Game Theory as Decision Theory

Brian Weatherson

2023-08-31

Table of contents

Preface

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

1 Introduction

1.1 Ten Features of a Good Decision Theory

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 an interesting and plausible theory, which I'll call Gamified Decision Theory (GDT), from these textbooks. There are ten characteristics of GDT (as I'll understand it) that I will focus on. I'll quickly list them here, then the bulk of the book will consist of a chapter describing and motivating each of the ten 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. **Selection**: The aim of decision theory is to generate a function from possible choices to choice-worthy options, not to generate a preference ordering over the options.
9. **Substantive Probability**; the ideal decider has rational credences.
10. **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 ten of these. 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 textbooks endorse 9 and 10, 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 to 8. 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).

1.2 Demons

A lot of contemporary philosophical decision theory revolves around what to do if there is a certain kind of demon around. Following Nozick (1969), such a demon is typically taken to be arbitrarily good at predicting what a human deliberater will do. I'll call our arbitrary deliberater Chooser, and whenever X is a choice Chooser can make, I'll use PX to mean that the demon predicts Chooser makes that choice. It's not so common to have problems where there are two such demons around, but I'll make heavy use of them, and in such cases I'll be clear about whether PX means that the first or the second demon predicted that Chooser will do X. These are predictions, and we assume that causation runs from past to future, so what Chooser does has no causal impact on what Demon predicts.

I'm squeamish about assigning probability 1 to predictions that are causally isolated from the thing being predicted; I have reductionist enough views about causation to think that if a prediction is correct with probability 1, that raises questions about whether causation does really run from past to future in this case. So I prefer to say that the Demon is correct with a probability close enough to 1 that it doesn't matter for the purposes of the problem being analysed. But this squeamishness, and the

associated reductionism about causation, is not part of GDT. If you're happy with having causally isolated Demons who are correct with probability 1, everything else I say should be acceptable. Indeed, some of the reasoning goes through even more smoothly with perfectly accurate, but causally isolated, Demons.

A generic binary choice problem involving Chooser and Demon looks like this.

Table 1.1: The demonic decision problem generated by a generic symmetric game.

	PA	PB
A	x	y
B	z	w

Chooser selects A or B, Demon predicts the choice, and there are four possible outcomes. I'll assume that the value of these outcomes can be measured numerically, with greater numbers being better. We'll come back to this assumption briefly in Chapter 2, and more substantively in Chapter 3. Following Nozick (1969), the most common problem that people discuss involving Demon is what Nozick dubbed "Newcomb's Problem", after the physicist who suggested the problem to him. A Newcomb problem is an instance of Table A.3 satisfying the following constraints.

- $z > x$
- $w > y$
- $x \gg w$

The standard example uses (more or less) the following values, but all that really matters are the three inequalities above.

Table 1.2: Newcomb's Problem.

	PA	PB
A	1000	0
B	1001	1

Option A and B are typically called 'one-boxing' and 'two-boxing' respectively, because they involve selecting either one or two boxes in the vignette Nozick gives to

go along with the story. But what really matters is the schematic form, not the details of the physical setup.

Nozick distinguishes two approaches to this problem you might take. He doesn't use the following terms, but they quickly became identified as Evidential Decision Theory, and Causal Decision Theory. Evidential Decision Theory (EDT) says that one should first assign values to each option using the following formulae. I'll just give the formulae for the case where there are two states of the world, PA and PB, but it should be clear how to generalise this to the case where there are m possible states. When X is a choice and Y a state, I'll use V(XY) to mean the value of choosing X in state Y. So for example in Newcomb's Problem, V(BPA) = 1001; if Chooser selects B and Demon predicts A, Chooser's payout is 1001. And I'll use Pr(Y | X) to mean the probability of being in state Y conditional on choosing X. Using this terminology, EDT says that the value of the choices is:

$$\begin{aligned} V(A) &= V(APA) \cdot Pr(PA \mid A) + V(APB) \cdot Pr(PB \mid A) \\ V(B) &= V(BPA) \cdot Pr(PA \mid B) + V(BPB) \cdot Pr(PB \mid B) \end{aligned}$$

So in Newcomb's Problem, if the Demon is, say, 90% reliable, we have:

$$\begin{aligned} V(A) &= 1000 \cdot 0.9 + 0 \cdot 0.1 = 900 \\ V(B) &= 1001 \cdot 0.1 + 1 \cdot 0.9 = 101 \end{aligned}$$

Then EDT says that higher valued options are better, so A is better than B, since $900 > 101$. And if the Demon is even more reliable than 90%, that gap just grows further.

Causal Decision Theory (CDT), on the other hand, is moved by the following argument. Whatever Demon has predicted, Chooser is better off choosing B than A. That, says CDT, settles things; Chooser should take option B. I think this is right; Chooser should choose B, and they should do so for this reason. But note that this is not anything like a complete theory of choice. Two people could agree with this little argument and have any number of different views about problems that not so easily disposed of. In this book, especially in Chapter 7, we'll spend a lot of time on problems like the following.

Table 1.5: The Stag Decision.[1]

	PA	PB
A	6	0
B	5	2

It turns out that among people who endorse the little argument for choosing B in Table 1.2, there are at least four distinct views about what to do in Table 1.5.

1. Frank Arntzenius (2008) and Johan E. Gustafsson (2011) recommend Choosing A.
2. Ralph Wedgwood (2012), Dmitri Gallow (2020), Abelard Podgorski (2022), and David Barnett (2022) recommend choosing B.
3. James Joyce (2012) says that what Chooser should do is a function of Chooser's probability distribution over their choices prior to deliberating about what to do.
4. Jack Spencer (2021b) and Melissa Fusco (n.d.) say that Chooser can rationally take either option.

I'm going to side with option 4. Though note that Spencer and Fusco disagree about what Chooser should do in several other cases, most notably in cases like Table 1.5 but with the payouts inverted, and GDT is going to agree more with Fusco than Spencer.

But I'm not going to argue for option 4, let alone my preferred version of option 4, or against any other options, just yet. Rather, I want to start with a terminological point. It's not obvious, either from the description of the problems or the history of the philosophical discussion, which if any of these theories should get the name "Causal Decision Theory". Some people write as if Joyce's view is the unique one that should get that name; indeed many of the people I've listed above describe themselves as critics of CDT who are offering an alternative to it. I think that's not the most helpful way to classify views. All of them accept that in Newcomb's Problem, Chooser should choose option B, and that Chooser should choose it because Chooser can't make a causal difference to whether PA or PB happens, and either way, B is better than A. That's the core idea behind Causal Decision Theory.

[1] I say much more about why the problem has this label in Appendix A.

A decision theory should say what to do not just in one problem, but across a family of problems. It should say what to do in Table 1.5, for example. As I'm using the term, Causal Decision Theory, as such, is neutral between the four possible approaches to Table 1.5. So it isn't a theory. Rather, it is a family of theories, that all agree about what to do in Newcomb's Problem, and about why to do it, but disagree in different problems.

So as I'm using the term, Causal Decision Theory is not a theory. That might be surprising, since it has the word 'Theory' in the name. But we're used to things like the United States of America which includes parts that are neither States nor in America (e.g., Guam). We can live with Causal Decision Theory not being a theory, and instead being a family that agree about what to do, and why to do it, in Newcomb's Problem. The bulk of this book will be an in house dispute between causal decision theories, though I'll spend some time objecting to EDT, and also some time objecting to other theories that reject both CDT and EDT.[2]

1.3 Gamified Decision Theory

The actual theory I will defend, GDT, is a version of what's sometimes called causal ratificationism.

The 'causal' in causal ratificationism means that there are constraints on the proper formulation of a decision problem. EDT says it does not matter how we divide the world into states; decision theory should give the same verdict. If we rewrite Newcomb's Problem with the states being that Demon predicted correctly, and that Demon predicted incorrectly, EDT gives the same recommendation, for essentially the same reason. GDT, like all causal theories, rejects this. The correct formulation of a decision problem requires that the states, like PA and PB, be causally independent of the choices that Chooser makes. I have a fairly strong version of this independence constraint, which I'll discuss more in Chapter 4.

