playing U_1U_2 or D_1U_2, both of which are dominated and hence absurd strategies.

Spencer argues that since the Demon is almost certainly accurate, Chooser should play U_2, so they should play a dominated strategy, so the Dual Mandate doesn't apply. (This assumes that synchronic choice rules out strictly dominated options in

cases like this, but Spencer agrees that it does.) This argument only goes through if Chooser doesn't have access to mixed strategies; i.e., if Chooser is not ideally practically rational. If Chooser does have access to mixed strategies, they should play a 50/50 mix of U_1 and D_1, then choose D_2. That is ratifiable as long as Chooser believes Demon plays PU_1 with probability 0.45, and PD_1 with probability 0.55. Since that's the only ratifiable play for Demon, it's reasonable for Chooser to believe this. If mixed strategies are allowed, this is not a case where the Dual Mandate fails.

In general, if mixed strategies are not allowed, the Dual Mandate is implausible. But that's because without mixed strategies, cases like Table 8.3 are dilemmas; they have no ratifiable choices. And it's true that the Dual Mandate is implausible in dilemmas. Think back to the sinner described in Chapter 6. Imagine that sinner will in fact say that they get d days in heaven. Now complicate the case; they are offered a choice of $d!$ days in heaven, or to make their own choice. If they will in fact choose d, they should simply take $d!$, even though there are strategies available, like choosing $d!!$ days, that are better. Weird things happen when there are dilemmas around, and we shouldn't judge decision theories against these cases.

The Dual Mandate is also implausible if Chooser thinks they will be irrational, or that they will have different preferences. Indeed, it is implausible if Chooser thinks they might either change or lose their mind. For example, Odysseus binds himself to the mast because he does not approve of future-Odysseus's preferences. Professor Procrastinate[6] cite turns down a referee request because he does not trust his future self to be practically rational. Both of them deliberately turn down strategies that would be better than where they end up, because they do not trust their future selves to carry them out. They are alienated in this way from their future selves. When one does not endorse one's future preferences, or does not trust one's rationality in the future, it makes sense to be alienated from one's future self in this way. In such cases, one's future self is just another part of the world that must be predicted and worked around. And so it might make sense to forego, as Odysseus and Procrastinate forego, strategies that one's future self will not be so kind as to carry out.

My main claim here is when neither of those two conditions obtain, i.e., when one knows that one's future self will be rational and have the same preferences, one's choices should make strategic sense. That is, they should satisfy the fairly weak condition that they are part of some strategy that one could choose if one was simply choosing a strategy for the whole tree. Unless one fears future irrationality, or future

[6] A famous character in Jackson and Pargetter (1986).

change of preference, one should not be alienated from one's future self. If Chooser takes 1 rather than play Table 8.2, they are alienated in this way. They have to think, I know I'd be better off if I played U. But that fool future-me will play D instead, and blow up the plan. But future-them is not a fool; by hypothesis they are known to be ideally rational. So it isn't coherent to think this way, and that reveals that it is incoherent to 'rationally' take the 1. And that is why the Dual Mandate requires that one's strategy be rational, and not just the moves that make up the strategy.

9 Substantive

Here are two interesting characters. Piz wants to put mud on his pizza. This won't bring him joy, or any other positive emotions; he has a non-instrumental desire for mud pizza. Za wants to eat a tasty pizza, and believes that putting mud on his pizza will make it tasty. There is a long tradition of saying that the point of philosophical decision theory is not to evaluate beliefs and desires, but merely to say what actions those beliefs and desires do or should issue in. On such a view, both Piz and Za should (or at least will) put mud on their pizzas. Here is David Lewis expressing such a view.[1]

> The central question of decision theory is: which choices are the ones that serve one's desires according to one's beliefs? (Lewis 2020, 472)

We need one caveat on this. Philosophical decision theories typically do not issue verdicts unless the chooser satisfies some coherence constraints. So it's not quite that the theory says nothing about what the beliefs and desires should be. It's that it says nothing *substantive* about what the beliefs and desires should be. Purely structural constraints, like transitivity of preferences, or belief in the law of excluded middle, may be imposed.

At least sometimes, game theorists impose non-structural, substantive conditions on the beliefs of players. Most notably, the "intuitive criterion" of Cho and Kreps (1987) is meant to be continuous with other equilibrium conditions, and is a substantive constraint. Someone who violates it has coherent beliefs that don't conform

[1] I'm using Lewis as an example of the orthodox view that decision theory does not care about whether beliefs and desires are substantively rational, just that they are coherent. But note that Lewis has an idiosyncratic view in the neighbourhood of this one. He denies that the point of decision theory is to guide or judge action. He thinks that decision theory is primarily description, not normative. I agreed with that in Chapter 2. But he thinks its descriptive role is primarily in defining belief and desire; I think it is in explaining social phenomena.

to their evidence. The intuitive criterion takes some time to set up, but I'll get to a simplified version of it later in this section.

