

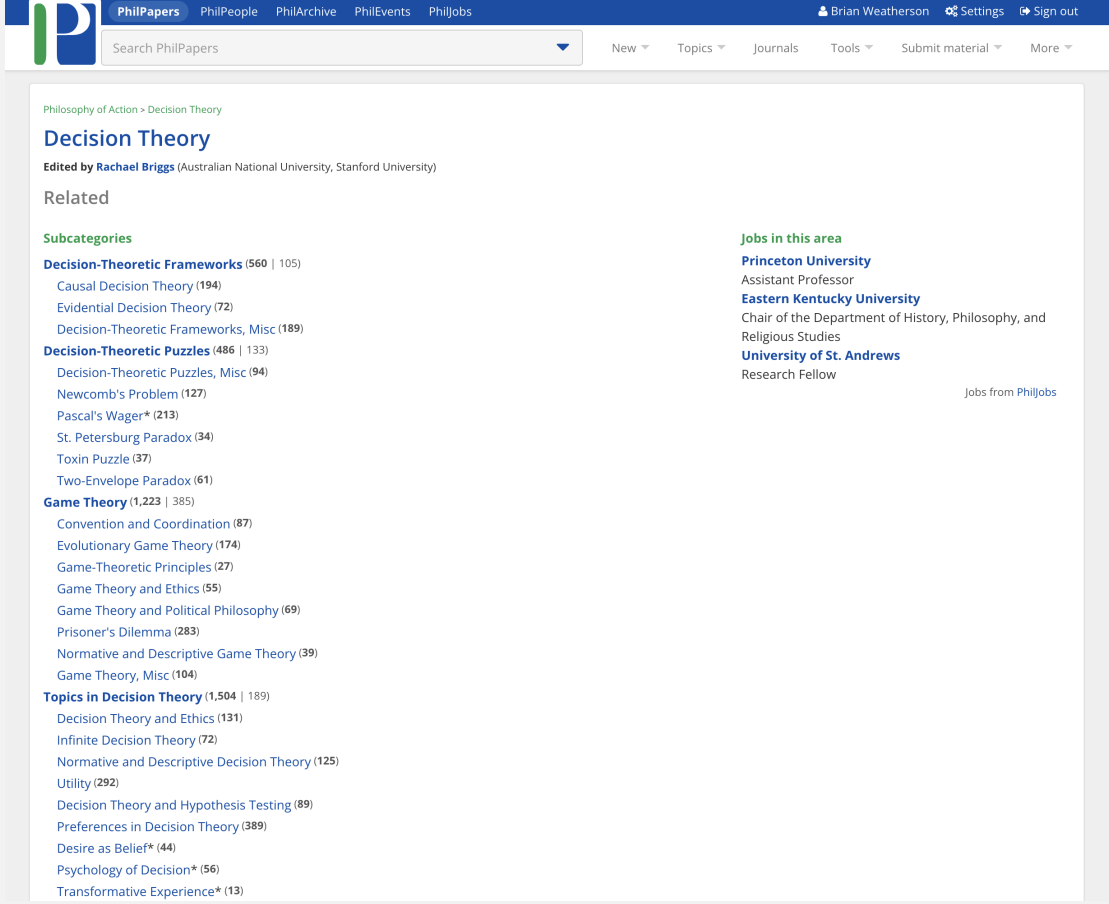
The End of Decision Theory

Brian Weatherson

Overview

Decision Theory

This talk is about decision theory, i.e., this field:



The screenshot shows the PhilPapers website interface. At the top is a navigation bar with links for PhilPapers, PhilPeople, PhilArchive, PhilEvents, and PhilJobs. On the right of the navigation bar are links for Brian Weatherston, Settings, and Sign out. Below the navigation bar is a search bar labeled "Search PhilPapers" and a dropdown menu. The main content area is titled "Philosophy of Action > Decision Theory" and "Decision Theory". It is edited by Rachael Briggs (Australian National University, Stanford University). The page is divided into two main sections: "Related" and "Jobs in this area". The "Related" section lists subcategories and topics with their respective counts. The "Jobs in this area" section lists job openings at Princeton University, Eastern Kentucky University, and the University of St. Andrews.