The 'ratificationism' in causal ratificationism means that Chooser will ratify their choice once they make it, i.e., that Chooser will not regret a rational choice as soon

[2]The most notable of these will be the Functional Decision Theory of Levinstein and Soares (2020), and the non-expectationist theories of Quiggin (1982) and Buchak (2014).

as it is made. Formally, this means that Chooser will only choose A in cases like Table A.3 if the following inequality holds.[3]

$$V(APA) \cdot Pr_A(PA) + V(APB) \cdot Pr_A(PB) \geq V(BPA) \cdot Pr_A(PA) + V(BPB) \cdot Pr_A(PB)$$

By Pr_A I mean the rational probabilities that Chooser has after choosing A. If there is more than one rational probability that Chooser could have, all that matters is that the inequality hold for one such probability function.[4] In somewhat technical English, what this inequality says is that once A is chosen, the expected value of choosing A is at least as great as the expected value of having chosen B. That's what I mean by ratifiability; once Chooser selects A, they think it was for the best (or at least equal best) that they chose it.

I'm far from the first to endorse ratifiability as a constraint on decisions. It's defended by William Harper (1986), in a paper that was a central inspiration for this project, both because of its conclusions, and because of the way it connected decision theory to game theory. I'll talk about the ratifiability constraint much more in Chapter 6.

GDT, as I'm defining it, has three extra features beyond this causal ratification constraint, and I'll end this chapter with a brief discussion of each of them.

GDT says that permissible choices are not weakly dominated. An option weakly dominates another if it could be better, and couldn't be worse. So in Table 1.7, A is not a permissible choice because it is weakly dominated by B.

Table 1.7: An example of weak dominance.

	PA	PB
A	2	0

[3] In general, the sum on each side of the inequality ranges over all possible states, so if there are more than two states, there will be more than two summands on either side. And A must be ratified compared to all alternatives, so if there are more than two options, this inequality must hold if you replace B with C, D, or any other choice.

[4] If there is more than one alternative to A, and more than one rational probability function, the rule is that there is some probability function such that A does better than every possible alternative, if we put that function into the inequality above. It's not enough that for each alternative there is some probability function that judges A to be better than the alternative.

	PA	PB
B	2	1

Since B could be better than A, if Demon predicted B, and could not be worse than A, at worst they produce the same outcome if Demon predicts A, B weakly dominates A. And weakly dominated actions are not rational choices. So in this problem the only rational choice is B. This is not particularly intuitive, but I don't think agreement with first pass intuition is a particularly strong constraint on decision theories, for reasons I'll go over in Chapter 3. And I'll have much more to say about weak dominance, and in particular why I reject an iterated version of the weak dominance constraint, in Chapter 11.

In dynamic choices, GDT says that Chooser must satisfy two constraints. First, the plan they make for what to do over time, what we'll call a strategy, must be a permissible choice of strategy. Second, at each point in time, they must choose an option that would be permissible were the dynamic choice problem to have started at that point, with that set of options. These two constraints, which I'll discuss much more in chapter Chapter 8, have some surprising consequences. Imagine that Chooser has the following two-stage problem. At stage 1, they can choose to Exit or Continue. If they Exit, they get 4. If they continue, they make a choice in the following Demonic problem.[5]

Table 1.8: The second stage of a dynamic problem.

	PA	PB
A	3	3
B	0	5

The plan of Continuing, then choosing A, is not a sensible plan. Chooser knows from the start that there is a plan which is guaranteed to be better, namely Exiting. If they Exit, they are guaranteed to get 4, if they Continue then choose A, they are guaranteed to get 3. So they may not Continue then choose A. But this does not mean that they must Exit. They may Continue and choose B. Now here's the surprising

[5] In this problem, the Demon makes a prediction after Chooser opts to Continue, but this prediction is only revealed after Chooser selects A or B.

part. If they faced Table 1.8 as the first choice they have to make, they could choose B, but they also could choose A. In Table 1.8, A is ratifiable and not weakly dominated. So GDT is not a purely forward looking decision theory. By that I mean that sometimes, choices that Chooser makes earlier in a dynamic choice situation constrain which choices are rational later in the game. A lot of versions of CDT do not specify how they are to be extended into theories of dynamic choice, but my impression is that many philosophers do think decision theory should be purely forward looking, and GDT disagrees with them on this point.

There are some decision theorists who agree with GDT that decisions should not be strictly forward-looking. These include the resolute theorists, in the sense of McClennan (1990), and the functional theorists, in the sense of Levinstein and Soares (2020). But GDT disagrees with them as well. Those theorists think that the only thing Chooser must do is choose a sensible plan, and then at each stage Chooser should just carry it out. Here is a case where GDT disagrees with them. It will be a three stage game, and the role of Demon will be somewhat different to their role in previous examples.

1. First, Chooser chooses to scratch their ear or not scratch. The choice is revealed to Demon, but it makes no difference to anyone's payouts.
2. Second, Demon predicts whether Chooser will select A, B or C at stage 3. Demon's prediction is partially revealed to Chooser. If Demon predicts C, Chooser is told this. If Demon predicts A or B, Chooser is just told that Demon did not predict C.
3. Chooser selects A or B, knowing what Demon predicted. And then Chooser's payouts are given by Table 1.9.

Table 1.9: Payouts in the three stage game.

	PA	PB	PC
A	2	0	3
B	2	1	0
C	0	1	1

The strategy of scratching one's ear, then choosing AB whatever Demon announces, is a permissible choice of strategy. By that, I mean that if Chooser were able (counterfactually) playing a game where they just announced a strategy and it was carried out

automatically, the strategy scratch then play A whatever happens, would be permissible.[6] In that strategic form of the game, this strategy is ratifiable and not weakly dominated. But in the actual dynamic game Chooser is playing, GDT says that it is not permissible. After all, were Chooser to learn that Demon did not predict C, then Chooser would be back in Table 1.7, and in that example A is impermissible.

I really don't think this example will motivate people to prefer GDT either to its forward looking alternatives, or to extant backward looking theories like resolute choice or functional decision theory. I'm not sure what intuition says about this puzzle, but I really don't think it says that choosing B if Demon announces they have not predicted C is the only rational alternative. That, however, is what GDT says; the only rational choice here is to do whatever one likes about scratching or not scratching, then choose A if Demon predicts C, and B if Demon does not predict C.

The point of this chapter is to describe GDT, not argue for it. And the point of this example is to show that GDT differs from theories like resolute or functional choice, which say that choosing A whatever one learns from Demon is at least permissible, and perhaps mandatory.

Finally, my version of GDT says that what matters for rational choice is what probabilities over states are rational, not which probabilities Chooser happens to endorse. GDT is a theory of rational choice simpliciter, not a theory of rational choice given possibly irrational beliefs. I'll have more to say about this in Chapter 10.

[6] So would not scratch then play A whatever Demon announces. The scratching is just there to make it clear that Chooser has a choice to make before Demon makes any prediction. I'll mostly drop this device in later examples, and just stipulate that the dynamic problem may start before Chooser's first choice.