First, I'll note some general reasons for scepticism about this use of the substantive-structural distinction. One obvious point is that Piz and Za do not look like rational choosers. Another is that this draws distinctions between overly similar characters, such as these two, Cla and Sic. Both of them have taken classes in classical statistics, but only skimmed the textbooks without attending to the details. Cla came away with the belief that any experiment with a P value less than 0.05 proved that its hypothesis is true. Sic came away with a standing disposition to belief the hypothesis whenever there was an experiment with a P value less than 0.05. Cla is incoherent; there is no possible world where that belief is true. Sic is coherent; any one of their beliefs could be true. It's just they just have a disposition to often form substantially irrational beliefs. Personally, I don't think the difference between Cla and Sic is important enough to be philosophically load bearing. Lastly, it has proven incredibly hard to even define what makes a norm structural. The most important recent attempt is in Alex Worsnip's book *Fitting Things Together: Coherence and the Demands of Structural Rationality* (Worsnip 2021). Here's his definition:

> *Incoherence Test.* A set of attitudinal mental states is jointly incoherent iff it is (partially) constitutive of (at least some of) the states in the set that any agent who holds this set of states has a disposition, when conditions of full transparency are met, to revise at least one of the states. (Worsnip 2021, 132)

This won't capture nearly enough. If probabilism is correct, then non-probabilists about uncertainty like Glenn Shafer (1976) endorse incoherent views. If expectationalism is correct, then non-expectationalist decision theorists, like Lara Buchak (2013), endorse incoherent views. If classical logic is correct, then intuitionist logicians like Crispin Wright (2021) are incoherent. Those three all seem to meet Worsnip's conditions of full transparency, and don't seem disposed to revise their beliefs. Maybe this is just a problem with Worsnip's definition, but it is also a reason to be sceptical that there even is a distinction to be drawn here. Wooram Lee (n.d.) raises some different challenges for Worsnip, and offers a rival theory. But for that theory to work, Lee requires that when a dialethist proposes to solve the Liar Paradox by saying the liar sentence is both true and not true, they are being insincere. The idea is that sincerely saying p requires believing p and not believing its negation. But this simply isn't part of the concept of sincerity, and as much as I find the dialethist

solution to the Liar implausible, I think the dialethists I know have been perfectly sincere in offering it. Maybe there is some theory of coherence waiting to be found, but the search for one feels like a degenerating research program.[2]

Even if the substantive/structural distinction can be made precise, and shown to do philosophical work, it won't track the notion game theorists most care about. We can see this with a version of the beer-quiche game Cho and Kreps (1987), here translated into decision-theoretic language.

There are five steps in the game.

1. A coin will be flipped, landing Heads or Tails. It is biased, 60% likely to land Heads. It will be shown to Chooser, but not to Demon.
2. Chooser will say either Heads or Tails.
3. Demon, knowing what Chooser has said, and being arbitrarily good at predicting Chooser's strategy[3]. will say Heads if it is more probable the coin landed Heads, and Tails if it is more probable the coin landed Tails.[4]
4. Chooser is paid $30 if Demon says Heads, and nothing if Demon says Tails.
5. Chooser is paid $10 if what they say matches how the coin landed, and nothing otherwise. This is on top of the payment at step 4, so Chooser could make up to $40.

If you prefer things in table form, here are the payouts chooser gets, given what happens at steps 1-3.

Table 9.1: The coin game.

Coin	Chooser	Demon	Dollars
H	H	H	40
H	H	T	10
H	T	H	30
H	T	T	0
T	H	H	30
T	H	T	0
T	T	H	40

[2] See also Heinzelmann (n.d.) for a different set of reasons to be sceptical that there is a notion of coherence that can do the work its philosophical defenders want.

[3] That is, what Chooser will do if Heads, and what they will do if Tails.

[4] If both are equally likely, Demon will flip a fair coin and say how it lands.

T	T	T	10

What will Chooser do? There are two coherent things for Chooser to do, though each of them is only coherent given a background belief that isn't entailed by the evidence.

1. Chooser could say Heads however the coin lands. Demon gets no information from Chooser, so their probability that the coin landed Heads is 0.6, so they will say Heads. Further, Chooser believes that if they were to say Tails, Demon would say Tails, so saying Heads produces the best expected return even after seeing the coin.
2. Chooser could say Tails however the coin lands. Demon gets no information from Chooser, so their probability that the coin landed Heads is 0.6, so they will say Heads. Further, Chooser believes that if they were to say Heads, Demon would say Tails, so saying Tails produces the best expected return even after seeing the coin.

While both of these are coherent, there is something very odd, very unintuitive about option 2. I guess we've been trained to be sceptical when philosophers report intuitions, but here we have a very large data pool to draw on. Cho and Kreps reported essentially the same intuition. Their paper has been cited tens of thousands of times, and I don't think this intuition has been often questioned. Option 2, while coherent, is unintuitive. It is the kind of option that the theory of rationality behind game theory, and behind decision theory, should rule out.

