

The End of Decision Theory

Brian Weatherson

What is Decision Theory a Theory Of?

If you're reading a journal like this, you're probably familiar with seeing papers defending this or that decision theory. Familiar decision theories include:

- Causal Decision Theory [GibbardHarper1979; Lewis1980; Skyrms1990; Joyce199x];
- Evidential Decision Theory [Ahmed2013];
- Benchmark theory [Wedgwood2013];
- Risk-Weighted theory [BuchakRiskRationality];
- Tournament Decision Theory [Podgorski2020]; and
- Functional Decision Theory [LevinsteinSoares2020]

Other theories haven't had snappy 'isms' applied to them, such as the non-standard version of Causal Decision Theory that Dmitri Gallow2020 defends, or the pluralist decision theory that Jack Spencer2021 defends, or the broadly ratificationist theory that Melissa Fuscon2021 defends

This paper isn't going to take sides between these nine or more theories.¹ Rather it is going to ask a prior pair of questions.

1. If these are the possible answers, what is the question? That is, what is the question to which decision theories are possible answers?
2. Why is that an interesting question? What do we gain by answering it?

On 1, I will argue that decision theories are answers to a question about what an ideal decider would do. The 'ideal' here is like the 'ideal' in a scientific idealisation, not the ideal in something like an ideal advisor moral theory. That is, the ideal decider is an idealisation in the sense of being simple, not in the sense of being perfect. The ideal decision maker is ideal in the same way that the point-masses in the ideal gas model are ideal; they are (relatively) simple to work with. The main opponent I have in mind is someone who says that in some sense decision theory tells us what decisions we should make.

¹The arguments here are intended to support a theory like Fusco's, but in a fairly roundabout way, but the connection between what I say here and Fusco's theory would take a paper as long as this one to set out.

On 2, I will argue that the point of asking this question is that these idealisations play important roles in explanatorily useful models of social interactions, such as the model of the used car market that Georg @Akerlof1970 described. Here, the main opponent I have in mind is someone who says that decision theory is useful because it helps us make better decisions.

There is another pair of answers to this question which is interesting, but which I won't have a lot to say about here. David Lewis said that decision theory was in the business of setting out a key part of the notion of rationality that is constitutive of having mental states. It is important on this view that we get decision theory right because the details might matter to the right theory of mental content. I think the arguments I give for my preferred view provide some reasons to be sceptical of this view, but they are far from conclusive reasons against it. My main purpose is to argue against the claim that decision theory says anything about what we should do, or that it provides us with advice.²

The nine theories I mentioned above disagree about a lot of things. In philosophy we typically spend our time looking at cases where theories agree. Not here! I will focus almost exclusively on two cases where those nine theories all say the same thing. I'll assume that whatever question they are asking, the correct answer to it in those two cases must agree with all nine theories. That will be enough to defend the view I want to defend, which is that a decision theory is correct iff it is true in the right kind of idealisation.

Three Cases

Betting

Chooser has \$110, and is in a sports betting shop. There is a basketball game about to start, between two teams they know to be equally matched. Chooser has three options: bet the \$110 on Home, bet it on Away, keep money. If they bet and are right, they win \$100 (plus get the money back they bet), if they are wrong, they lose the money. Given standard assumptions about how much Chooser likes money, all the decision theories I'm discussing say Chooser should not bet.

From this it follows that decision theory is not in the business of answering this question: *What action will produce the best outcome?* We know, and so does Chooser, that the action that produces the best outcome is to bet on the winning team. Keeping their money in their pocket is the only action they know will be sub-optimal. And it's what decision theory says to do.

²Lewis's views about the role of decision theory are set out in In a letter to ... he says that the advisory role of decision theory is one that he would "downplay", but he still seems to make it more central than I would do.

This is to say, decision theory is not axiology. It's not a theory of evaluating outcomes, and saying which is best. Axiology is a very important part of philosophy, but it's not what decision theorists are up to.

So far this will probably strike you, dear reader, as obvious. But there's another step, that I think will strike some people as nearly as obvious, that I'm at pains to resist. Some might say that decision theorists don't tell Chooser to bet on the winner because this is lousy advice. Chooser can't bet on the winner, at least not as such. That, I'll argue, would be a misstep. Decision theorists do not restrict themselves to answers that can be practically carried out.