2 Idealised

2.1 Introducing Ideal Theory

Game theorists, like philosophical decision theorists, are doing ideal theory. To see that they are doing ideal theory, compare what they say about two problems: Salesman and Basketball. The first is a version of what Julia Robinson dubbed the 'travelling salesman' problem.[1]

> **Salesman** Chooser is given the straight line distance between each pair of cities from the 257 represented on the map below. Using this information, Chooser has to find as short a path as possible that goes through all 257 cities and returns to the first one. The longer a path Chooser selects, the worse things will be for Chooser.

```
Loading required package: tidyverse
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.1     v readr     2.1.4
v forcats   1.0.0     v stringr   1.5.0
v ggplot2   3.4.2     v tibble    3.2.1
v lubridate 1.9.2     v tidyr     1.3.0
v purrr     1.0.1
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
```

[1] The dubbing is in Robinson (1949). For a thorough history of the problem, see Schrijver (2005). For an accessible history of the problem, which includes these references, see the wikipedia page on 'Traveling Salesman Problem'.

```
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflic
Loading required package: TSP

Loading required package: maps


Attaching package: 'maps'


The following object is masked from 'package:purrr':

    map
```

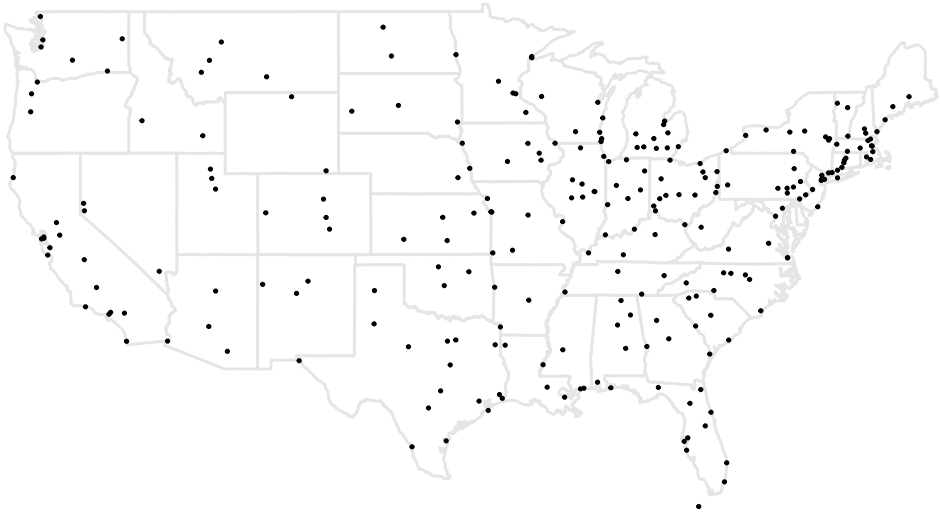

Figure 2.1: The 257 cities that must be visited in the Salesman problem.

Since there are 256! possible paths, and $256! \approx 10^{727}$, Chooser has a few options here.[2] Game theorists, and philosophical decision theorists, start with the assumption that the people in their models can solve these problems in zero time and at zero cost. That's not even approximately true for any actual person without technological

[2]The 256 cities are the cities in the lower 48 states from the 312 cities in North America that John Burkardt mapped in his dataset Cities, available at people.sc.fsu.edu/~jburkardt/datasets/cities/cities.html.

assistance. Even with knowledge of the problem and a good computer, there are not that many actual people who you could properly model as being able to solve it in zero time and at zero cost.

The question of how to think about people who do have to spend time and resources to solve a problem like this is an interesting one. We might call that problem one in *non-ideal decision theory*.[3] I won't say much about non-ideal decision theory in the body of this book, though I'll come back to it in Appendix B. What I mostly want to do now is use Salesman to say something about what the difference between ideal and non-ideal theory is. And that difference is brought up vividly by the following problem.

> **Basketball** Chooser is at a casino, and a basketball game is about to start. Chooser knows that basketball games don't have draws or ties; one side will win. And Chooser knows the teams are equally balanced; each team is 50% likely to win. Chooser has three options. They can bet on the Home team to win, bet on the Away team to win, or Pass, and not bet. If they bet, they win $100 if the team they bet on wins, and lose $110 if that team loses. If they Pass, they neither gain nor lose anything.

Ideal decision theory says that in Basketball, Chooser should Pass. That's not the optimal outcome for Chooser. The optimal outcome is that they bet on the winning team. But since they don't know who that is, and either bet will, on average, lose them money, they should Pass rather than bet on Home or Away. We could have a theory that just evaluated the possible outcomes in any decision. I'll call this Outcome Evaluation Theory. Contrast this with two other theories. Game theory says that the ideal agent chooses the shortest route, whatever it is, in Salesman and does not bet in Basketball. If an ordinary reasonable person was advising a friend facing these two problems, they would give the same advice as the game theorist about Basketball, but in Salesman they would not simply say *Choose the shortest path!*, since that's useless advice. Rather they would suggest something about how to solve the problem, possibly by looking up strategies.

[3] I'm borrowing the term 'non-ideal' from work in political philosophy. See Valentini (2012) for a good survey of some relevant work, and Mills (2005) for an important critique of the centrality of ideal theory in political philosophy. Critics of ideal theory, such as Mills, and Sen (2006), argue that we shouldn't base non-ideal theory on ideal theory. I'm going to agree, but my focus is primarily in the other direction. I'm going to argue that it isn't a constraint on ideal theory that it is useful in constructing a non-ideal theory.

So we have three theories on the table: Outcome Evaluation Theory; the game theory approach, which I'll call Ideal Decision Theory; and the ordinary reasonable person approach, which I'll call Non-Ideal Decision Theory. We can distinguish these three theories by what they say to do in two examples introduced so far: Salesman and Basketball.

Table 2.1: How three kinds of theories handle two problems.

Theory	**Salesman**	**Basketball**
Outcome Evaluation	Shortest route	Bet on winner
Ideal Decision	Shortest route	Pass
Non-Ideal Decision	Study optimization	Pass

Game theory agrees with the middle row. GDT, the theory I'm developing in this book, does so too. And so do almost all decision theorists working in philosophy.[4] So in trying to convince philosophers to adopt GDT, I'm not asking them to change their view on this point. But still, this is odd. What is the benefit of a theory of decision that does not produce the best outcomes, and does not produce useful, reasonable advice?

We could say that if Chooser were ideal, they would agree with Ideal Decision Theory. But why we should care about what would have if Chooser were ideal, since Chooser is not in fact ideal? 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 announce the best answer they have after spending no time calculating the solution to Salesman, and resembling the ideal agent in that respect will make Chooser worse.[5] And there is a separate problem. Why say it is ideal to make a choice in Basketball that Chooser knows will lead to a sub-optimal outcome? We can make progress on both these problems, what it means to say something is

[4]The exceptions are people working in 'descriptive decision theory' (Chandler 2017). But that's normally not taken to be a normative theory in any respect; it isn't about what people should do in problems like Salesman, but what they actually do.

[5]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.

ideal, and why we should care about the ideal, but stepping back and asking what we even mean by 'ideal', and 'idealisation'.

In philosophy, it turns out 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.[6]

The best way to understand game theorists, and most philosophical decision 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.

2.2 Uses of Ideal Theory

Still, this approach raises two pressing questions. One is why we should be interested in a model that is so idealised. The other is why we don't idealise even further, idealising away from informational limitations as well as computational ones.[7]

All social sciences use idealised models of some kind or other. The fact that real humans can't solve problems like Salesman, but the modeled humans can, isn't in itself a problem. It might be in some cases, like if you are giving a model of when humans fail at maximisation problems, but in itself it isn't a problem. The real

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

[7] John Conlisk (1996) stresses that explaining the asymmetry here is a big part of the challenge. That paper had a big influence in how I'm thinking about the problem, and several of the citations below are from it.

challenge is that the idealisation is useless. If all we end up saying is that when it's more likely to rain, more people take umbrellas, we don't need books full of math to say that. Here's how Keynes puts the complaint, in a closely related context.

> But this *long run* is a misleading guide to current affairs. *In the long run we are all dead*. Economists set themselves too easy, too useless a task if in tempestuous seasons they can only tell us that when the storm is long past the ocean will be flat again. (Keynes 1923, 80, emphasis in original)

Don't focus on the temporal connotations of Keynes's terminology of 'long run'. What's characteristic of his long run is not that it takes place in the distant future. What is characteristic of it instead is that it takes place in a world where some sources of interference are absent. It's a world where we sail but there are no storms. It's a study where we abstract away from storms and other unfortunate complications. And that's what's characteristic of Ideal Decision Theory. We know that people cannot easily solve hard arithmetic problems, but we abstract away from that fact. Does this leave the resulting theory "easy and useless"?

