The End of Decision Theory

Anon

What question should decision theorists be trying to answer, and why is it worth trying to answer it? A lot of philosophers talk as if the aim of decision theory is to describe how we should make decisions, and the reason to do this is to help us make better decisions. I disagree on both fronts. The aim of decision theory should be to describe how a certain kind of idealised decider does in fact decide. And the reason to do this is that this idealisation, like many other idealisations, helps generate explanations of real-world behaviour. We shouldn't do what these ideal deciders do, or try to be more like them, because a lot of what they do only makes sense because of the differences between us and them. Still, sometimes those differences are small enough that they can be ignored in explanations, and that's when decision theory is useful.

Keywords: decision theory, model, idealisation, advice, second-best.

1 What is Decision Theory a Theory Of?

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

- Causal Decision Theory (Gibbard and Harper 1978; Lewis 1981; Skyrms 1990; Joyce 1999);
- Evidential Decision Theory (Ahmed 2014);
- Benchmark theory (Wedgwood 2013);
- Risk-Weighted theory (Buchak 2013);
- Tournament Decision Theory (Podgorski 2022); and
- Functional Decision Theory (Levinstein and Soares 2020)

Other theories haven't had snappy 'isms' applied to them, such as the non-standard version of Causal Decision Theory that Dmitri Gallow (2020) defends, or the pluralist decision theory that Jack Spencer (2021) defends, or the broadly ratificationist theory that Melissa Fusco (2024) 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 could the question be? That is, what question could decision theorsts be asking such that these are plausible answers to it?
2. Why is that an interesting question? What do we gain by answering it?

On 1, I will argue that decision theories should be understood as 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 the best decision theory tells us what decisions we should make.

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 George Akerlof (1970) 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 held that "central question of decision theory is: which choices are the ones that serve one's desires according to one's beliefs?" (Lewis [1989] 2020, 472). That's not far from the view I have, though I'd say it's according to one's evidence. But I differ a bit more from Lewis as to the point of this activity. For him, a central role for decision theory is supplying a theory of constitutive rationality to an account of mental content (Lewis 1994, 321–22). I think the resulting theory is too idealised to help there, and that's before we

[1] The 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.

get to questions about whether we should accept the approach to mental content that requires constitutive rationality. That said, the view I'm defending is going to be in many ways like Lewis's: the big task of decision theory is describing an idealised system, not yet recommending it.

The nine theories I mentioned above disagree about a lot of things. In philosophy we typically spend our time looking at cases where theories disagree. 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 could be 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 the best decision theory will be one that correctly describes an idealised version of actual deciders.

The resulting theory has a lot in common with the view that Joe Roussos has defended about ethics (Roussos (2022)) and, especially, formal epistemology (Roussos (2025)). He says that we should think of philosophical work in these areas as modeling rather than theorizing. I agree. If decision theory is a theory of anything, it's a theory of how some very strange creatures behave. Why we care about those creatures is not immediately obvious. The best reason, I'll argue, to care about these creatures is that they help us understand some aspects of the behaviour of real humans. Of course, real humans are not ideal. But sometimes they are close enough, in relevant respects, to the ideal that learning that idealised humans do something helps us understand why actual humans do it. This is to broadly agree with Roussos's main claims. If anything, I think the case for a view like his is even stronger in decision theory than in ethics or formal epistemology, and the point of this paper is to make that case.

2 Two Cases

2.1 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.

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.

2.2 Salesman

We'll focus on a version of what Julia Robinson (1949) called the travelling salesman problem.[2] 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 1.

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

[2] For a thorough history of the problem, see Schrijver (2005). For an accessible history of the problem, which includes these references, see the Wikipedia article on the Travelling salesman problem (2024).

[3] The 257 cities are the cities in the lower 48 states from the 312 cities in North America that John Burkardt (2011) mapped in his dataset USCA312.