PhilPapers PhilPeople PhilArchive PhilEvents PhilJobs Brian Weatherston Settings Sign out

Search PhilPapers

Philosophy of Action > Decision Theory

Decision Theory

Edited by Rachael Briggs (Australian National University, Stanford University)

Related

Subcategories

- Decision-Theoretic Frameworks** (560 | 105)
 - Causal Decision Theory (194)
 - Evidential Decision Theory (72)
 - Decision-Theoretic Frameworks, Misc (189)
- Decision-Theoretic Puzzles** (486 | 133)
 - Decision-Theoretic Puzzles, Misc (94)
 - Newcomb's Problem (127)
 - Pascal's Wager* (213)
 - St. Petersburg Paradox (34)
 - Toxin Puzzle (37)
 - Two-Envelope Paradox (61)
- Game Theory** (1,223 | 385)
 - Convention and Coordination (87)
 - Evolutionary Game Theory (174)
 - Game-Theoretic Principles (27)
 - Game Theory and Ethics (55)
 - Game Theory and Political Philosophy (69)
 - Prisoner's Dilemma (283)
 - Normative and Descriptive Game Theory (39)
 - Game Theory, Misc (104)
- Topics in Decision Theory** (1,504 | 189)
 - Decision Theory and Ethics (131)
 - Infinite Decision Theory (72)
 - Normative and Descriptive Decision Theory (125)
 - Utility (292)
 - Decision Theory and Hypothesis Testing (89)
 - Preferences in Decision Theory (389)
 - Desire as Belief* (44)
 - Psychology of Decision* (56)
 - Transformative Experience* (13)

Jobs in this area

- Princeton University**
Assistant Professor
- Eastern Kentucky University**
Chair of the Department of History, Philosophy, and Religious Studies
- University of St. Andrews**
Research Fellow

Jobs from PhilJobs

PhilPapers subject directory for Decision Theory

Example Puzzle

Imagine that Chooser is going to choose 1 or 2, and there is a Demon who is very good at predicting Chooser's choice.

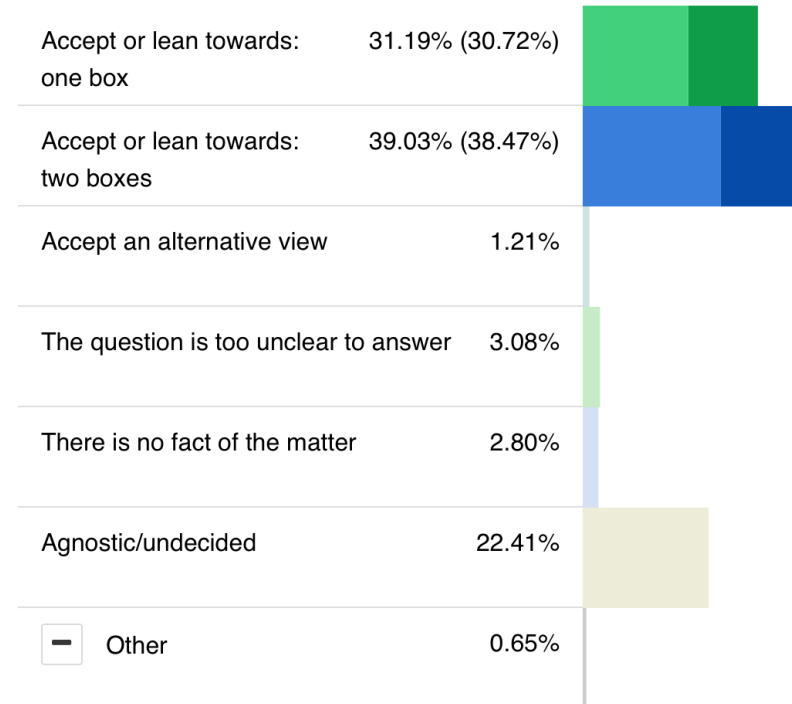
Chooser's payoff is a function of their choice and Demon's prediction, as shown on this table.

Newcomb's
Problem

	P1	P2
1	100	0
2	101	1

Survey Data

Newcomb's problem: one box or two boxes?



N = 1071, excluding skipped & insufficiently familiar (714 respondents)

PhilPapers subject directory for Decision Theory

Two Big Questions

1. How do we turn those noun phrases into sentences? What exactly is the question here, and what are the answers saying?
2. Why is it a worthwhile question to ask and answer?

General Puzzle

What question is decision theory asking, when it puts forward tables like this and offers ‘solutions’ to them?