To see that it's not "easy", it simply suffices to take a casual glance at any economics journal. But what about Keynes's suggestion that it is "useless"? It turns out there are some surprising results that we need the details of something like GDT to generate. One nice case of this is the discussion of Gulf of Mexico oil leases in Wilson (1967).[8] Another 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) 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

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

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

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.)

And here's a simpler surprising prediction that you need something like GDT to get, and which is relevant to some debates in philosophical decision theory.[9] 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[10]. Simple rules like "When behaviour is rewarded, it happens more often" don't always work in strategic settings, and it takes some care to tell when they do work.

The point of decision theory is not to advise people on what to do in Rock-Paper-Scissors, or in Salesman. In each case, it would give bad advice. Really you should try to read your opponent's body shape for clues in Rock-Paper-Scissors, and find some good software in Salesman, neither of which the theory says. Rather, the point is be part of explanations like why there was such a large discount on used cars in the

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

[10] The proof is in Appendix C.

20th century, and why the bystander's gambit won't work in my modified version of Rock-Paper-Scissors. David Lewis gives a similar account of the purpose of decision theory in a letter to Hugh Mellor. The context of the letter, like the context of this section, is a discussion of why idealisations are useful in decision theory. Lewis writes,

> We're describing (one aspect of) what an ideally rational agent would do, and remarking that somehow we manage to approximate this, and perhaps – I'd play this down – advising people to approximate it a bit better if they can. (Lewis 2020a, 432)

2.3 Why This Idealisation

Still, there are a lot of ways to idealise away from the details of individual humans. Why do we delete the differences from rationality, and not the differences from full-informedness, or the differences from something that lacks normative significance? One simple answer is that people have used this idealisation and it has (to some extent) worked. But there is a little more to say.

Take some generalisation about human choosers that isn't particularly rational, is true in most but not all cases, and which it would simplify our description of various cases to say it holds in all cases. Why don't we use the idealisation that says it does in fact hold in all cases? The answer here depends a bit on the 'we'. Some generalisations about less than ideally rational behavior are useful in empirical studies of consumer choice.[11] But there is a reason that philosophers and more theoretical economists have focussed on rational idealisations. The thought is that a lot of deviations from rationality are short-cuts, that are sensible to use when the stakes are low. But in high-stakes situations, humans will more closely approximate ideally rational agents. (This might be coupled with the suggestion that over time they will do this better, simply because the ones who more closely approximate the rational choice will increase their market share.) And getting correct predictions and explanations in high-stakes cases might be particularly important in understanding society and the economy. So while non-rational idealisations might be crucially important in understanding store design (e.g., why supermarkets have produce at the entrance),

[11] Here's one relatively recent example, picked more or less at random - https://www.sciencedirect.com/science/article/pii/S0969698923000796.

rational idealisations are needed for understanding the nature of stock markets and business investment.

That reasoning looks like it might over-generate. In high-stakes cases, people are not only more careful with their decision making process, they are more careful about acquiring information before they decide. If our focus is high-stakes decision making, and I think it has to be to motivate rational idealisations, why don't we also abstract away from informational limitations of the deciders? After all, the decider will try to remove those limitations before deciding in these high-stakes cases. The answer is that in some cases, and these are the cases that decision theory is most useful in explaining, there are in principle reasons why the decider can't do anything about certain informational limitations. The information might be a fact about the result of a chance-like process that is unknowable either in principle, or in any practical way. Or there might be someone else who has just as much incentive to keep the information hidden as the decider has to seek it out. The latter is what happens when someone is selling a lemon, for example. I don't have anything like a proof of this, but I suspect that most uses of game theory or decision theory to explain real-world phenomena will fall into one or other of these categories: there are relevant facts that the decider can't know, either because they have to decide before decisive evidence is revealed, or because someone just as well resourced as them is determined to prevent them getting the information.

2.4 Two Bonus Uses

There are two other advantages to using the particular idealisation that game theorists and decision theorists have settled on, i.e., idealise away from computational but not informational limitations. The first can be seen from this famous quote from Frank Knight, an early proponent of the view of idealisations in decision theory that I'm endorsing here.

> It is evident that the rational thing to do is to be irrational, where deliberation and estimation cost more than they are worth. That this is very often true, and that men still oftener (perhaps) behave as if it were, does not vitiate economic reasoning to the extent that might be supposed. For these irrationalities (whether rational or irrational!) tend to offset each other. The applicability of the general "theory" of conduct to

> a particular individual in a particular case is likely to give results bordering on the grotesque, but *en masse* and in the long run it is not so. The *market* behaves *as if* men were wont to calculate with the utmost precision in making their choices. We live largely, of necessity, by rule and blindly; but the results approximate rationality fairly well on an average. (Knight 1921, 67n1)

I don't agree with everything Knight says here; I think he's much too quick to assume that deviations from rationality will "offset".[12] But that's something to be worked out on a case-by-case basis. We should not presuppose in advance either that the imperfections be irrelevant or that they will be decisive.

Despite that, I quoted Knight here because there is an important point to I do agree with, and deserves emphasis. If we don't act by first drawing Marshallian curves and solving optimisation problems, how do we act? As Knight says, we typically act "by rule". Our lives are governed, on day-by-day, minute-by-minute basis, by a series of rules we have internalised for how to act in various situations. The rules will typically have some kind of hierarchical structure - do this in this situation unless a particular exception arises, in which case do this other thing, unless of course a further exception arises, in which case, and so on. And the benefit of adopting rules with this structure is that they, typically, produce the best trade off between results and cognitive effort.

One useful role for idealised decision theory is in the testing and generation of these rules. We don't expect people who have to make split-second decisions to calculate expected utilities. But we can expect them to learn some simple heuristics, and we can expect theorists to use ideal decision theory to test whether those heuristics are right, or whether some other simple heuristic would be better. This kind of approach is very useful in sports, where athletes have to make decisions very fast, and there is enough repetition for theorists to calculate expected utilities with some precision. But it can be used in other parts of life, and it is a useful role for idealised decision theory alongside its roles in prediction and explanation.

The other benefit of idealised decision theory is that it has turned out to be theoretically fruitful in ways that I would never have expected. It turns out that sometimes one gets a powerful kind of explanation from very carefully working out the ideal theory, and then relaxing one of the components of the idealisation. At a very high

[12] See Conlisk (1996) for many, many examples from both theory and practice where they do not.

level of abstraction, that's what happened with the Eyster and Rabin's development of the notion of cursed equilibrium (Eyster and Rabin 2005). The explanations they give for certain kinds of behavior in auctions are completely different from anything I'd have expected, but they seem to do empirically fairly well.[13] Their 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. A priori, this is not very plausible. But if it fits the data, and it seems to, 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, that's not in fact how it was developed. In fact the development of certain kinds of ideal models[14] was theoretically fruitful in the understanding of very non-ideal behavior.

2.5 Summary

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, or the upsides of an alternative, 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, and they know they'll have that knowledge whatever choices they face. 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

[13] And there are even more empirically successful theories that build on their work, such as in Fong, Lin, and Palfrey (2023) and Cohen and Li (2023).

[14] The ideal models they use, which involve the notion of Bayesian Perfect Equilibrium, are slightly more complicated than any model I'll use in this book; they were not the simple models from the first day of decision theory class.

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. 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. And 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.

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. So if the states are H and T, the expected value of a choice X is V(XH)Pr(H) + V(XT)Pr(T), and the expectationist says to choose the choice with the

[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.

highest expected value.[2] Second, it says that something like this expected value plays an important role in the theory of abnormal decision problems. What's an important role is 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. I go over the details of this in Appendix D. The other reason is dialectical.