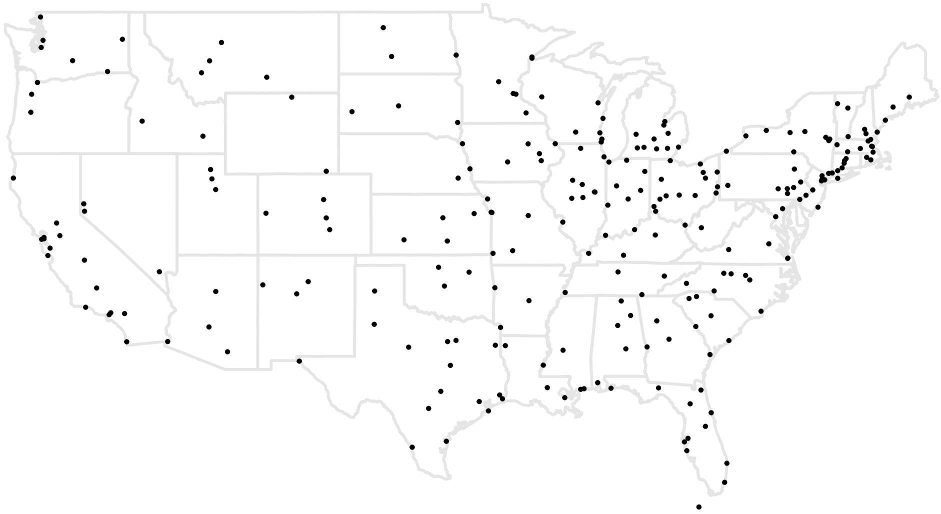

Figure 1: 257 American cites; our task is to find the shortest path that goes through all of them.

problems like this, you can't draw the shortest path through the map. At 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 (Mandelkern, Schultheis, and Boylan 2017).

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.[4] 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 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 2 shows the result of one output of this algorithm.[5]

[4] To implement both this algorithm and the optimisation I'll mention below, I've used the TSP package by Michael Hashler and Kurt Hornik (2007). The description of the two steps owes a lot to their summaries in the package documentation.

[5] The algorithm is silent on which city you start with, and typical implementations of it choose the starting city randomly.

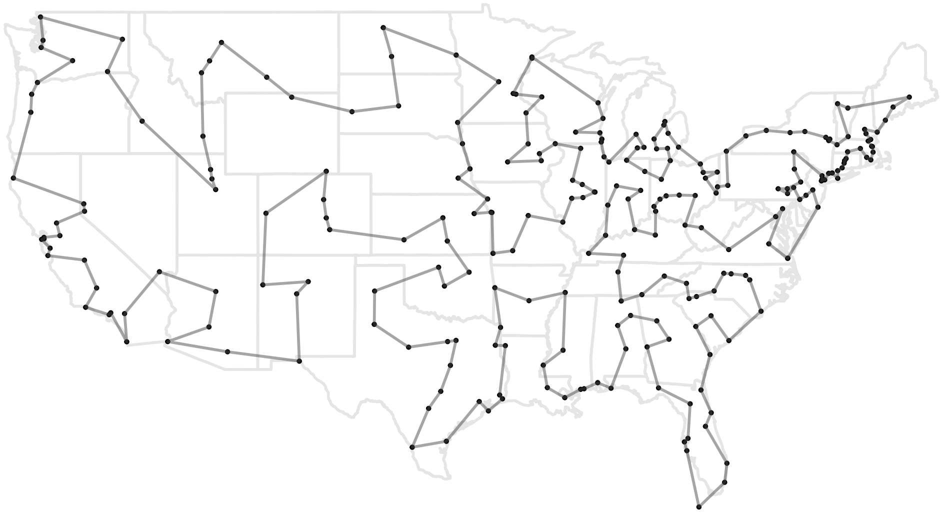

Figure 2: An output of the Farthest Insertion Algorithm, with a length of 21075 miles.

The path in Figure 2 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 3 is the output from applying this strategy to the path in Figure 2.

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. 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,

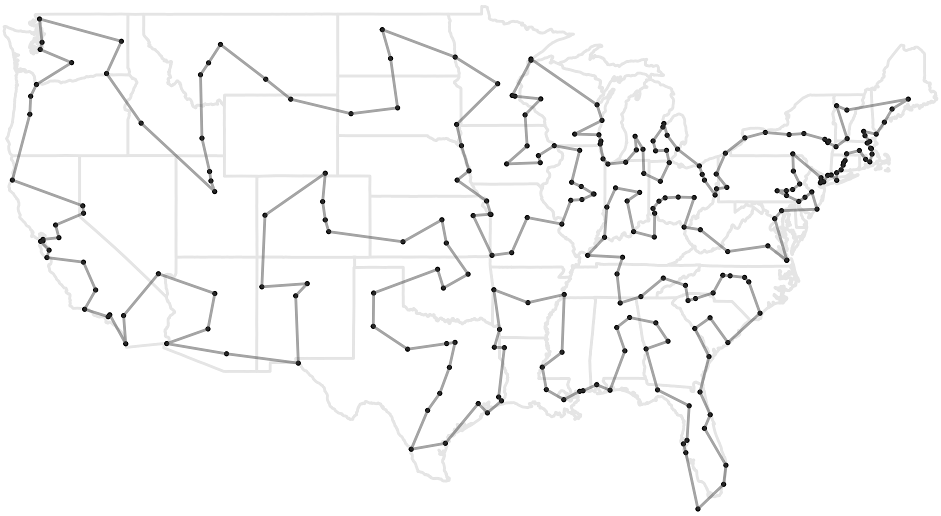

Figure 3: The output of an optimisation process, which reduced the path length to 20891 miles.