What Question

Decision theory is about trying to describe what a certain kind of idealised decider will do.

The idealisation here is more like the idealisations of science than of ethics.

When we talk about an **ideal decider**, that's more like talking about an **ideal gas** than about an **ideal advisor**.

Why Care About That Question

Negative claim

Not because it helps us make decisions in anything like normal circumstances.

Positive claim

Because it helps us predict and understand what people will do in an interesting range of cases.

Aside on David Lewis

Lewis defended a third answer to the questions.

His decision theory was about articulating part of the theory of rationality that goes into the theory of mental content, via the notion of constitutive rationality.

I don't entirely agree, but I'm not going to engage with that here. His answers are closer to mine than the ones I'm mostly opposing here.

Getting at the Question

Methodology

- Figure out what question decision theorists must be asking by looking at what answers they give.
- Concentrate on questions that everyone (in the field) answers the same way.
- I said earlier there was a four-way disagreement on one particular puzzle; here I'll turn to questions where all the parties to that debate agree.

Betting Example

- 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.

Betting Example

- Given standard assumptions about how much Chooser likes money, all the decision theories I'm discussing say Chooser should not bet.
- So decision theory is not in the business of answering this question:
- *What action will produce the best outcome?*

Axiology

- We do have a discipline in philosophy that is all about evaluating outcomes: axiology.
- It's a worthwhile project.
- But it's not what decision theorists are up to.

Why Not Axiology?

Intuitive Answer

It isn't very practical in this case. Chooser can't bet on the winner.¹

¹. Actually, it's a bit more complicated than that, but let's stick with this characterisation for now.

Helpful Advice

- Philosophical decision theory is not in the business of providing helpful advice to choosers.
- We can see this by another example.



- Task: find the shortest path that goes through each of these cities.

Answer

- All the decision theories I'm discussing say that one should choose the path that's actually shortest.
- That's not particularly helpful advice!

Helpful Advice

Now as it turns out there are various helpful things you can say here.

Farthest Insertion Algorithm

Start with an arbitrary city. At each stage, add a city to the path by finding the point to insert it into the path that will add the least distance. The city you add should be the one **farthest** from the existing path.



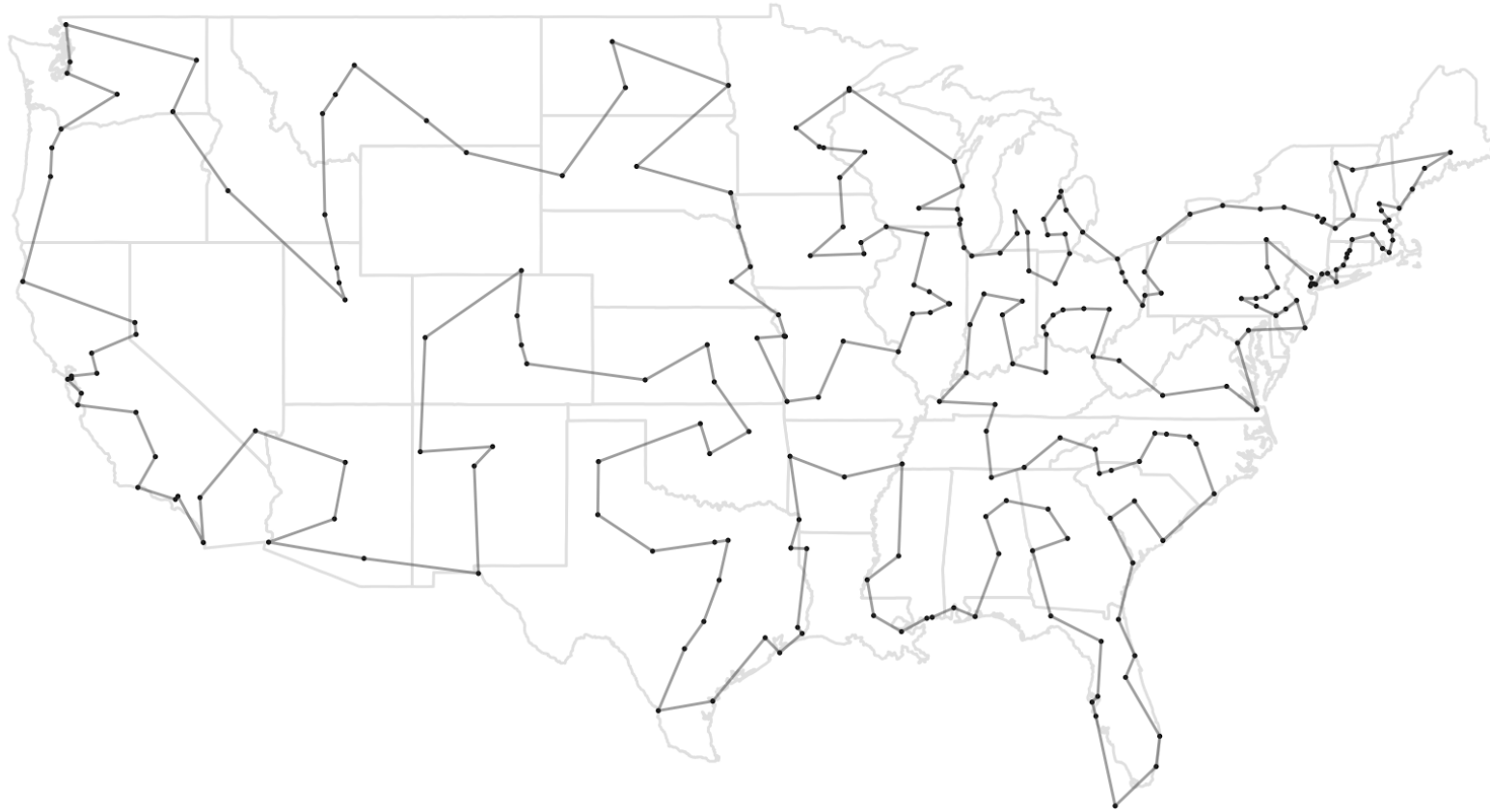
Tour length: 21075 miles.