Salesman

We'll focus on a version of what Julia @Robinson1949 called the traveling salesman problem. Given some points on a map, find the shortest path through them. We'll focus on the 257 cities shown on the map in Figure ??.

Loading required package: tidyverse

```
-- Attaching core tidyverse packages ----- tidyverse 2.0.0 --
v dplyr      1.1.4      v readr      2.1.5
v forcats    1.0.0      v stringr    1.5.1
v ggplot2    3.5.1      v tibble     3.2.1
v lubridate  1.9.3      v tidyr      1.3.1
v purrr      1.0.2
```

```
-- Conflicts ----- tidyverse_conflicts() --
```

```
x dplyr::filter() masks stats::filter()
```

```
x dplyr::lag()     masks stats::lag()
```

```
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become
```

Loading required package: TSP

Loading required package: maps

Attaching package: 'maps'

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

map

Loading required package: grid

Warning: A numeric `legend.position` argument in `theme()` was deprecated in ggplot2 3.5.0.

i Please use the `legend.position.inside` argument of `theme()` instead.

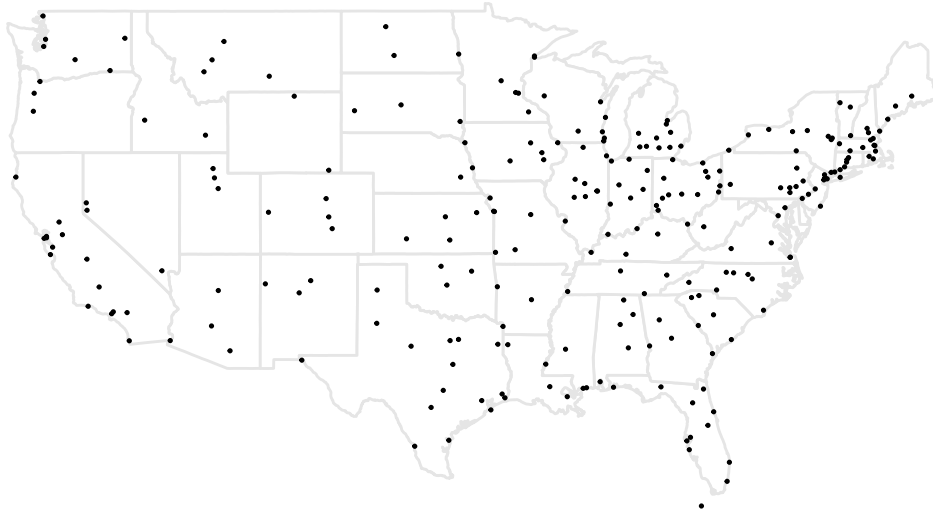


Figure 1: 257 American cities; we'll try to find the shortest path through them.

The task is to find the shortest path through those 257 cities.³

All nine of the decision theories I mentioned, and as far as I know every competitor to them in the philosophical literature, say the thing to do here is to draw whichever of the $256!$ possible paths is shortest. That is not particularly helpful advice. Unless you know a lot about problems like this, you can't draw the shortest path through the map. And least, you can't draw it as such. You can't draw it in the way that you can't enter the correct code on a locked phone [MandlekernEtAl2017].

One of the striking things about this puzzle is that it turns out there are some helpful things that can be said. One helpful bit of advice to someone trying to solve a problem like this is to use a Farthest Insertion Algorithm. Insertion algorithms say to start with a random city, then add cities to the path one at a time, at each time finding the point to insert the city into the existing path that adds the least distance. The Farthest Insertion Algorithm says that the city added at each stage is the one farthest from the existing path. Insertion algorithms in general

³Include citation as to where I got the map, and the packages used in this, and to where I learned some of this stuff about TS problems.

produce pretty good paths in a very short amount of time - at least on normal computers. And the Farthest Insertion Algorithm is, most of the time, the best Insertion Algorithm to use. Figure ?? shows the result of one output of this algorithm.⁴



Figure 2: An output of the Farthest Insertion Algorithm, with a length of 21197 miles.

The path in Figure ?? is not bad, but with only a bit of extra computational work, one can do better. A fairly simple optimisation algorithm takes a map as input, and then deletes pairs of edges at a time, and finds the shortest path of all possible paths with all but those two edges. The process continues until no improvements can be made by deleting two edges at a time, at which point you've found a somewhat resilient local minimum. Figure ?? is the output from applying this strategy to the path in Figure ??.