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 4.

I'm not sure if Figure 4 is as short as possible, but I couldn't find a shorter one. Still, for many purposes it wouldn't have been worth the trouble it took to find this map.

2.3 The Two Cases

Table 1 summarises the examples from the last two sections.

Table 1: How three approaches to decision theory handle the two cases

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

	Betting	Salesman
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 could it be doing?

It won't do to simply say that decision theory is a theory of rational choice. The person who uses a farthest insertion algorithm in Salesman, possibly supplemented with an edge-deletion optimisation, is being pretty rational. If we say that this person is being irrational, we must be using a somewhat non-standard notion of rationality. And now we're back to the problems I started with. What could that notion be, and why should we care about it? The next two section try to answer those questions.

3 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. Call this person IC, for Idealised Chooser.

Here's one important fact about IC: they decline to bet in Betting, and they choose the optimal path in Salesman. That is, they do exactly what the nine decision theories say. Why will they do that?

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

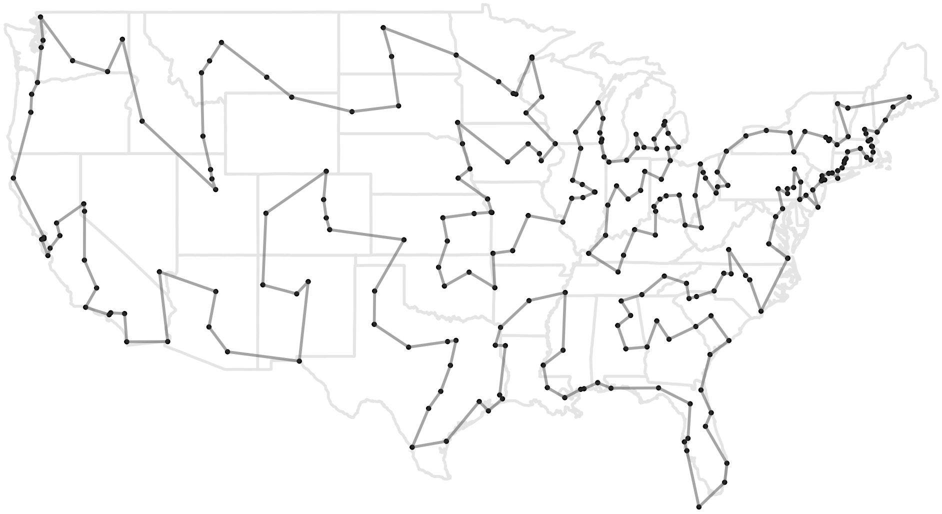

Figure 4: The shortest path I could find, with a distance of 20301 miles.

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 that 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.

Here's one important fact about IC: they decline to bet in Betting, and they choose the optimal path in Salesman. That is, they do exactly what the nine decision theories say. Why will they do that? In the betting case, adding unlimited computing power doesn't tell you who is going to win the game. So IC will still avoid betting. But in the Salesman case, adding unlimited computing power is enough to solve the problem. IC doesn'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 that 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 by hypothesis IC can do them freely and instantaneously, so they will do them, and get the right answer.

Dropping the idealisation, it's notable that if we say that Chooser should maximise expected utility, and we expect them to compute that, then we're asking Chooser to perform a task that is one step harder than calculating the shortest path in a Salesman problem. To calculate an expected utility, for each option one looks up a probability and a utility for each state[6], multiplies the two together, then adds the results to get a value for the option. One repeats that for each state, and finds an extreme value. Calculating the shortest path is exactly the same, except one only has to look up one number (a distance) rather than two (a probability and a utility), and there is no multiplication. Solving for the shortest path is strictly easier than finding the maximum expected utility. And yet finding the shortest path is practically impossible.

This is one reason I focussed on Salesman problems rather than other mathematical claims that Chooser is, in the standard models, assumed to know. I didn't ask Chooser to bet on the Twin Primes conjecture. It's possible one could come up with a model where finding the maximum expected utility is typically possible but resolving the Twin Primes conjecture is not; it's really hard to see how an agent who could always calculate expected utilities couldn't solve a Salesman problem.