Expectationism does a surprisingly bad job at matching untutored intuition about cases. In some important sense, it says that risk-aversion is irrational.[3] But intuition says that risk-aversion is rational. To be sure, expectationism does have things to say here. There is something like risk-aversion which makes sense given the declining marginal utility of money.[4] And, I think, expectationism can explain why risk-aversion seems rational to people who primarily think about bets in terms of goods with declining marginal value. But still, the expectationist is playing defence here. It seems intuitively like risk-aversion is rational, and as Allais (1953) showed, this implies that expectationism says unintuitive things about some fairly simple cases.

Given that, it is surprising how many expectationists rely on intuitions about cases when assessing the merits of different theories of decision in abnormal cases. It

[2] Some theorists will put conditional probabilities in place of probabilities in this formula, so they'll use Pr(H|X) rather than Pr(H). But since we're restricting attention to normal problems, this is a distinction without a difference.

[3] More precisely, it says the following. If B is between A and C in value, and the difference in value between A and B equals the difference in value between B and C, then a rational agent will be indifferent between getting B for sure, and a 50/50 chance of getting A or C. The risk-averse agent, in the sense of risk-aversion at issue here, will prefer B.

[4] To modify the example of the previous footnote, if A, B and C are monetary rewards, and B is halfway between A and C in terms of its monetary value, then expectationism says that an agent can, and probably should, prefer B to a 50/50 chance of getting A or C.

often seems, when reading this part of the literature, that philosophical decision theorists endorse the following argument schema.

1. The correct decision theory is the one that best tracks intuitions about cases.
2. The decision theory that best tracks intuitions about cases is T (the preferred theory of the person making the argument).
3. Therefore, the correct decision theory is T.

If the philosopher is an expectationist, they can't really endorse this argument. Premise 2 can't possibly be true. No matter how well theory theory does at tracking intuitions about abnormal cases, a modified version of their theory that allows for risk-aversion will do an even better job, especially when normal cases are considered. And for that reason, they can't really believe premise 1. After all, if premise 1 is true, then the correct decision theory for normal decision problems is not expectationist.

So this means that arguments like this one should not be used in decision theory. Of course, we can't entirely depart from intuitions about cases. If our theory disagrees too much with common sense it starts becoming a theory of something else (Jackson 1998, Ch. 2). The role of intuitions is like the drawing on a wanted poster. It's not true that the criminal is the person who best resembles the drawing, but you should be very sceptical of a theory that the criminal looks nothing at all like the poster. Still, you should also be sceptical of a theory that the criminal is someone who lacks what we thought were necessary skills for committing the crime, even if the person who most resembles the poster lacks those skills. One aim of this book is to develop several plausible preconditions on a good theory of decision, most importantly the Exit Principle of Chapter 7, and argue that only GDT (or something like it) satisfies those preconditions. That's more philosophically significant than whether GDT best tracks intuitions about cases.

If expectationism does not maximise agreement with intuition, why is it so popular? It is because there are several plausible principles that are consistent with expectationism, but which are not consistent with the best alternative, namely the Quiggin-Buchak theory. These include[5]:

[5]Philosophers often attribute the result that expectationism implies News is Valuable to Good (1967), but it's really just a reformulation of a result due to David Blackwell (1951), which I think is a better attribution. That said, Das (2023) notes that a similar result is in an old note by C. S. Peirce (1967), first published in an academic setting in 1967, but (according to Wible (1994)), in a US government report in 1879. So maybe the result is very old.

News is Valuable It is never worse to have more information before making a decision, and it is typically better to have more information.

Sure Thing If A is better than B conditional on p being true, and A is better than B conditional on p being false, then A is better than B.

Substitution of Indifferents If Chooser is indifferent between A and B, then Chooser should be indifferent between any two gambles that have the same payouts in all but one possibility, and in that possibility one of them returns A, and the other returns B.

Exit Principle If the choice Chooser makes does not make a difference to the payout Chooser gets if p is false, then it should not matter to Chooser whether they make their choice before or after finding out whether p is true.

These four principles suggest four arguments for expectationism. First, argue that either expectationism is true, or the Quiggin-Buchak theory is true. Second, argue that one of these principles is a plausible second premise. Then conclude, since the principle rules out the Quiggin-Buchak theory, that expectationism must be true.

It's certainly not a requirement of any expectationist that they endorse all four of these arguments. It isn't even a requirement that they endorse any of these four; they could have some other argument. But since expectationism is not the intuitive theory, it is a requirement that they endorse some argument or other, probably something like one of these, to the conclusion that expectationism is true.

This is harder than it looks. EDT, for example, rejects both News is Valuable and Sure Thing in Newcomb's Problem. The news about what Demon has predicted is not, according to EDT, valuable. If Chooser learns what Demon predicted, they will (rationally) choose B, and get, in expectation, a worse return than if they had not learned this and chosen A. And choosing B is preferable conditional on either the Demon predicting, or not predicting, that Chooser will choose A. But A is preferable overall.

That said, it's not like all versions of CDT endorse these principles either. There are a lot of intuitive counterexamples to News is Valuable. It's good to avoid spoilers for movies or football matches. In asymmetric coordination games (such as Table A.9 in Appendix A), it's bad to have it be conventional wisdom that you know what the other person will do. You're sure not to get the best result that way. Das (2023) argues that even given expectationism, the argument for News is Valuable fails on

externalist conceptions of evidence. There is more to say about each of these cases, but the problems for News is Valuable are substantial enough that it doesn't seem like a good premise in an argument against a decision theory.

And, perhaps more surprisingly, there are versions of CDT that reject Sure Thing. Dmitri Gallow (n.d.) argues that any version of CDT which is 'stable' in his sense will reject it. I suspect the argument he gives can generalise to some theories that are not stable in his sense as well. GDT avoids his argument only by the expedient of not offering a preference ordering over alternatives; it just says which choices are choice-worthy, not which choices should and should not be preferred to others.[6] So Sure Thing doesn't look like a safe starting point either, even if something like it might turn out to be true.

——FINISH CHAPTER—-

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.[7] 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.[8]

[6] I'll have much more to say about this in Chapter 9.

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

[8] 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 (2007a). 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.

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. Both demons will predict what Chooser does before anything else happens. 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. And Chooser knows what they are playing, that's part of what it is to play a game, so Demon-1's prediction will be announced, though Demon-2's prediction will be secret. After Chooser makes a decision, 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.

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 object that this is unfair because at the time they make decisions, the EDTer and the CDTer have different evidence. After all, they will know what Demon-1 predicted and it will (almost certainly) be different in each case. Can we get rid of that step? We can, but it's a bit complicated and has some other complications. The example I'll use to illustrate this is a version of a signalling game of the kind introduced by Lewis (1969). And in particular it's a version of the broadly adversarial kinds of signalling games that are central to the plot of Cho and Kreps (1987). It will involve a human Chooser, and a Demon who is excellent at predictions, and the game will have three stages.

At the first stage a fair coin is flipped, and the result shown to Chooser, but not to Demon. At the second stage, Chooser will choose Up or Down, and the choice will be publicly announced. At the third stage, Demon will try to guess what the coin showed. Demon knows the payoff table I'm about to show you, and is arbitrarily good at predicting Chooser's strategy for what to do given how the coin appears. This prediction is causally independent of Chooser's choice, but Demon's guess is not independent; it could be affected by the choice. The payoffs to each player are a function of what happens at each of the three steps, and are given by table Table 4.2. (The payoffs here are all in utils.)

Table 4.2: Payouts for the coins and signals game

Coin	Chooser	Demon	Chooser Payoff	Demon Payoff
H	U	H	40	1
H	U	T	400	0
H	D	H	0	1
H	D	T	0	0
T	U	H	40	0
T	U	T	28	1
T	D	H	0	0

Coin	Chooser	Demon	Chooser Payoff	Demon Payoff
T	D	T	44	1

Figure 4.1 shows the game they are playing in tree form. We start at the middle, then move left or right depending on the coin flip, up or down depending on Chooser's choice, and at one or other angle depending on Demon's choice. Demons's payoffs are just as you'd expect - they get rewarded iff they figure out how the coin landed. Chooser's payoffs are more complicated, but the big things to note are the huge payout if they get to the top-left and Demon does not make a correct prediction, and the generally poor payouts for choosing Down.