[1] FALSE

This optimisation tends to produce paths that look a lot like the original, but are somewhat shorter. For most practical purposes, the best advice you could give someone faced with a problem like this is to use a Farthest Insertion Algorithm, then optimise it in this way. Or, if they have a bit more time, they could do this a dozen or so times, and see if different starting cities led to slightly shorter paths.

While this is good advice, and indeed it's what most people should do, it's not typically what is optimal to do. For that reason, it's not what our nine decision theories would say to do.

⁴The algorithm is silent on which city you start with, and usually chooses this randomly.



long

Figure 3: An output of the Farthest Insertion Algorithm, with a length of 21197 miles.

If one had unlimited and free computing power available, hacks like these would be pointless. One would simply look at all the possible paths, and see which was shortest. I do not have free, unlimited computing power, so I didn't do this. Using some black box algorithms I did not particularly understand, I was able to find a shorter path, however. It took some time, both of mine and my computer's, and for most purposes it would not have been worth the hassle of finding it. Still, just to show it exists, I've plotted it as Figure ??.

[1] TRUE

I'm not sure if Figure ?? is as short as possible, but I couldn't find a shorter one. Still, unless those 200 miles really make a difference, it wouldn't have been worth the trouble it took to find this map.

The Two Cases

Let's summarise these two cases in a table.

	<div> <div>Betting</div> <div>Salesman</div> </div>	
Best outcome	Bet on winner	Shortest path

	Betting	Salesman
Decision theory	Pass	Shortest path
Best advice	Pass	Learn algorithms

The first row says which action would produce the best outcome in the two cases. The third row says what advice one ought give someone who had to choose in the two cases. And the middle row says what all the decision theories say about the two cases. Notably, it agrees with neither the first nor third row. Decision theory is neither in the business of saying what will produce the best result, nor with giving the most useful advice. So what is it doing?

Decision Theory as Idealisation

Imagine a version of Chooser with, as Rousseau might have put it, their knowledge as it is, and their computational powers as they might be. That is, a version of Chooser who has unlimited, and free, computational powers, but no more knowledge of the world than the actually have - save what they learn by performing deductions from their existing knowledge.

Decision theories describe what that version of Chooser would do in the problem that Chooser is facing. In the betting case, adding unlimited computing power doesn't tell you who is going to win the game. So that version of Chooser will still avoid betting. But in the Salesman case, adding unlimited computing power is enough to solve the problem. They don't even have to use any fancy techniques. To find the shortest path, all it takes is finding the length of each path, and sorting the results. The first requires nothing more than addition; at least if, as was the case here, we provided the computer with the distances between any pairs of cities as input. The second just requires being able to do a bubble sort, which is technically extremely simple. To be sure, doing all these additions, then doing a bubble sort on the results, will take longer than most human lives on the kinds of computers most people have available to them. But a version of Chooser with unlimited, free, computational power will do these computations no problem at all.

If we say that

Technical Detour

Most philosophical decision theory concerns decisions under uncertainty, not decisions like Salesman that are made under certainty.

- But the structure is still the same.



Figure 4: An output of the Farthest Insertion Algorithm, with a length of 21197 miles.

Technical Detour

They say that for each option, you should loop through the possible states of the world, in each case multiplying something (usually a probability) by something else (usually a utility), and then summing the results. Then you choose the maximum.

- That's exactly the same technical task as solving Salesman by brute force.⁵

Summary

Decision theory describes what a particular kind of idealised agent **will** do.

- I've bolded **will** because it's going to turn out that's the important modal to use here; as opposed to *should*.
- If there is any normativity here, it's in the **idealised** part of that sentence, not the modal.

⁵Actually one step harder because of the multiplication, but otherwise the same.

Idealisations as Life Goals

A Modest Proposal

Decision theory is relevant to how we should act because:

1. It tells us that idealised people do use decision theory, and
 2. We should try to be like idealised people.
- C. We should try to use decision theory.
-



Figure 5: I think this stands for What Would Jeffrey Do?

First Objection - Knowing the Inputs

To use decision theory as a guide to action, I need to know the utility of the possible states.

- Knowing ordering isn't enough, need cardinality of each utility.

- I can only ever tell that the utility of A is half way between that of B and C by thinking about whether A is better or worse to take than a 50/50 bet on B or C.
- I need to make decisions to get the inputs to decision theory.
- And I think this is the usual case.