There are two other things that are distinctively interesting about this problem which I'll simply note here, and defer longer discussion of them to another day. First, it is possible to give practical useful advice about how to solve Salesman problems. I've repeated some of the

[6] Exactly which probability it is, or indeed whether it even strictly is a probability, varies by which theory one chooses. But the basic idea that Chooser multiples something probability like by a utility is common across theories

better advice I've heard in the previous section. Second, when someone follows this advice and does badly, as can happen with carefully designed maps, it seems they are unlucky in just the same way that someone who maximises expected utility but gets a low amount of actual utility is unlucky. This raises some interesting questions about the normative significance of expected utility maximisation that will be in the background of the rest of the discussion here; hopefully I'll return to them in later work.

At this point you might complain that I've talked about decision theories asking Chooser to *calculate* expected utilities. They do no such thing. This is a point that Frank Knight made a century ago.

> Let us take Marshall's example of a boy gathering and eating berries ... We can hardly suppose that the boy goes through such mental operations as drawing curves or making estimates of utility and disutility scales. (Knight 1921, 66–67)

And Knight does not say this is irrational. As long as the boy gets enough berries, he's doing fine. In other terminology, we might say that decision theory provides a criteria of rightness, not a deliberation procedure.[7] As long as one follows the rules of decision theory, even if one follows them largely instinctually like Marshall's boy, one is rational.

Once again, this move just brings us back to the original problem. It's easy to understand the distinction in Sidgwick. The criterion of rightness is that one actually produces the best outcome. Which decision procedure actually produces that outcome is hard to determine in advance, though there are good reasons for suspecting that aiming for the best outcome as such is not the optimal procedure. Why, however, should we think that maximising *expected* utility is a criteria of rightness? What benefits does it have, over the standard of maximising actual utility, as such a criteria?

In Betting, a typical Chooser can maximise expected utility, while they can't maximise actual utility. Could that be the reason to say that maximising expected utility is the criteria of

[7] I'm taking this distinction from Peter Railton (1984), though his isn't the earliest use of the distinction. Alastair Norcross (1997) notes that the phrase "criterion of rightness" is used in the context of drawing this distinction by Sidgwick (1907, bk. 4, Chapter 1, §1).

rightness? Hardly. After all, in Salesman, Chooser can neither maximise expected nor actual utility. There must be some other reason that don't bet is the right answer, and not just a useful answer, in Betting.

One reason we might suppose that these theories provide the right answer is that expected utility maximisation, or whatever one's favourite decision theory endorses, is a goal; it is something we should try to achieve. On this picture, decision theory is relevant because it tells us what our ideal selves are like, and it recommends we try to be like them. In practice we can't always be like them, as in the Salesman problem, but we should try.

The problem with this answer is that it is not, in general, good to try to be like the ideal. The key point goes back to Lipsey and Lancaster's *General Theory of the Second Best* (Lipsey and Lancaster 1956). Often times, the right thing to do is something whose value consists in mitigating the costs of our other flaws. It's not true in general, indeed it's rather rare that it's true in practice, that approaches which differ from the ideal in one respect are better than all approaches which differ from the ideal in two respects. For example, us non-ideal agents 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. After all, stopping is costly, and the ideal agent gets no gain from incurring that cost.

In general, we differ from the ideal agent in any number of ways. Some of these are respects in which we'd be better off being more like them. For example, they hedge against costly but realistic risks, and typical humans don't take out as many such hedges as they should. But some of these are respects in which we'd be worse off being more like them. For instance, they never stop to have a think, or put in effort to get better at calculations. 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 hedging rather than more like never trying to get better at calculating. That, unfortunately, is not something which we can really figure out from within the idealised approach to decision theory that is standard these days.

Let's recap. So far I've argued that what decision theories do is describe what characters like

IC are like. I haven't yet said anything about why describing IC is a worthwhile activity. I've argued that a few attempts to provide a broadly normative explanation of this activity don't work. Learning more about IC does not provide useful advice, or tell us which of the options we can do are best, or provide us with a goal worth aiming at. I'll argue in the next section that the reason to learn about IC is not to get better at decision making, but to get better at understanding decision makers.

4 Idealisations as Models

At the start I said that the word 'idealised' gets used differently in ethics and in philosophy of science. The main claim I want to make in this section is that we should understand the idealisations in decision theory in the latter sense. In particular, we should understand them as simplifications. Michael Weisberg (2007) identifies three kinds of idealisations in science: Galilean, which distort the situation to make computation easier; minimalist, which only include the factors one takes to be causally significant to a situation; and multiple models, where one tries to understand a situation by considering different minimal idealisations with different strengths and weaknesses. The idealisations in decision theory are the second kind. They aren't particularly computationally tractable, unlike the Galilean idealisations, and there is typically just the one of them.