Not a bad path, but not the best.

More Good Advice

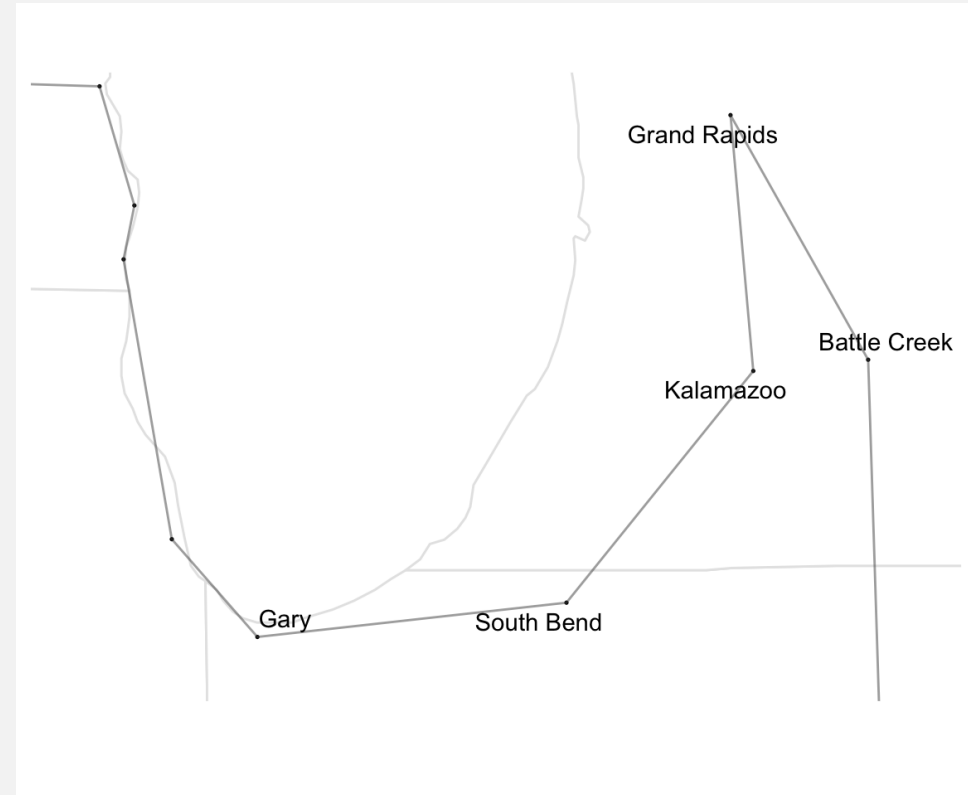
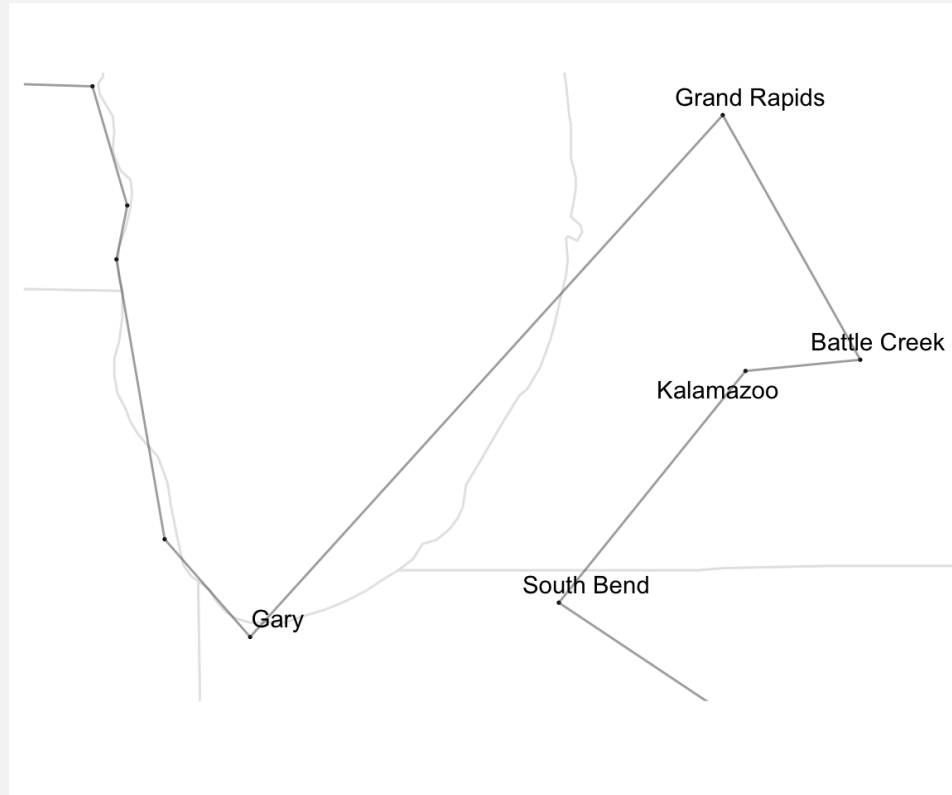
Delete pairs of edges and find the optimal replacement for that pair, until there are no benefits from doing this deletion.