But what about it is incoherent? One might think it is because it has an expected return of \$34, while option 1 has an expected return of \$36. But we showed in section Indecisive ref that using expected returns to choose between coherent options leads to implausible results. Moreover, if you change the payout in the bottom row to \$50, the intuition doesn't really go away, but the expected return of option 2 is now \$38; higher option 1's payout.[5] Alternatively, one might think it is because option 2 requires Chooser to believe a counterfactual that is not entailed by the evidence. But option 1 also requires Chooser to believe a counterfactual that is not entailed by the evidence. That can't be the difference between them, but it is closer to the truth.

[5] I believe if you change that payout to \$65, the various regret based theories I discussed in Chapter 7 also start preferring option 2. But applying these theories to complex cases is hard, so I'm not quite sure about this.

What Cho and Kreps argue, persuasively, is that the difference between the options is that in one case the counterfactual belief is reasonable, and in the other it is unreasonable. Assume Chooser plans to adopt option 1. But when it becomes time to play, they change their mind, and say Tails. What would explain that? Not the coin landing Heads - given their plan, they will get the maximum possible payout by sticking to the plan (assuming Demon has done their job). No, the only plausible explanation is the coin landed Tails, and Chooser was (foolishly) chasing the extra $10. In option 1, Chooser believes the counterfactual that's grounded in Demon picking an explanation that makes sense. What about in option 2? Here, everything is back to front. If Chooser is ever going to depart from their plan, it's when the coin lands Heads. Then Chooser might chase the extra $10 by saying Heads. But Chooser has to believe that were they to depart from the plan, Demon would draw the explanation that makes no sense whatsoever, that they gave up on their plan even though it was about to lead to the best possible outcome. This makes no sense at all. And in fact it makes less sense the more you increase the payout in line 8.

So that's why decision theory requires substantive rationality. The right decision theory should say to take option 1. And the argument against option 2 is not that it is incoherent, but that carrying it out requires believing Demon will do things that make no sense given Demon's evidence. It is substantive, not structural, rationality that rules out option 2. And yet, as the game theorists have insisted, option 2 must be ruled out. So decision theory should be sensitive to substantial rationality.

10 Weak Dominance, Once

An option *a* weakly dominates another option *b* if *a* is at least as good as *b* in all states, and better than *b* in some states. Just what role weak dominance has in decision theory is one of the most unsettled topics in game theory. There are three natural positions, and all of them are occupied. One is that weak dominance is of no significance. A second is that ideal agents do not choose weakly dominated options, and that's the only role weak dominance has. And a third is that ideal agents do not choose options that are eliminated by an iterative process of deleting weakly dominated strategies. I'm going to argue in favour of the middle position. I'm not going to try to argue this is the standard game-theoretic move; as I said, I think you can find prominent support for all three options. To argue for the middle position requires making two cases: first, that weakly dominated options are not ideally chosen; and second, that options that would be eliminated by iterative deletion of weakly dominated options are ideally chosen. I'll argue for these in turn.

Start with Table 10.1; what would the ideal chooser do?

Table 10.1: A ratifiable, weakly dominated, option.

	PU	**PD**
U	1	1
D	0	1

On the one hand, D is ratifiable, as long as Demon is sufficiently reliable. If Demon will in fact get the predictions right, D gets a return of 1, and had Chooser played U, they would have still received 1. So they would not regret playing D, so by ratifiability it is fine to play it. Against this, there are three reasons to not play D.

First, it is rather unrealistic to think that the probability that Demon will make an accurate prediction is 1. And even if Demon's prediction is correct with probability 1 - □, then D will not be ratifiable. I'm inclined to rule out, on broadly Humean grounds,

the very possibility of a Demon whose predictions are correct with probability 1, but is causally independent of Chooser's choice. Temporally backwards causation is not a logical impossibility, and a world where the predictions are correct with probability 1 seems like a world which has backwards causation. I don't want to rest the case for GDT on contentious metaphysics, so I won't lean on this point.

Second, playing D involves taking on an uncompensated risk. It might be that we don't have a good way of capturing within probability theory just what this risk is. Perhaps you think that it makes sense to say that Demon is correct with probability 1. Still, in some sense D has a risk of failure that U lacks. One should not take on a risk without some compensation. So one should not play D in this case. This, I think, is the most persuasive argument against D.

Third, it has been argued by game theorists that we should always allow for the possibility that one or other player in a game will make some kind of performance error. This idea is at the heart of Reinhard Selten's notion of trembling hand equilibrium (R. Selten 1975), and Roger Myerson's notion of proper equilibrium (Myerson 1978). If a strategy would not make sense if the probability of an error by one or other player was positive, even if it was arbitrarily low, it should not be played. Since D only makes sense if the probability of an error by Demon is 0, that means D should not be played.

If it is good to remove weakly dominated options, then one might think it follows straight away that it is good to keep doing this.[1] Think about Table 10.2.

Table 10.2: An example of iterated weak dominance.

	PU	**PD**	**PX**
U	1	1	0
D	0	1	1
X	0	0	1

In Table 10.2, X is weakly dominated by D. So it shouldn't be played. But if X isn't played, then PX is weakly dominated by both PU and PD. Demon can't make a correct prediction by playing PX, since by hypothesis it won't be played, so it can't be better than PU or PD. But both PU and PD can be better than PX. So PX is now

[1]This suggestion is made by, for example Hare and Hedden (2015).

weakly dominated. So let's remove it as well. If both X and PX are deleted, we're back to Table 10.1, in which we said Chooser should play U.