The idealisations in minimalist models like, say, ideal gas theory, are *simplifications* rather than *perfections*. We do not think that having volume is an imperfection. Maybe some religious traditions think this, but it isn't baked into introductory chemistry. Nor do we think that they are things we should aim for. Introductory chemistry does not imply a *Smaller the better!* rule for molecules. Rather, it says that volumeless molecules with perfectly elastic collisions are simpler to work with, and that some of the phenomena of real gases can be explained by looking at this simpler model.

We can make sense of what decision theory, as a discipline, is doing if we take it to be engaged in the same style of project. (Whether we can make sense of what individual decision theorists are doing this way is a harder question, one I'll return to below.) Just like the point masses we use in the ideal gas law, decision theories provide a good description of a certain kind of simplification. The idealisation need not be a perfection for this to be interesting. Indeed, even the idealised creatures studies in decision theory are not completely perfect. They have similar informational limitations to what we do.

This is the point of the basketball example. The idealised self that gets used in decision theory is god-like god-like in one respect - computational ability - but human-like in another - informational awareness. That's a common feature of idealised models; one doesn't idealised away from absolutely everything.

That still doesn't tell us why we build these models. Part of the reason is similar to the reason we ever use minimal models. Minimal models are explanatorily powerful when the difference between the minimal model and reality is not relevant to predicting, explaining, or understanding what happens in the real world. The same thing is true in decision theory. The idealised models of decision theory are, at least sometimes, relevant to predicting, explaining, or understanding what happens in the real world.

When could such an idealised account be relevant? They are relevant when the differences between real people and idealised people are small, relative to the size of the problem. It's tempting to identify these cases with high stakes situations. After all, in high stakes situations deciders are disposed to throw enough computational resources at the problem that the differences between ordinary people and ideal agents shrinks. But that is isn't quite right. After all, in many high stakes cases, the decider also throws enough investigative resources at the problem that holding actual knowledge fixed is a bad modelling assumption.

Instead, the cases where decision theory is most helpful for modelling real life situations are ones where there are principled limitations to the decider's informational capacities. There are two kinds of cases where there are such principled limitations. One is where the information concerns the future, and the decision must be made now. And the other is where the information that someone else has (or at least may have) just as much incentive to suppress the information as the decider has to find it. Most textbook examples of the usefulness of decision theory concern the first kind, though they don't always make explicit why it matters that the case is future directed. I'm going to work through a case of the second kind that I think is enlightening about the way decision theory is valuable.

Until very recently, used cars sold at a huge discount to new cars, even when the cars were just a few months old with almost no usage. For a long time there was no agreed upon explanation for this phenomenon. The most common theory was that it reflected a preference, or perhaps a prejudice, on the part of buyers. George Akerlof (1970) showed how this discount could be explained in a model of perfectly rational agents. His model makes the following assumptions.

1. Cars vary a lot in quality, even cars that come from the same production line.
2. Sellers of used cars know how good the particular car they are selling is.
3. Buyers of used cars do not know how good any token car is; they only know how good that type of car generally is.
4. People rarely sell cars they just bought.
5. Everyone involved is an expected utility maximiser.

Based on these five assumptions, Akerlof built a formal model of the market for recently used cars. In the model, the most common reason to sell a car one just bought is the discovery that it was a bad token of that type of car. Knowing that this was the main reason to sell, buyers of used cars demanded a big discount in exchange for the possibility they were buying a dud. But as long as there are enough forced sellers of good recently purchased cars, who prefer whatever money they can get for their car to keeping the car, there can be an equilibrium where lightly used cars sell at a heavy discount to new cars, and it is rational for (some) owners to sell into this market, and for (some) buyers to buy in this market.

If Akerlof was right, and I think he was largely correct, you'd expect the used car discount to fall if either of the following things happened. First, it would fall if production lines got more reliable, and cars off the same line were more similar to one another. And second, it would fall if buyers had access to better tools[8] to judge the quality of used cars. By 2020 both of those things had happened, and the used car discount was almost zero.[9]

The philosophical significance of this is that one can't build models like Akerlof's without a theory of rational action under uncertainty. The usefulness of philosophical decision theory is that it's an essential input to useful models, like the Akerlof model. Since those models are useful, getting the inputs to them right is useful.