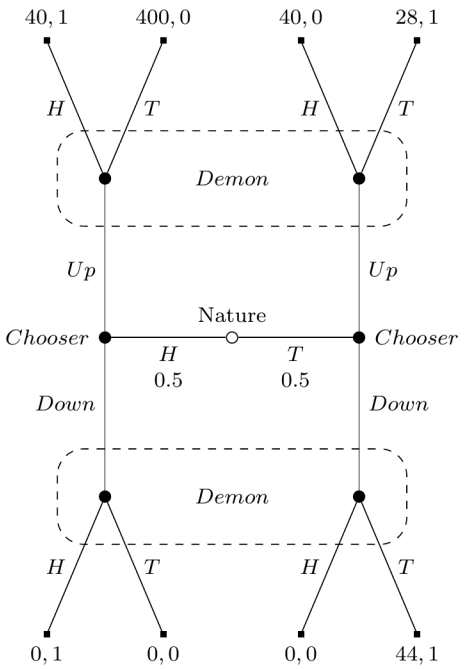

Figure 4.1: Tree Diagram of the Coins and Signals Game

I intend the Demon to be a rational player in a game-theoretic sense. But to translate that into decision-theory terms, it's important to make a few stipulations.[4] Demon predicts Chooser's strategy, that is Chooser's plan about what to do if the coin lands

[4] I'll make the same stipulations in subsequent cases that involve Demon and games that unfold over time.

Heads and what to do if the coin lands Tails, before the game starts. They make their guess about how the coin landed after seeing Chooser's actual choice, and updating their prior beliefs (about both the coin and Chooser) with this information. If they predict that Chooser will do the same thing however the coin lands, they will have no useful information about the coin, so they will flip their own coin to make a guess. In that case it will be 50/50 whether Demon says Heads or Tails. Also, Demon is surprised by what Chooser does, i.e., if they had predicted Chooser would do one thing however the coin lands but Chooser does the other thing, Demon will also flip their own coin to make a guess.(^A key part of the discussion in Cho and Kreps 1987 is that in sone cases we can say substantuve things about what a player will do if they are surprised in this sense. But Figure 4.1 is not one of these cases.) Finally, Demon's predictions are arbitrarily accurate. For simplicity, I'll assume Demon is correct with probability 1, though it doesn't matter if you allow for probability ε that Demon gets it wrong.

Now I want to analyse what Chooser will do if they follow EDT. It should be fairly clear that if the coin lands Heads, Chooser should say Up. The worst possible return from Up is 40, the best possible return from Down is 0. So that's what any theory would recommend, and Chooser will do that whether or not they follow EDT. Indeed, this is so clear that we should assume Demon will predict that Chooser will play Up if the coin lands Heads. So what happens if the coin lands Tails? There are four possibilities here: the two things Chooser might do crossed with the two predictions Demon might make. The expected return to Chooser in these four possibilities is given in Table 4.3.

Table 4.3: The expected payout to Chooser in four cases if the coin lands Tails

	Predict Up	Predict Down
Up	34	40
Down	18	44

The numbers in Table 4.3 aren't entirely obvious; I'll spell out how I got them.

- If Demon predicts Up, Demon will flip a coin. That's because they'll either get no information (if Chooser plays Up), or will be surprised (if Chooser plays Down). So Chooser will get the average of lines 5 and 6 in Table 4.2 if they play Up, and the average of lines 7 and 8 if they play Down.

- If Demon predicts Down, and Chooser plays Up, Demon will think (falsely) that the coin must have landed Heads, since Demon will have predicted that Chooser will only say Up if Heads. So Demon will say Heads. So we'll definitely be at line 5 of Table 4.2, where Chooser gets 40.
- If Demon predicts Down, and Chooser plays Down, Demon will think (correctly) that the coin must have landed Tails. So Demon will say that, and we'll be at line 8 of Table 4.2.

In a decision problem like Table 4.3, EDT says that all that matters is which of the top-left and bottom-right cells is largest. In this case, it's the bottom-right, so EDT says to play Down. That isn't absurd in this case; it gets the best possible payout of 44. So that's our analysis of the game for EDT: Chooser plays Up if Heads, Down if Tails, gets 40 if Heads and 44 if Tails (plus/minus a small amount in expectation if Demon has ε chance of being wrong), and on average gets 42.

GDT does not give any clear verdict about what to do in Table 4.3; it says either Up or Down is permissible. So following GDT doesn't mean you'll do better than EDT in this game; you might do exactly as well as EDT. But all it takes to get a "Why Ain'cha Rich?" argument going is to show that one theory does better than EDT. And the version of CDT that Dmitri Gallow (2020) endorses implies that one should play Up in Table 4.3. So someone following his theory will play Up however the coin lands. So Demon will always flip a coin to decide what to do. So all of the top four outcomes in Figure 4.1 are equally likely, and Chooser will on average get a return of 127. Since $127 > 42$, that means that on average if Chooser follows Gallow's theory, they will on average be much richer than if they follow EDT. So if "Why Ain'Cha Rich?", they show that EDT should be rejected in favor of Gallow's theory.

Ian Wells (2019) has earlier offered an example where EDT predictably does worse than (all versions of) CDT. His case involves a two-step game, where the EDTer will, at step 2, make a decision that everyone, whether they believe in CDT or EDT or any other plausible theory, think is bad from the perspective of the player at step 1. At round 1 the players can pay to tie their hands at round 2, and the EDTer will make this payment. (As would the CDTer who thinks they will become an EDTer before round 2 starts.) Arif Ahmed (2020) responds that this is an unfair criticism. In Wells's cases, he says, the EDT and CDT deciders are not in equivalent situations in round one. The EDTer knows that they will use EDT in later rounds, and the CDTer knows that they will use CDT in later rounds. So they have different evidence about

what will happen at some later time in a way that's relevant to their current decision, so it's not a like-for-like comparison between CDT and EDT at the first stage.

I don't think this is a fair criticism of Wells, or a successful defence of EDT. But even if you think EDT survives Wells's criticisms, that response doesn't work here. Chooser will definitely choose Up if the coin lands Heads, whether they follow EDT, Gallow's theory, or any remotely plausible theory. And this is common knowledge. Demon knows this, and Chooser knows that Demon knows it, and so on. The only difference is that if Chooser follows EDT, they will play Down if Tails. And that's good as far as it goes; they'll probably get the highest possible payoff they can get at that point. More importantly for this debate, they will have the same subjective states if Tails is true whether they follow EDT, Gallow's theory, or anything else. They will believe that they would have played Up if Heads, and that the Demon would have predicted that. So the different choices they make if the coin lands Tails can't be traced back to differences in their subjective states. So the complaint that Ahmed makes about Wells's examples can't be made here (even setting aside the question of whether it is fair complaint). Nonetheless, the EDTer ends up with less money in the long run than the follower of Gallow's theory when playing this game.

All of this is to say that while followers of EDT end up with more money than followers of causal theories *when playing Newcomb's Problem*, this is not because of the distinctive money-making powers of EDT. It's because Newcomb's Problem is designed to leave causally based decision theories badly off. Design a case to leave evidential theories badly off, and they'll be badly off. The "Why Ain'Cha Rich?" consideration tells against everyone, so it overgenerates, so it should be rejected.

Well, not quite everyone. It doesn't tell against some kind of 'resolute' decision theories which recommend one-box in Newcomb's Problem, Up in Table 4.1, and always Up in Figure 4.1. Those theories leave their proponents well off in all three cases. The so-called 'foundational decision theory' that Levinstein and Soares (2020) endorse also endorse the same three 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 2007a), 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

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

rejection 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.[4] 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.[5] 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.[6]

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).