So does it make sense to say that U is the only play in Table 10.2? I think not, for three reasons.

First, as Bonanno (2018, 37) points out, in general iterative deletion of weakly dominated strategies is not a well defined decision procedure. It turns out that in two player games, the order that weakly dominated strategies are deleted can affect which choices one ends up with. There are ways of fixing this problem, by specifying one or other order of deletion as canonical, but they all feel somewhat artificial.

Second, iterative deletion of weakly dominated strategies leads to a single solution to the money-burning game described by Ben-Porath and Dekel (1992). But, as Stalnaker (1998) showed, this game has multiple rational solutions, and arguments to the contrary turn on conflating indicative and subjunctive conditionals.

Third, the reasons we gave for avoiding the weakly dominated option in Table 10.1 simply don't carry over to Table 10.2. In the latter game, D is not an uncompensated risk. It's true that D loses if Demon makes an incorrect prediction and plays PU. But U loses if Demon makes an incorrect prediction and plays PX. Unless one thinks that PX is particularly unlikely to be played, it seems U and D are just as risky as each other. So both of them look like rational plays.

So I conclude that we just need one round of deleting weakly dominated options to get rid of irrational plays. D is irrational in Table 10.1, but not in Table 10.2.

11 Conclusion

Given all that, here is the positive theory, what I'm calling Gamified Decision Theory (GDT). The core is that rational choices are ratifiable. That is, they maximise expected utility from the perspective of someone who chooses them. Formally, that means that in a particular problem, with choices o_1, ..., o_m, and causally independent states s_1, ... s_n, a choice o is rational iff there is some probability function Pr that is a rational credal distribution over the s after choosing o, and such that for all o_j, $\Sigma \Pr(s_i)\mathrm{V}(os_i) \geq \Sigma \Pr(s_i)\mathrm{V}(o_j s_i)$.

There are two extra clauses. First, o is choiceworthy only if it is not be weakly dominated by any other option. Second, if one is making a series of decisions in a tree, and one knows that one will be rational and not change one's preferences throughout, then both the decisions one makes at any given time, and the set of decisions one collectively makes, must satisfy all the other conditions of rational choice. That is, they must be ratifiable and not weakly dominated.

The big advantage of this theory is that it satisfies the nine conditions I mentioned at the start, and that it is consistent with Exit Principle. These turn out to be very sharp constraints on a theory. For example, any theory that associates some number with each choice, and says to maximise the number, will violate the Indecisiveness constraint. Any theory that simply the chooser's credences as given will violate the Substantiveness constraint. The vast bulk of decision theories on the market in philosophy do at least one of these two things, and typically both. So we have strong reason to prefer GDT to them.

GDT does require that mixed strategies are available to choosers, on pain of saying that a lot of decision problems are dilemmas. That is not a problem for GDT, since ideal agents can perform mixed strategies. It is a shortcoming to not be able to perform them, and ideal agents don't have these kinds of shortcomings. But it does suggest an important research program in working out how GDT should be altered

for agents who lack one or other idealisation. In fact it suggests two research programs: a descriptive one, setting out what choosers who don't satisfy one or other idealisation in fact do; and a normative one, setting out what these choosers should do. But those projects are for a very different paper.[1]

[1]Thanks to many people for conversations on these topics, especially Dmitri Gallow and Ishani Maitra, and audiences and ACU and UBC, and students in my group choices classes at University of Michigan.

Appendix One: Rock-Paper-Scissors

This appendix shows how to find the equilibrium of Table 2.2b, the version of Rock-Paper-Scissors where it is common knowledge that the players will get a bonus of $c > 0$ if they will while playing rock. The game is symmetric, so we'll just work out Column's strategy, and the same will go for Row.

There is no pure strategy equilibrium of the game, so we have to find a mixed strategy for each player. And a mixed strategy equilibrium requires that every option that has positive probability has equal expected returns. (If that didn't happen, it wouldn't make sense to mix between them.) Let x be the probability (in equilibrium) that Column plays Rock, y that they play Paper, and z that they play Scissors. Given that, the expected return of the three options for Row are:

$$\begin{aligned} V(Rock) &= z(1+c) - y \\ V(Paper) &= x - z \\ V(Scissors) &= y - x \end{aligned}$$

We know that these three values are equal, and that $x + y + z = 1$. From this we can start making some deductions.