5 Theories and Theorists

So far I've argued that decision theory is best understood as the project of developing models of agents that are useful in certain kinds of explanations. I argued that it isn't obvious what question could have the answers "Don't bet in Betting! Be perfect in Salesman!", and in particular what question that we should care about could have those answers. My suggestion is that we should take the question to be about what a certain kind of idealised agent does, and we should care about it because it's a useful input to explanations.

But, one might object, this isn't what individual decision theorists take themselves to be doing. Some decision theorists, perhaps most of them, disagree with the arguments in Section 3 that decision theory does not have normative implications. They take themselves to be saying how ordinary people should act, or offering advice to ordinary people. Isn't this in tension with my claims about the aims of decision theory as a practice? I think it isn't, for four reasons.

First, there isn't any inconsistency between saying that an institution has an aim, and denying that this is the aim of most people who make up the institution. Still, we'd like to have more to say than just that this is consistent.[10]

Second, if I'm right about the aims of decision theory, you'd expect that this would typically *not* show up in everyday practice. To see this, imagine we have two decision theorists, Daniels

[8] Better that is than a drive around the block test drive.

[9] Then during the pandemic very strange things happened in the used car market and the 'discount' arguably went negative. Whatever was happening there was not explained by the Akerlof model.

[10] Though if no one disagreed with my claims about the aim of the institution, this paper would be useless for a different reason.

and O'Leary, and Daniels agrees with me about the aims of decision theory, while O'Leary says decision theory is about how ordinary people should act. The two of them can still debate ordinary cases, and can still say that this or that decision would be the ideal one to make. They might mean different things by 'ideal', but that won't matter for most purposes. It's harmless enough for Daniels to describe his idealised agent as 'rational', and if he does there will be even less conflict with O'Leary. There will be some differences between them, and I'll describe some of them soon, but for the most part they won't show up, even if the view Daniels and I have is correct.

Third, there are cases where what O'Leary says is correct. While it's not *always* true that ordinary people should resemble the ideal more than they actually do, it is *sometimes* true. So Daniels can even agree on a case by case basis that some arguments in decision theory do show us what ordinary people should do. He'll just think that the argument from decision theory to claims about ordinary rationality has one extra step, namely that in the case in question ordinary people should try to be more like the ideal, in the respect being discussed. This isn't always true, as Salesman shows, but it is often enough so obviously true that it's fine to leave it out of the argument. For instance, as mentioned above, when thinking about what kinds of insurance to buy, it's a fair bet that one should do is what the idealised agent in decision theory actualy does.

Fourth, even if we're interested in modelling, there is a reason that we'd be interested in the kinds of agents who are computationally perfect, and hence that O'Leary would take to be a rational role model. If we spend a lot of resources coming up with the optimal model for agents with our actual computational limitations, then we risk being superseded when computational capacities change. And given that most computation these days is outsourced to machines, and the machines are getting better literally every year, that's a very real risk. Looking at the cases when we can treat actual computational resources as close enough to ideal means that we don't have to throw our work out every time Moore's Law kicks in. Put another way, in my story of Daniels and O'Leary, Daniels has an incentive from within his own theory to work on problems where he and O'Leary won't disagree.

So there are a lot of reasons that Daniels and O'Leary wouldn't have to get into disputes about the point of decision theory, even while doing decision theory. Relatedly, Daniels can take O'Leary's first-order work within decision theory to be excellent contributions to the modelling project he's engaged in, even if O'Leary takes himself to be engaged in a different, normative, project, and Daniels thinks that project rests on implausible foundations.

But there will ultimately be some disagreements between them, not just about the point of decision theory, but about its practice. I'll end this paper with a discussion of what those might be, and why Daniels's side of the disagreement is plausible.

6 Consequences

On my view, standard approaches in decision theory provide a kind of minimal model. It is rare that only one minimal model is suitable for a whole area of study. The ideal gas law is useful, but it is more useful when supplemented with theories about when it breaks down, and about what happens in those cases. The same is true in decision theory.

So on the modelling approach to decision theory, you'd like to see the standard kinds of models, which make those distinctive recommendations in Betting and Salesman, supplemented by other models. We are already seeing that in philosophy in work on limited awareness (Steele and Stefánsson 2021). And we are seeing it in economics in work on cursed equilibrium (Eyster and Rabin 2005). On the modelling approach, these are not rivals to the standard model, but useful supplements to it.

The previous paragraph mentioned two kinds of models where we drop some of the idealisations in the standard model. There are also interesting cases where we add in even more idealisations.