[4] IMPORTANT NOTE TO SELF: This isn't right. In cases where there are multiple equilibria, earlier choices might rule out some later choices. E.g., when there is an exit choice that is guaranteed to be better than some earlier choice. Gotta fix all this.

[5] 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.

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

Table 8.4: Strategic form of Ahmed Insurance.

	$\mathbf{PU_1}$	$\mathbf{PD_1}$
$\mathbf{U_1U_2}$	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[7] 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

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

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 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 Selection

- Solution concepts are not preference ordering
- Talk about Sen for a bit, and how there are principled constraints on selection functions
- One reason why not preference ordering: Indecisiveness. The permissible choices are not things the chooser is indifferent between
- Another, better, reason why not preference ordering: doesn't play an explanatory role.
- Possible objection: It does play an explanatory role, it explains what they would do with fewer choices.
- This is I think the most important point, and it's why I've put off this discussion for so long.
- First reply: Gotta specify whether the loss of option is common knowledge or not, and both answers have flaws
- Second reply: Taking away mixed strategies might take us out of the realm of rational choice (if I'm right in mixed strategies chapter)
- Third reply: Taking away some strategies in dynamic cases might be really weird.
- Example: Two rounds. At each round a $1 bill is on table, and Chooser takes it or leaves it.
- What does it mean to remove the strategy TTT, i.e., take at R1, and then take at R2 whether you take or leave at R1.
- Counter: This is an artifact of the definition of strategies
- My reply: Need this definition of strategies to explain simple (3 step) centipede problems
- Counter: Stalnaker says that this is wrong
- My reply: Eh, that's a point. I disagree here, but whatever
- This is why I reject strategic form normal form

10 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 2020b, 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 10.1: 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.

11 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 11.1; what would the ideal chooser do?

Table 11.1: 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 11.2.

Table 11.2: An example of iterated weak dominance.

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

In Table 11.2, 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 11.1, in which we said Chooser should play U.

So does it make sense to say that U is the only play in Table 11.2? 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 11.1 simply don't carry over to Table 11.2. 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 11.1, but not in Table 11.2.

12 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.

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.

Blackwell, David. 1951. "Comparison of Experiments." *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability* 2 (1): 93–102.

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.

———. 2014. "Belief, Credence and Norms." *Philosophical Studies* 169 (2): 285–311.

https://doi.org/10.1007/s11098-013-0182-y.

Chandler, Jake. 2017. "Descriptive Decision Theory." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Winter 2017. https://plato.stanford.edu/archives/win2017/entries/decision-theory-descriptive/; Metaphysics Research Lab, Stanford University.

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.

Cohen, Shani, and Shengwu Li. 2023. "Sequential Cursed Equilibrium." https://arxiv.org/abs/2212.06025.

Conlisk, John. 1996. "Why Bounded Rationality?" *Journal of Economic Literature* 34 (2): 669–700.

Das, Nilanjan. 2023. "The Value of Biased Information." *British Journal for the Philosophy of Science* 74 (1): 25–55. https://doi.org/10.1093/bjps/axaa003.

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

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

———. 2007a. "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.

Fong, Meng-Jhang, Po-Hsuan Lin, and Thomas R. Palfrey. 2023. "Cursed Sequential Equilibrium." https://arxiv.org/abs/2301.11971.

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/jphil2020011739.

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

Gibbard, Allan, and William Harper. 1978. "Counterfactuals and Two Kinds of Expected Utility." In *Foundations and Applications of Decision Theory*, edited by C.

A. Hooker, J. J. Leach, and E. F. McClennen, 125–62. Dordrecht: Reidel.
Good, I. J. 1967. "On the Principle of Total Evidence." *British Journal for the Philosophy of Science* 17 (4): 319–21. https://doi.org/10.1093/bjps/17.4.319.
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. 1998. *From Metaphysics to Ethics: A Defence of Conceptual Analysis*. Clarendon Press: Oxford.
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.
Joyce, James M. 2012. "Regret and Instability in Causal Decision Theory." *Synthese* 187 (1): 123–45. https://doi.org/10.1007/s11229-011-0022-6.
Keynes, John Maynard. 1923. *A Tract on Monetary Reform*. London: Macmillan.
Knight, Frank. 1921. *Risk, Uncertainty and Profit*. Chicago: University of Chicago 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/jphil20201

17516.
Lewis, David. 1969. *Convention: A Philosophical Study*. Cambridge: Harvard University Press.
———. 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.
———. 2020a. "Letter to Hugh Mellor Dated 14 October 1981." In *Philosophical Letters of David k. Lewis*, edited by Helen Beebee and A. R. J. Fisher, 2: Mind, Language, Epistemology:432–34. Oxford: Oxford University Press.
———. 2020b. "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.
Mills, Charles W. 2005. ""Ideal Theory" as Ideology." *Hypatia* 20 (3): 165–84. https://doi.org/10.1111/j.1527-2001.2005.tb00493.x.
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.
O'Connor, Cailin. 2019. *The Origins of Unfairness: Social Categories and Cultural Evolution*. Oxford: Oxford University Press.
Peirce, C. S. 1967. "Note on the Theory of the Economy of Research." *Operations Research* 15 (4): 643–48.
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 Problem)." Santa Monica, CA: The RAND Corporation.
Schrijver, Alexander. 2005. "On the History of Combinatorial Optimization (till 1960)." *Handbooks in Operations Research and Management Science* 12: 1–68. https://doi.org/10.1016/S0927-0507(05)12001-5.
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.
Sen, Amartya. 2006. "What Do We Want from a Theory of Justice?" *Journal of Philosophy* 103 (5): 215–38. https://doi.org/10.5840/jphil2006103517.
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.
Smullyan, Raymond. 1978. *What Is the Name of This Book? The Riddle of Dracula and Other Logical Puzzles*. Englewood Cliffs, NJ: Prentice-Hall.
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.
Valentini, Laura. 2012. "Ideal Vs. Non-Ideal Theory: A Conceptual Map." *Philosophy Compass* 7 (9): 654–64. https://doi.org/10.1111/phco.2012.7.issue-9.
Wedgwood, Ralph. 2012. "Justified Inference." *Synthese* 189 (2): 273–95. https://doi.org/10.1007/s11229-011-0012-8.
———. 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.
Wible, James R. 1994. "Charles Sanders Peirce's Economy of Research." *Journal of Economic Methodology* 1 (1): 135–60. https://doi.org/10.1080/13501789400000009.
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.

A Games as Decisions

Much of what happens in this book comes from seeing demonic decision problems as games and, conversely, seeing games as potential demonic decision problems. So I want to spend a little time setting out how the translation between the two works. This is intended largely for people who want to use the existing resources in game theory, which are voluminous, as a source for decision theoretic ideas.

Say that a demonic decision problem is a problem where the states are sensitive to the predictions of some kind of demon. So Newcomb's Problem is the classic demonic decision problem, but it's hardly the only one. Indeed, almost every decision problem in this book is demonic. Transforming a demonic decision problem into a game is easy. As I noted, you just replace the states generated by Demon's choices with moves for Demon, and give them payout 1 if they predict correctly, and 0 otherwise.

You might worry that this only gives you cases where Demon is approximately perfect, but we also want cases where the demon is, say, 80% accurate. But that's easy to do as well. In fact there are two ways to do it.

The first is what I'll call the Selten strategy, because it gives the demon a 'trembling hand' in the sense of R. Selten (1975). Instead of letting Demon choose a state in the original problem, let Demon choose one of n buttons, where n is the number of choices the (human) chooser has. Each button is connected to a probabilistic device that generates one of the original states. If you want Demon to be 80% accurate when option o_i is chosen, say the button b_i associated with option o_i outputs state s_i with probability 0.8, and each of the other states with probability $\frac{0.2}{n-1}$. And still say that Demon gets payout 1 for any i if the chooser selects o_i and the button generates state s_i, and 0 otherwise.