Since $x - z = y - x$, it follows that $x = \frac{y+z}{2}$. And that plus the fact that $x + y + z = 1$ implies that $x = \frac{1}{3}$. So we've already shown one of the surprising results; adding in the bonus c will not change the probability with which Rock is played. Substituting this value for x into the fact that Rock and Paper have the same payout, we get the following.

$$
\begin{aligned}
& & \frac{1}{3} - z &= z(1+c) - y \\
\Rightarrow & & \frac{1}{3} + y &= z(2+c) \\
\Rightarrow & & z &= \frac{y + \frac{1}{3}}{2+c}
\end{aligned}
$$

Now we can substitute the values for x and z into the fact that $x + y + z = 1$.

$$
\begin{aligned}
& & x + y + z &= 1 \\
\Rightarrow & & \frac{1}{3} + y + \frac{y + \frac{1}{3}}{2+c} &= 1 \\
\Rightarrow & (2+c) + 3y(2+c) + (3y+1) &= 3(2+c) & \quad \text{Multiply both sides by } 3(2+c) \\
\Rightarrow & 3cy + 9y + c + 3 &= 3c + 6 \\
\Rightarrow & 3cy + 9y &= 2c + 3 \\
\Rightarrow & y &= \frac{2c+3}{3c+9} \\
\Rightarrow & z &= \frac{3}{3c+9} & \quad \text{From previous derivation for } z
\end{aligned}
$$

So each option has expected payout $\frac{c}{3c+9}$. And there is one unsurprising result, namely that the expected return to the players increases as c increases. But note that x, the probability that a player plays Rock, is invariant as c changes. And z, the probability that a player plays Scissors, goes down as c goes up.

It is intuitive that announcing the reward makes each player less likely to play Scissors. And that in turn puts down downward pressure on playing Rock. What you need some theory (and algebra) to show is that this downward pressure is exactly as strong as the upward pressure that comes from the incentive for playing Rock supplied by the bystander. Intuition alone can tell you what the various forces are that are acting on a chooser; the role of theory is to say something more precise about the strength of these forces.

Appendix Two: Risk-Weighted Utility

This appendix goes over a problem for Lara Buchak's risk-weighted utility theory, based around the Exit Principle from Chapter 7. Buchak's theory concerns normal decision problems, where there are no demons lying around, so we have to modify Exit Principle a little to make it apply. The modifications still leave it recognisably the same principle though. And the main point of this appendix is to show that it is possible to theorise about normal and abnormal decision problems using the same tools.

The core of Buchak's theory is a non-standard way of valuing a gamble. For simplicity, we'll focus on gambles with finitely many outcomes. Associate a gamble with a random variable O, which takes values $o_1, \ldots, o_n$, where $o_j > o_i$ iff $j > i$. Buchak says that the risk-weighted expected utility of O is given by this formula, where r is the agent's risk-weighting function.

$$REU(O) = o_1 + \sum_{i=2}^{n} r(\Pr(O \geq o_i))(o_i - o_{i-1})$$

The decision rule then is simple: choose the gamble with the highest REU.

The key notion here is the function r, which measures Chooser's attitudes to risk. If r is the identity function, then this definition becomes a slightly non-standard way of defining expected utility. Buchak allows it to be much more general. The key constraints are that r is monotonically increasing, that $r(0) = 0$ and $r(1) = 1$. In general, if $r(x) < x$, Chooser is some intuitive sense more risk-averse than an expected utility maximiser, while if $r(x) > x$, Chooser is more risk-seeking. The former case is more relevant to everyday intuitions.

There are a number of good reasons to like Buchak's theory. Standard expected utility theory explains risk-aversion in a surprisingly roundabout way. Risk-aversion simply falls out as a consequence of the fact that at almost all points, almost all

Table 11.1: The abstract form of an exit problem with coins.

(a) Exit Parameters

Exit Payout	e
$\Pr(H_1)$	y
$\Pr(H_2)$	x

(b) Round 2 game

	H_2	T_2
Up	a	b
Down	c	d

goods have a declining marginal utiility. This is theoretically elegant - risk-aversion and relative satiation are explained in a single framework - but has a number of downsides. For one thing, it doesn't allow rational agents to have certain kinds of risk-aversion, such as the kind described by Allais (1953). For another, it doesn't seem like risk-aversion just is the same thing as the declining marginal utility of goods. Buchak's theory, by putting attitudes to risk into r, avoids both these problems.

Unfortunately, Buchak's theory runs into problems. Our focus will be on two-stage problems where Chooser's choice only makes a difference if the game gets to stage 2. The general structure will be this.

1. A coin with probability y of landing Heads will be flipped. If it lands Tails, Chooser gets the Exit Payout, and the game ends.
2. If the game is still going, a second coin, with probability x of landing Heads, will be flipped.
3. Chooser's payout will be a function of whether they chose Up or Down, and the result of this second coin.

I'll write H_1 and T_1 for the propositions that the first coin lands Heads and Tails respectively, and H_2 and T_2 for the propositions that the second coin lands Heads and Tails respectively. I'll mostly be interested in the case where Up is a bet on H_2, and Down is declining that bet, but the general case is important to have on the table. The general structure of these problems is given by Table 11.1.

Then we get a version of Exit Principle that applies to games like this.

- **Exit Principle**: Whether a choice is rational for Chooser is independent of whether Chooser chooses before or after they are told the result of the first coin flip.

Table 11.2: An exit game with exit payout 0.