Second Objection - The General Theory of the Second Best

In general, it's not true that one should try to approximate what the ideal is like.

The General Theory of Second Best¹

There is an important basic similarity underlying a number of recent works in apparently widely separated fields of economic theory. Upon examination, it would appear that the authors have been rediscovering, in some of the many guises given it by various specific problems, a single general theorem. This theorem forms the core of what may be called *The General Theory of Second Best*. Although the main principles of the theory of second best have undoubtedly gained wide acceptance, no general statement of them seems to exist. Furthermore, the principles often seem to be forgotten in the context of specific problems and, when they are rediscovered and stated in the form pertinent to some problem, this seems to evoke expressions of surprise and doubt rather than of immediate agreement and satisfaction at the discovery of yet another application of the already accepted generalizations.

In this paper, an attempt is made to develop a *general* theory of second best. In Section I there is given, by way of introduction, a verbal statement of the theory's main general theorem, together with two important negative corollaries. Section II outlines the scope of the general theory of second best. Next, a brief survey is given of some of the recent literature on the subject. This survey brings together a number of cases in which the general theory has been applied to various problems in theoretical economics. The implications of the general theory of second best for piecemeal policy recommendations, especially in welfare economics, are considered in Section IV. This general discussion is followed by two sections giving examples of the application of the theory in specific models. These examples lead up to the general statement and rigorous proof of the central theorem given in Section VII. A brief consideration of the existence of second best solutions is followed by a classificatory discussion of the nature of these solutions. This taxonomy serves to illustrate some of the important negative corollaries of the theorem. The paper is concluded with a brief discussion of the difficult problem of multiple-layer second best optima.

I A GENERAL THEOREM IN THE THEORY OF SECOND BEST²

It is well known that the attainment of a Paretian optimum requires the simultaneous fulfillment of all the optimum conditions. The general theorem for the second best optimum states that if there is introduced into a general equilibrium system a constraint which prevents the attainment of one of the Paretian conditions, the other Paretian conditions, although still attainable, are, in general, no longer desirable. In other words, given that one of the Paretian optimum conditions cannot be fulfilled, then an optimum situation can be achieved only by departing from all the other Paretian conditions. The optimum situation finally attained may be termed a second best optimum because it is achieved subject to a constraint which, by definition, prevents the attainment of a Paretian optimum.

From this theorem there follows the important negative corollary that there is no *a priori* way to judge as between various situations in which some of the Paretian optimum

¹ The authors are indebted to Professor Harry G. Johnson for a number of helpful suggestions relating to this paper. The appellation, "Theory of Second Best," is derived from the writings of Professor Meade; See Meade, J. E., *Trade and Welfare*, London, Oxford University Press, 1955. Meade has given, in *Trade and Welfare*, what seems to be the only attempt to date to deal systematically with a number of problems in the theory of second best. His treatment, however, is concerned with the detailed case study of several problems, rather than with the development of a general theory of second best.

² See section VII for formal proofs of the statements made in this section.

This is one of the most philosophically important economics papers ever published.

Second Best

Often times, the right thing to do is something whose value consists in mitigating the costs of our other flaws.

- We should, especially in high stakes settings, stop and have a little think before acting.
- The “ideal agent” of decision theory never stops to have a think.
- Stopping is costly, and **they** don’t gain anything from it.

Second Best

- The ideal agent does lots of things we don’t do.
- They always take reasonable hedges against costly possibilities, and they never stop to have a think.
- Knowing that the ideal agent is F doesn’t tell us whether we should try to be F unless we also know that F is more like the first of these than the second.
- And decision theory, in **anything like its current form**, is not particularly helpful on this score.

Third Objection - The Yoda Objection

Decision theory doesn’t say what one should try or not try, it says what one should do.

- So it’s weird to infer something about trying from a theory about doing.

Yoda

I think there’s something importantly right about this - decision theory gives criteria of correctness not methods of deliberation - but that in turn shows us why it might be useful.

Idealisations as Models

Two Notions of Idealisation

In philosophy we use the word ‘idealisation’ for two rather different kinds of thing.

1. Perfect

2. Simple

The point particles in ideal gas theory are not perfect - having volume is not an imperfection.

Nor are they things to aim for - high school chemistry does not imply a rule: **Smaller the better.**

But they are simple.

Idealisations in Decision Theory

Decision theory provides idealisations in the second sense - they are **simplifications**.