The second is what I'll call the Smullyan strategy, because it involves a Knights and Knaves puzzle of the kind that play a role in several of Smullyan's books, especially his (1978). Here the randomisation takes place before Demon's choice. Demon is

assigned a type Knight or Knave. Demon is told of the assignment, but Chooser is not. If Demon is assigned type Knight, the payouts stay the same as in the game where Demon is arbitrarily accurate. If Demon is assigned type Knave, the payouts are reversed, and Demon gets payout 1 for an incorrect prediction.

There are benefits to each approach, and there are slightly different edge cases that are handled better by one or other version. I find the Selten strategy a little easier to use, especially if Demon's expected accuracy is different with different choices by Chooser. But in general either will work for turning a demonic decision problem into a game.

Turning games into demonic decision problems is a bit more interesting. Start with a completely generic two-player, two-option, simultaneous move, symmetric game, as shown in table Table A.1. We won't only look at symmetric games, but it's a nice way to start.

Table A.1: A generic symmetric game.

	A	B
A	x, x	y, z
B	z, y	w, w

In words, what this says is that each player chooses either A or B. If they both choose A, they both get x. If they both choose B, they both get w. And if one chooses A and the other chooses B, the one who chooses A gets y and the one who chooses B gets z. (Note that the payouts list row's payment first, if you're struggling to translate between the table and the text.) A lot of famous games can be defined in terms of restrictions on the four payout values. For example, a game like this is a Prisoners' Dilemma if the following constraints are met.

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

Some books will also add 2_x_ > $y + z$ as a further constraint, but I'll stick with these three.

Now to turn a game into a demonic decision problem, first replace column's payouts with 1s and 0s, with 1s along the main diagonal, and 0s everywhere else. Table Table A.2 shows what a generic symmetric game looks like after this transformation.

Table A.2: The demonic version of a generic symmetric game.

	A	B
A	x, 1	y, 0
B	z, 0	w, 1

The next step is to replace Demon's moves with states that are generated by Demon's predictions. As before, I'll put 'P' in front of a choice name to indicate the state of that choice being predicted. The result is table Table A.3, which we already saw back in the introduction.

Table A.3: The demonic decision problem generated by a generic symmetric game.

	PA	PB
A	x	y
B	z	w

If we add the constraints $x > z$, $y > w$, $w > x$, this is a Newcomb Problem. I'm a long way from the first to point out the connections between Prisoners' Dilemma and Newcomb's Problem; it's literally in the title of a David Lewis paper (Lewis 1979). But what I want to stress here is the recipe for turning a familiar game into a demonic problem.

We can do the same thing with Chicken. The toy story behind Chicken is that two cars are facing off at the end of a road. They will drive straight at each other, and at the last second, each driver will choose to swerve off the road, which we'll call option A, or stay on the road, which we'll call option B. If one swerves and the other stays, the one who stays is the winner. If they both swerve they both lose and it's boring, and if they both stay it's a fiery crash. So in terms of the payouts in the general symmetric game, the constraints are:

- $x < z$

- $y \gg w$
- $x \gg w$

Just what it means for one value to be much more than another, which is what I mean by '$\gg$', is obviously vague. Table A.4 gives an example with some numbers that should satisfy it.

Table A.4: A version of Chicken.

	A	B
A	0, 0	0, 1
B	1, 0	-100,-100

Replace the other driver, the one who plays column in this version, with a Demon, who only wants to predict row's move. The result is Table A.5.

Table A.5: A demonic version of Chicken.

	A	B
A	0, 1	0, 0
B	1, 0	-100,1

All I've done to generate table Table A.5 is replace column's payouts with 1s on the main diagonal, and 0s elsewhere. The next step is to replace the demonic player with states generated by Demon's choices. The result is table Table A.6.

Table A.6: A demonic decision problem based on Chicken.

	PA	PB
A	0	0
B	1	-100

And Table A.6 is just the Psychopath Button example that Andy Egan (2007b) raises as a problem for Causal Decision Theory.

Another familiar game from introductory game theory textbooks is matching pennies. This is a somewhat simplified version of rock-paper-scissors. Each player has a penny, and they reveal their penny simultaneously. They can either show it with the heads side up (option A), or the tails side up (option B). We specify in advance who wins if they show the same way, and who wins if they show opposite ways. So let's say column will win if both coins are heads or both are tails, and row will win if they are different. The payouts are shown in Table A.7.

Table A.7: The game matching pennies.

	A	B
A	0, 1	1, 0
B	1, 0	0, 1

This isn't a symmetric game, but it is already demonic. Column's payouts are 1 in the main diagonal and 0 elsewhere. So we can convert it to a demonic decision problem fairly easily, as in Table A.8.

Table A.8: Matching Pennies as a decision problem.

	PA	PB
A	0	1
B	1	0

And Table A.8 is the familiar problem Death in Damascus from Gibbard and Harper (1978).

Let's do one last one, starting with the familiar game Battle of the Sexes.[1] Row and Column each have to choose whether to do R or C. They both prefer doing the same thing to doing different things. But Row would prefer they both do R, and Column would prefer they both do C. The original name comes from a version of the story where Row and Column are a heterosexual married couple, and Row wants to do some stereotypically male thing, while Column wants to do some stereotypically

[1] O'Connor (2019) calls this the Bach-Stravinsky game, which is a better name.

female thing. That framing is tiresome at best, but the category of asymmetric coordination games is not, hence my more abstract presentation. So Table A.9 is one way we might think of the payouts.

Table A.9: A version of Battle of the Sexes.

	R	C
R	4, 1	0, 0
C	0, 0	1, 4

As it stands, that's not a symmetric game. But we can make it a symmetric game by relabeling the choices. Let option A for each player be doing their favored choice, and option B be doing their less favored choice. That turns Table A.9 into Table A.10.

Table A.10: Battle of the Sexes, relabeled.

	A	B
A	0, 0	4, 1
B	1, 4	0, 0

After making that change, change column's payouts so that it is a demonic game. The result is Table A.11.

Table A.11: A demonic version of battle of the sexes.

	A	B
A	0, 1	4, 0
B	1, 0	0, 1

Finally, replace Demon's choices with states generated by (probably accurate) predictions, to get the decision problem in Table A.12.

Table A.12: A demonic decision problem based on Battle of the Sexes.

	PA	PB
A	0	4
B	1	0

That decision problem is the asymmetric version of Death in Damascus from Richter (1984).

The point of this section has not just been to show that we can turn games into decision problems by treating one of the players as a predictor. That's true, but not in itself that interesting. Instead I want to make two further points.

One is that most of the problems that have been the focus of attention in the decision theory literature in the past couple of generations can be generated from very familiar games, the kinds of games you find in the first one or two chapters of a game theory textbook. And the generation method is fairly similar in each respect.

The second point is that most of the simple games you find in those introductory chapters turn out to result, once you transform them this way, in demonic decision problems that have been widely discussed. But there is just one exception here. There hasn't been a huge amount of discussion of the demonic decision problem you get when you start with the game known as Stag Hunt. One of the aims of this book has to been to remedy that. Particularly in Chapter 7 I've relied on demonic decision problems that you get by starting with Stag Hunt and applying the transformations discussed in this appendix.

But there are so many more interesting examples from game theory that could be used to generate interesting decision problems. The beer-quiche game that Cho and Kreps (1987) developed is not nearly as widely discussed as Chicken or Stag Hunt, but as we saw in Chapter 4, it has interesting consequences for decision theory. I'm sure that there are more interesting results that can be generated by transforming other games into decision problems.

B Non-Ideal Decision Theory

C 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 $x + y + z = 1$.

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

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.

D 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 D.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 utility. 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 D.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 D.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 D.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 D.3, where y is arbitrary.

Table D.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 D.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 D.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 D.4, which is just the same as Table D.3, except the exit payout is now 2.

In Table D.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.

E Against Uniqueness