(a) Exit Parameters

Exit Payout	0
$\Pr(H_1)$	y
$\Pr(H_2)$	x

(b) Round 2 game

	H_2	T_2
Up	$\frac{1}{r(x)}$	0
Down	1	1

Again, the argument for this turns on reflections about conditional questions. If Chooser is asked before the first coin flip, they are being asked what they want to do if the first coin lands Heads; if they are asked after that flip, they are being asked what they want to do now that the first coin landed Heads. These questions should get the same answer. I'll show that REU-maximisation only gets that result if r is the identity function, i.e., if REU-maximisation just is expected utility maximisation.

As before, I'll refer to Chooser's Early and Late choices, meaning their choices before and after being told the result of the first coin. I'll write $REU_E(X)$ to be the risk-weighted expected utility of X before finding out the result of the first coin toss, and $REU_L(X)$ to be the risk-weighted expected utility of X after finding out the result of the first coin toss. So Exit Principle essentially becomes this biconditional, for any gambles X and Y.

$$REU_E(X) \geq REU_E(Y) \leftrightarrow REU_L(X) \geq REU_L(Y)$$

I'll first prove that this implies that r must be multiplicative, i.e., that $r(xy) = r(x)r(y)$ for all x, y. This isn't a particularly problematic result; the most intuitive values for r, like $r(x) = x^2$, are multiplicative. Consider the Exit Problem shown in Table 11.2, where x and y are arbitrary.

It's easy to check that $REU_L(U) = REU_L(D) = 1$. So by Exit Principle, $REU_E(U) = REU_E(D)$. Since $REU_E(U) = \frac{r(xy)}{r(x)}$, and $REU_L(D) = r(y)$, it follows that $r(xy) = r(x)r(y)$, as required.

Define m, for midpoint, as $r^{-1}(0.5)$. Intuitively, m is the probability where the risk-weighting agent is indifferent between taking and declining a bet that stands to win and lose the same amount. Since r is monotonically increasing, and goes from 0 to 1, m must exist. Consider now Table 11.3, where y is arbitrary.

Table 11.3: An exit game with exit payout 1.

(a) Exit Parameters

Exit Payout	1
$\Pr(H_1)$	y
$\Pr(H_2)$	m

(b) Round 2 game

	H_2	T_2
Up	2	0
Down	1	1

Table 11.4: An exit game with exit payout 2.

(a) Exit Parameters

Exit Payout	2
$\Pr(H_1)$	y
$\Pr(H_2)$	m

(b) Round 2 game

	H_2	T_2
Up	2	0
Down	1	1

In Table 11.3, it's also easy to see that $REU_L(U) = REU_L(D) = 1$. So by Exit Principle, $REU_E(U) = REU_E(D)$. And it's also clear that $REU_E(D) = 1$, since that's the only possible payout for Down. So $REU_E(U) = 1$. So we get the following result.

$$\begin{aligned} REU_E(U) &= r(1 - y(1 - m)) + r(ym) \\ &= 1 \\ \therefore r(1 - y(1 - m)) &= r(ym) \end{aligned}$$

That doesn't look like a particularly notable result, but it will become useful when we discuss our last case, Table 11.4, which is just the same as Table 11.3, except the exit payout is now 2.

In Table 11.4, it's again easy to see that $REU_L(U) = REU_L(D) = 1$. So by Exit Principle, $REU_E(U) = REU_E(D)$. But the early values are more complicated. $REU_E(D) = 1 + r(1 - y)$, and $REU_E(U) = 2r(1 - y(1 - m))$. Using what we've discovered so far, we can do something with that last value.

$$
\begin{aligned}
REU_E(U) &= 2r(1-y(1-m)) & \\
&= 2(1-r(ym)) & \text{from previous calculations} \\
&= 2-2r(ym) & \\
&= 2-2r(y)r(m) & \text{since } r \text{ is multiplicative} \\
&= 2-r(y) & \text{since } r(m)=0.5
\end{aligned}
$$

Putting all this together, we get

$$
\begin{aligned}
REU_E(D) &= REU_E(U) & \Rightarrow \\
1+r(1-y) &= 2-r(y) & \Rightarrow \\
r(y)+r(1-y) &= 1 &
\end{aligned}
$$

So r is a monotonic increasing function satisfying $r(0)=0, r(1)=1, r(xy)=r(x)r(y)$ and $r(y)+r(1-y)=1$. The only such function is $r(x)=x$. So the only version of risk-weighted expected utility theory that satisfies Exit Principle is where $r(x)=x$, i.e., where risk-weighted expected utility just is old-fashioned expected utility.

This doesn't yet prove expectationism. I haven't shown that there is no other alternative to expected utility theory that satisfies Exit Principle. There are such other theories out there, such as the Weighted-linear utility theory described by Bottomley and WIlliamson (n.d.). But it's a guide to how we could start defending expectationism in a way consistent with how we handle decision problems involving demons.

References

Ahmed, Arif. 2012. "Push the Button." *Philosophy of Science* 79 (3): 386–95. https://doi.org/10.1086/666065.