If one takes decision theory to be giving a normative theory of rational choice, it's a very hard problem to say what the options are that Chooser has to choose between (Hedden 2012). The problem is that given the structure of the theory, we want to only have options that Chooser will certainly carry out if they choose them, and that's hard to guarantee. We can't say that Chooser chooses to take a holiday in Paris, because the flight might be cancelled, or Chooser might have a medical emergency, and so on. It's tempting to say that Chooser only ever chooses to try to do things. But this is seriously unintuitive. On the modelling approach, things are easier. We can just say that taking a holiday in Paris is one of Chooser's options, and have the conditional *If Chooser chooses Paris, they will go to Paris* be a harmless enough idealisation.

More controversially, I think the same argument can be given for an idealisation that is standard in economics, but not in philosophy: the availability of mixed strategies. In economics, it is usually assumed that if A and B are among Chooser's options, so is any probabilistic mixture of A and B. Philosophers frequently describe decision problems where this is explicitly ruled out. After all, they say, Chooser might not have any randomizing device available. Setting aside the vexed question of whether such a device is needed to carry out a mixed strategy, a question which turns on just what we take a mixed strategy to be, the fact that such a device *might* not be available isn't relevant on the modelling approach. As long as it is a reasonable idealisation that such a device is available, and as long as Chooser has a smartphone it surely is, it will make sense to include the ability to use it in Chooser's idealised capacities.

But, a philosopher might continue, couldn't we also ask the question of what Chooser should do if they can't randomise? In practice, this usually comes up in the context of games that roughly resemble playing rock-paper-scissors against a near perfect predictor (Spencer and Wells 2019). In those cases if you can't randomise I think the best advice is to learn to lose graciously. But why couldn't we say more? Why shouldn't decision theory apply to them? On the modelling approach I favour, denying someone access to mixed strategies is like denying them access to free and limitless computation. There are things we can say with such a limitation imposed, e.g., it's good to learn insertion strategies in Salesman. But what we don't say is that standard decision theory should apply in such a case. The modelling approach opens up the possibility, a possibility I suspect we should take, of treating the ability to randomize perfectly as on a par with the ability to calculate perfectly. It's a frequently useful idealisation, and it's helpful to impose it to understand some behaviour.[11] But once it is dropped, we are engaged in a different kind of project to standard decision theory.

That's not to say that's a bad project to engage in. Quite the contrary, it's an excellent project. In Section 5 I mentioned one reason Daniels might want to be cautious about spending too many resources on particular kinds of optimisations projects for computationally limited agents: the agents will probably get more capacities by the time the project is done. While that's a good reason for caution, there are two kinds of projects in that vein that are more promising. David Thorstad (2024) has a good recent book that addresses both of these. One is the question of what the criterion of rightness is for decision making under computational limitations.[12] A second is whether there are things which can be said about how to manage the limits that are imposed by the architecture of our cognitive systems.

The point I'm making is not that the projects described in the last paragraph are valuable. That's something O'Leary could, and probably would, agree with. Rather, my claim is that they are more continuous with ordinary decision theory than is often appreciated. What I want to reject is the division between "ideal" decision theory, as exemplified by the nine theories mentioned at the top, and "non-ideal" decision theory, as exemplified by work like Thorstad's. On that picture, the theories output different kinds of oughts, or perhaps one outputs an ought and the other a purely descriptive claim. On my view, both kinds of theory involve different kinds of idealisation. (Really most social science involves some idealisation; it's hard to say anything without simplifying somewhere.) The difference between the theories is not how they judge us, but in what they explain. Neither theory is particularly promising as a universal theory of human behaviour. But both theories are very useful as explanations of particular phenomenon, and that should be good enough.[13]

[11] For a really nice example of this, see the explanation of bidding behaviour on oil rights in Sutton (2000).

[12] I favour a version of reliabilism here, but I won't argue for that in this paper.

[13] **Declarations**: I have no conflicts of interest to report. I used Anthropic's LLM Sonnet 4.5 to check the document for spelling errors and to help with the code used to produce the figures.

References

Ahmed, Arif. 2014. *Evidence, Decision and Causality*. Cambridge: Cambridge University Press.

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.

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

Burkardt, John. 2011. "Cities." https://people.sc.fsu.edu/~jburkardt/datasets/cities/cities.html.

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

Fusco, Melissa. 2024. "Absolution of a Causal Decision Theorist." *Noûs* 58 (3): 616–43. https://doi.org/10.1111/nous.12459.

Gallow, J. Dmitri. 2020. "The Causal Decision Theorist's Guide to Managing the News." *The Journal of Philosophy* 117 (3): 117–49. https://doi.org/10.5840/jphil202011739.