- Just like the point masses we use in the ideal gas law, they say not what should happen, but what would happen in the absence of certain complications.

Idealisations in Decision Theory

Why do I say this idealisation is a simplification not a perfection?

1. Allocating zero seconds to hard math problems is not a perfection.
2. The idealised self isn't absolutely perfect - they have very restricted information.

Idealisations in Decision Theory

The idealised self that gets used is god-like in one respect - computational ability - but human-like in another - informational awareness.

- That's a common feature of idealised models.
- You abstract away from one feature, but not others.

Why Care?

That's what we do, but why do we do it?

- Because sometimes these models are enlightening.
- Sometimes, the fact that we have computational limitations is not relevant to predicting/explaining/understanding what we will do.

Really, Why Care?

It's tempting to identify these with high stakes situations, since those are ones where we'll throw enough computational resources at the problem that we have god-like powers.

- But that isn't quite right.
- In some high stakes cases, we also throw enough investigative resources at the problem that holding actual knowledge fixed is a bad modeling assumption.

Informational Limitations

What we need are cases where there are principled limitations to our informational capacities, such as,

1. Cases where the information concerns the future; or
2. Cases where someone has (or may have) just as strong an incentive to hide information from us.

I'll end with a discussion of an important instance of the second.⁶

⁶Photo of George Akerlof on next slide by Yan Chi Vinci Chow.

Akerlof on Lemons





Akerlof on Lemons





Lemon.

A 20th Century Puzzle

Used cars sold at a huge discount to new cars, even when the cars were just a few months old with almost no usage.

- There was no agreed upon explanation for this, with the most common theory being that it reflected a preference/prejudice on the part of buyers.

Akerlof's Theory

Make the following assumptions.

1. Cars vary a lot in quality, even coming from the same production line.
2. Sellers of used cars know how good this token car is.
3. Buyers of used cars do not; they only know how good the model is in general.
4. People rarely sell cars they just bought.
5. Everyone involved is an expected utility maximiser.

Akerlof's Theory

Akerlof built a formal model with the following properties.

- The most common reason to sell a car one just bought is the discovery that it was a bad instance of that kind of car.
- Knowing this, buyers of used cars demanded a big discount in exchange for the possibility they were buying a dud.
- Everyone is acting rationally within the model, given their asymmetric information.

Akerlof's Theory

If he was right (and I basically think he was) you'd expect the used car discount to fall if either of the following things happened.

1. Production lines got more reliable, and cars off the same line were more similar to one another; or
2. Buyers had access to better tools to judge the quality of used cars.

By 2020 both of those things had happened, and the used car discount was almost zero. (Then in 2021 it went negative for weird reasons.)

Back To Philosophy

You can't build models like this without a theory of rational action under uncertainty.

- And that's the payoff of philosophical decision theory.
- It's an essential input to useful models, like this one.

Consequences for Decision Theory

The thing about explanatory models is that they can have very limited scope.

- There are lots of properties of gases that you cannot explain with the ideal gas model.

Consequences for Decision Theory

This matters because a lot of people in decision theory assume that a good decision theory will have something to say about every possible choice situation.

- But if you're in the business of explanation, it's fine to say that the theory only applies in some cases, and it only provides explanations in those cases.

Consequences for Decision Theory

There are (at least) two interesting notions around here:

1. What the ideal decider would do in a particular situation.
 2. What it would be advisable for a real life human to do in this situation.
- These come apart in Traveling Salesman cases, and we should keep open the possibility that they also come apart in cases that philosophers talk about.

For Another Day

At this point you might expect that I'd have a theory that does go silent on a bunch of hard cases, and which explains away a bunch of intuitions about cases as intuitions about advisability, not about what an ideal decider does, and you'd be right on both counts.

- And I'm happy to talk about that theory (and those cases) at literally any length people want.
- But it's mostly for another talk.

For Yet Another Day

You might also wonder at this point whether there are other idealisations we could make, which are useful in different circumstances to the standard computationally perfect, informationally limited model.

- I used to think the answer was no on broadly a priori grounds.
- But this was wrong, and the work on cursed equilibrium shows it was wrong.
- There are some potentially really interesting questions here for philosophy of economics that engages seriously with 21st century economics.

Conclusions

- Decision theory provides idealisations.
- These are not things we should aim for, but simplifications that play a role in explanations.