———. 2020. "Equal Opportunities in Newcomb's Problem and Elsewhere." *Mind* 129 (515): 867–86. https://doi.org/10.1093/mind/fzz073.

Akerlof, George. 1970. "The Market for "Lemons": Quality Uncertainty and the Market Mechanism." *Quarterly Journal of Economics* 84 (3): 488–500. https://doi.org/10.2307/1879431.

Alcoba, Natalie. 2023. "In Argentina, Inflation Passes 100% (and the Restaurants Are Packed)." *The New York Times*, June 19, 2023. https://www.nytimes.com/2023/06/19/world/americas/argentina-inflation-peso-restaurants.html.

Allais, M. 1953. "Le Comportement de l'homme Rationnel Devant Le Risque: Critique Des Postulats Et Axiomes de l'ecole Americaine." *Econometrica* 21 (4): 503–46. https://doi.org/10.2307/1907921.

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.

Barnett, David James. 2022. "Graded Ratifiability." *Journal of Philosophy* 119 (2): 57–88. https://doi.org/10.5840/jphil202211925.

Ben-Porath, Elchanan, and Eddie Dekel. 1992. "Signaling Future Actions and the Potential for Sacrifice." *Journal of Economic Theory* 57 (1): 36–51. https://doi.org/10.1016/S0022-0531(05)80039-0.

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

Bottomley, Christopher, and Timothy Luke WIlliamson. n.d. "Rational Risk-Aversion: Good Things Come to Those Who Weight." Philosophy and Phenomenological Research. https://doi.org/doi.org/10.1111/phpr.13006.

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

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

Cho, In-Koo, and David M. Kreps. 1987. "Signalling Games and Stable Equilibria."

The Quarterly Journal of Economics 102 (2): 179–221. https://doi.org/10.2307/1885060.

Davey, Kevin. 2011. "Idealizations and Contextualism in Physics." *Philosophy of Science* 78 (1): 16–38. https://doi.org/10.1086/658093.

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.

Eyster, Erik, and Matthew Rabin. 2005. "Cursed Equilibrium." *Econometrica* 73 (5): 1623–72. 10.1111/j.1468-0262.2005.00631.x.

Fusco, Melissa. n.d. "Absolution of a Causal Decision Theorist." Noûs. https://doi.org/10.1111/nous.12459.

Gallow, J. Dmitri. 2020. "The Causal Decision Theorist's Gudie to Managing the News." *The Journal of Philosophy* 117 (3): 117–49. https://doi.org/10.5840/jphil2020117 39.

———. n.d. "The Sure Thing Principle Leads to Instability." Philosophical Quarterly. https://philpapers.org/archive/GALTST-2.pdf.

Goodsell, Zachary. n.d. "Decision Theory Unbound." Noûs. https://doi.org/10.1111/nous.12473.

Grant, Simon, Guerdjikova Ani, and John Quiggin. 2021. "Ambiguity and Awareness: A Coherent Multiple Priors Model." *The B.E. Journal of Theoretical Economics* 21 (2): 571–612. https://doi.org/10.1515/bejte-2018-0185.

Gustafsson, Johan E. 2011. "A Note in Defence of Ratificationism." *Erkenntnis* 75 (1): 147–50. https://doi.org/10.1007/s10670-010-9267-6.

Hare, Caspar, and Brian Hedden. 2015. "Self-Reinforcing and Self-Frustrating Decisions." *Noûs* 50 (3): 604–28. https://doi.org/10.1111/nous.12094.

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

———. 1988. "Causal Decision Theory and Game Theory: A Classic Argument for Equilibrium Solutions, a Defense of Weak Equilibria, and a New Problem for the Normal Form Representation." In *Causation in Decision, Belief Change, and Statistics: Proceedings of the Irvine Conference on Probability and Causation*, edited by William Harper and Brian Skyrms, 25–48. Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-009-2865-7_2.

Heinzelmann, Nora. n.d. "Rationality Is Not Coherence." Philosophical Quarterly. https://doi.org/10.1093/pq/pqac083.

Jackson, Frank, and Robert Pargetter. 1986. "Oughts, Options, and Actualism."