- This will give you something very close to the original, but typically a bit shorter. And here it knocks off a few hundred miles.



Tour length: 20891 miles.

Making Adjustments



Salesman

For someone with normal amounts of computing power available, what they should do when faced with a Salesman problem is something like:

1. Run an insertion algorithm; and
2. Run a pairwise deletion and optimisation algorithm on the results of 1.
3. If they have the time and resources, repeat 1 and 2 a few times to see if they get a luckier draw.

This will typically not optimise, but it's what they should do.

Unlimited Computing

- If you have unlimited computer power, you could brute force your way through all $257!$ paths.
- Or even with a bit less computing power (but with more mathematical knowledge than before) you can come up with the following map.



Tour length: 20301 miles.

I'm not sure if this is best, but it's the best I could do, and it involved applying black box algorithms I don't understand.

Where We're At

Let's summarise these two cases in a table.

	Betting	Salesman
Best outcome	Bet on winner	Shortest path
Decision theory	Pass	Shortest path
Best advice	Pass	Learn algorithms

Decision theory is neither the theory of what is best to do, nor what is advisable to do.

So What Is It?

Imagine a version of Chooser with knowledge as it is, and computational powers as they might be.

- So for any mathematical problem, they can do it instantly.
- Ask, what would they do?

Decision Theory as Idealisation

They will pass in bet, and choose the shortest path.

- The mathematical work will be immense.
- They have to calculate the path length for each of $257!$ paths.
- And they have to find the minimum length among all of them.
- But setting computational costs to zero, this is easily do-able.

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.

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.¹

¹. Actually one step harder because of the multiplication, but otherwise the same.

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.

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.



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.

The General Theory of the Second Best, by R. G. Lipsey and Kelvin Lancaster, The Review of Economic Studies, 1956

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.¹

¹. Photo 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.