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.

Hahsler, Michael, and Kurt Hornik. 2007. "TSP—Infrastructure for the Traveling Salesperson Problem." *Journal of Statistical Software* 23 (2): 1–21. https://doi.org/10.18637/jss.v023.i02.

Hedden, Brian. 2012. "Options and the Subjective Ought." *Philosophical Studies* 158 (2): 343–60. https://doi.org/10.1007/s11098-012-9880-0.

Joyce, James M. 1999. *The Foundations of Causal Decision Theory*. Cambridge: Cambridge University Press.

Knight, Frank. 1921. *Risk, Uncertainty and Profit*. Chicago: University of Chicago Press.

Levinstein, Benjamin Anders, and Nate Soares. 2020. "Cheating Death in Damascus." *Journal* *of Philosophy* 117 (5): 237–66. https://doi.org/10.5840/jphil2020117516.

Lewis, David. 1981. "Causal Decision Theory." *Australasian Journal of Philosophy* 59 (1): 5–30. https://doi.org/10.1080/00048408112340011.

———. 1994. "Reduction of Mind." In *A Companion to the Philosophy of Mind*, edited by Samuel Guttenplan, 412–31. Oxford: Blackwell. https://doi.org/10.1017/CBO9780511625343.019.

———. (1989) 2020. "Letter to Jonathan Gorman, 19 April 1989." In *Philosophical Letters of David K. Lewis*, edited by Helen Beebee and A. R. J. Fisher, 2:472–73. Oxford: Oxford University Press.

Lipsey, R. G., and Kelvin Lancaster. 1956. "The General Theory of Second Best." *Review of Economic Studies* 24 (1): 11–32. https://doi.org/10.2307/2296233.

Mandelkern, Matthew, Ginger Schultheis, and David Boylan. 2017. "Agentive Modals." *Philosophical Review* 126 (3): 301–43. https://doi.org/10.1215/00318108-3878483.

Norcross, Alastair. 1997. "Consequentialism and Commitment." *Pacific Philosophical Quarterly* 78 (4): 380–403. https://doi.org/10.1111/1468-0114.00045.

Podgorski, Aberlard. 2022. "Tournament Decision Theory." *Noûs* 56 (1): 176–203. https://doi.org/10.1111/nous.12353.

Railton, Peter. 1984. "Alienation, Consequentialism, and the Demands of Morality." *Philosophy and Public Affairs* 13 (2): 134–71.

Robinson, Julia. 1949. "On the Hamiltonian Game (a Traveling Salesman Problem)." Santa Monica, CA: The RAND Corporation.

Roussos, Joe. 2022. "Modelling in Normative Ethics." *Ethical Theory and Moral Practice* 25: 865–89. https://doi.org/10.1007/s10677-022-10326-4.

———. 2025. "Normative Formal Epistemology as Modelling." *British Journal for the Philosophy of Science* 76 (2): 421–48. https://doi.org/10.1086/718493.

Schrijver, Alexander. 2005. "On the History of Combinatorial Optimization (till 1960)." *Handbooks in Operations Research and Management Science* 12: 1–68. https://doi.or g/10.1016/S0927-0507(05)12001-5.

Sidgwick, Henry. 1907. *The Methods of Ethics*. Seventh. London: Macmillan.

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

Spencer, Jack. 2021. "An Argument Against Causal Decision Theory." *Analysis* 81 (1): 52–61. https://doi.org/10.1093/analys/anaa037.

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.

Steele, Katie, and H. Orri Stefánsson. 2021. *Beyond Uncertainty: Reasoning with Unknown Possibilities*. Elements in Decision Theory and Philosophy. Cambridge University Press.

Sutton, John. 2000. *Marshall's Tendencies: What Can Economists Know?* Cambridge, MA: MIT Press.

Thorstad, David. 2024. *Inquiry Under Bounds*. Oxford University Press.

Travelling salesman problem. 2024. "Travelling Salesman Problem— Wikipedia, the Free Encyclopedia." https://en.wikipedia.org/w/index.php?title=Travelling_salesman_problem&oldid=1209291065.

Wedgwood, Ralph. 2013. "Gandalf's Solution to the Newcomb Problem." *Synthese* 190 (14): 2643–75. https://doi.org/10.1007/s11229-011-9900-1.

Weisberg, Michael. 2007. "Three Kinds of Idealization." *The Journal of Philosophy* 104 (12): 639–59. https://doi.org/10.5840/jphil20071041240.