Philosophical Review 95 (2): 233–55. https://doi.org/10.2307/2185591.
Jeffrey, Richard. 1983. "Bayesianism with a Human Face." In *Testing Scientific Theories*, edited by J. Earman (ed.). Minneapolis: University of Minnesota Press.
Lee, Wooram. n.d. "What Is Structural Rationality?" Philosophical Quarterly. https://doi.org/10.1093/pq/pqad072.
Levinstein, Benjamin Anders, and Nate Soares. 2020. "Cheating Death in Damascus." *Journal of Philosophy* 117 (5): 237–66. https://doi.org/10.5840/jphil2020117516.
Lewis, David. 1979. "Prisoners' Dilemma Is a Newcomb Problem." *Philosophy and Public Affairs* 8 (3): 235–40.
———. 1981. "Why Ain'cha Rich?" *Noûs* 15 (3): 377–80. https://doi.org/10.2307/2215439.
———. 2020. "Letter to Jonathan Gorman, 19 April 1989." In *Philosophical Letters of David K. Lewis*, edited by Helen Beebee and A. R. J. Fisher, 2:472–73. Oxford: Oxford University Press.
Lipsey, R. G., and Kelvin Lancaster. 1956. "The General Theory of Second Best." *Review of Economic Studies* 24 (1): 11–32. https://doi.org/10.2307/2296233.
McClennan, Edward. 1990. *Rationality and Dynamic Choice*. Cambridge: Cambridge University Press.
Myerson, R. B. 1978. "Refinements of the Nash Equilibrium Concept." *International Journal of Game Theory* 7 (2): 73–80. https://doi.org/10.1007/BF01753236.
Nash, John. 1951. "Non-Cooperative Games." *Annals of Mathematics* 54 (2): 286–95.
Nozick, Robert. 1969. "Newcomb's Problem and Two Principles of Choice." In *Essays in Honor of Carl G. Hempel: A Tribute on the Occasion of His Sixty-Fifth Birthday. Hempel: A Tribute on the Occasion of His Sixty-Fifth Birthday*, edited by Nicholas Rescher, 114–46. Riedel: Springer.
Podgorski, Aberlard. 2022. "Tournament Decision Theory." *Noûs* 56 (1): 176–203. 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.
Ramsey, Frank. 1990. "General Propositions and Causality." In *Philosophical Papers*, edited by D. H. Mellor, 145–63. Cambridge: Cambridge University Press.
Richter, Reed. 1984. "Rationality Revisited." *Australasian Journal of Philosophy* 62 (4): 393–404. https://doi.org/10.1080/00048408412341601.
Robinson, Julia. 1949. "On the Hamiltonian Game (a Traveling Salesman Prob-

lem)." Santa Monica, CA: The RAND Corporation.
Selten, R. 1975. "Reexamination of the Perfectness Concept for Equilibrium Points in Extensive Games." *International Journal of Game Theory* 4 (1): 25–55. https://doi.org/10.1007/BF01766400.
Selten, Reinhard. 1965. "Spieltheoretische Behandlung Eines Oligopolmodells Mit Nachfrageträgheit." *Zeitschrift für Die Gesamte Staatswissenschaft* 121 (2): 301–24.
Shafer, Glenn. 1976. *A Mathematical Theory of Evidence*. Princeton: Princeton University Press.
Skyrms, Brian. 1984. *Pragmatics and Empiricism*. New Haven, CT: Yale University Press.
———. 2004. *The Stag Hunt and the Evolution of Social Structure*. Cambridge: Cambridge University Press.
Spencer, Jack. 2021a. "An Argument Against Causal Decision Theory." *Analysis* 81 (1): 52–61. https://doi.org/10.1093/analys/anaa037.
———. 2021b. "Rational Monism and Rational Pluralism." *Philosophical Studies* 178: 1769–1800. https://doi.org/10.1007/s11098-020-01509-9.
———. 2023. "Can It Be Irrational to Knowingly Choose the Best?" *Australasian Journal of Philosophy* 101 (1): 128–39. https://doi.org/10.1080/00048402.2021.1958880.
Spencer, Jack, and Ian Wells. 2019. "Why Take Both Boxes?" *Philosophy and Phenomenological Research* 99 (1): 27–48. https://doi.org/10.1111/phpr.12466.
Stalnaker, Robert. 1998. "Belief Revision in Games: Forward and Backward Induction." *Mathematical Social Sciences* 36 (1): 31–56. https://doi.org/10.1016/S0165-4896(98)00007-9.
———. 2008. *Our Knowledge of the Internal World*. Oxford: Oxford University Press.
Strevens, Michael. 2008. *Depth: An Account of Scientific Explanations*. Cambridge, MA: Harvard University Press.
Sutton, John. 2000. *Marshall's Tendencies: What Can Economists Know?* Cambridge, MA: MIT Press.
Thoma, Johanna. 2019. "Risk Aversion and the Long Run." *Ethics* 129 (2): 230–53. https://doi.org/10.1086/699256.
Wedgwood, Ralph. 2013. "Gandalf's Solution to the Newcomb Problem." *Synthese* 190 (14): 2643–75. https://doi.org/10.1007/s11229-011-9900-1.
Weirich, Paul. 1985. "Decision Instability." *Australasian Journal of Philosophy* 63 (4): 465–72. https://doi.org/10.1080/00048408512342061.

Wells, Ian. 2019. "Equal Opportunity and Newcomb's Problem." *Mind* 128 (510): 429–57. https://doi.org/10.1093/mind/fzx018.
Wilson, Robert B. 1967. "Competitive Bidding with Asymmetric Information." *Management Science* 13 (11): 816–20. https://doi.org/10.1287/mnsc.13.11.816.
Worsnip, Alex. 2021. *Fitting Things Together: Coherence and the Demands of Structural Rationality*. Oxford: Oxford University Press.
Wright, Crispin. 2021. *The Riddle of Vagueness: Selected Essays 1975-2020*. Oxford: Oxford University Press.