The End of Decision Theory

Anon

What question are decision theorists 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 the decision theory is 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.[1] 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)

[1]Word count. Total: 6608. Body text plus footnotes: 5816; Abstract: 150; References: 552; Other: 90.

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.[2] 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.

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

[2]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.

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 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 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 is true in the right kind of idealisation.

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 rather unclear. If it is a model of how humans behave when certain constraints are not significant, then it is clear what we are doing. I'm calling this idealisation rather than modeling, but this is as much a terminological difference as anything else; I 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.[3] 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.

[3]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).

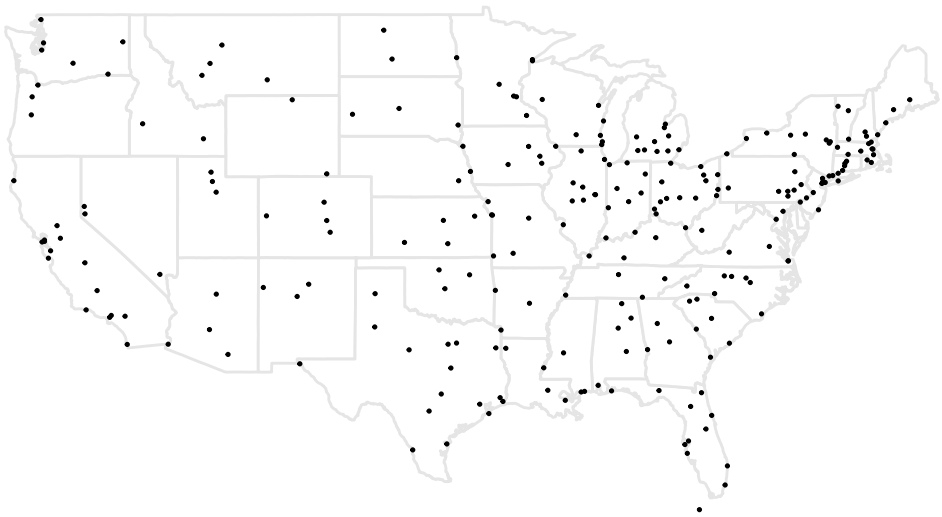

Figure 1: 257 American cites; our task is to find the shortest path that goes through all of them.

The task is to find the shortest path through those 257 cities.[4]

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 (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.[5] 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

[4]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.

[5]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.

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.[6]

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

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

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

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, 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.

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

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
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?

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.

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 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.

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[7], 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

[7] 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

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 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.[8] As long as one follows the rules of decision theory, even if one

[8] 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).

follows them largely instinctually like Marshall's boy, one is rational.

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? It is a somewhat easier rule to use, which makes it a better deliberation procedure. Unfortunately, as the Salesman cases show, there are other procedures that are better again qua deliberation procedures too. So what benefit does it have?

One possible answer to this challenge 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 idealised people 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 correctly hedge against costly but realistic risks. 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.

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.

Another way to put this is that the idealisations in, 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.

Decision theory is engaged in the same kind of project. 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. The idealisation here is not a perfection, for two reasons. First, allo-

cating zero seconds to hard but important math problems is not a perfection, it's a practical vice. Yet it's what the ideal agent does. Second, the idealised self is not in fact absolutely 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.

Why do we use these models? Part of the answer here comes back to the much discussed question of why we use models at all. I'm going to assume that part of the answer is that 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. So my hypothesis is that the idealised models of decision theory are, at least sometimes, relevant to predicting, explaining, or understanding what happens in the real world.

It's tempting to identify the situations where decision theory is relevant 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.

To find a case where decision theory is relevant, we need are cases where there are principled limitations to the decider's informational capacities. There are two kinds of cases that are relevant here. 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 phenomena. 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 the car is; they only know how good that model of cars 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 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. 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[9] to judge the quality of used cars. By 2020 both of those things had happened, and the used car discount was almost zero.[10]

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

[10] 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.

The philosophical significance of this is that one can't build models like Akerlof's without a theory of rational action under uncertainty. The big payoff 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 Conclusions

This has largely been a work of meta-philosophy. I've argued that decision theorists are building idealisations in the sense of simple models. And I've argued that this is a good project not because it issues in advice, or evaluation, but because it provides inputs to explanations. In particular, decision theoretic explanations are often accurate when people can behave somewhat like computationally ideal agents, but must still behave like informationally limited agents.

If I'm right, there are several consequences for first-order decision theory. I'll end the paper going over four of them.

5.1 The Value of Limited Theories

We use different styles of explanations for different phenomena. If a product routinely sells for $7.99, we might use a rational choice explanation for why the price is roughly $8 rather than roughly $10, and then a very different kind of explanation for why it is $7.99 rather than $8.01. It isn't always a weakness of an explanation that it does not generalise to as many cases as one might have hoped.

This matters for decision theory. If a decision theory goes silent on a certain kind of case, that isn't necessarily a bad thing. One sometimes hears theorists talk as if the fact that a theory doesn't say what to do in a particular situation is very bad, because the point of decision theory is to provide advice. But if decision theory goes silent on cases where we don't think decision theoretic explanations are likely to be good, that's not necessarily a bad thing.

5.2 The Ideal Agent

I've left off a lot of details about exactly what the ideal agent is like. I said they are computationally good, but informationally limited. This leaves open a lot of questions. Do they have perfect information about their own beliefs and desires, or about their own plans? Are they able to stick to a plan, and if so, which kinds of plans can they stick to?

One way to try answering these questions is by asking whether the inability to know one of these things, or do one of these things, is a kind of imperfection. If it is, we've discovered a new feature of the ideal, perfect agent.

If I'm right, that's the wrong way to go about answering the question. We should ask instead if assuming that the ideal decider has these features makes them too dissimilar to real people for explanatory purposes. For instance, I think we should allow that ideal deciders can play mixed strategies, because being able to play mixed strategies does not make the ideal decider that different to real people. In circumstances where real deciders have sufficient computational resources that ideal deciders are good models for them, real deciders also have sufficient resources to play mixed strategies.

Whether I'm right or wrong about mixed strategies, the point I want to really stress is the approach to answering these questions about idealisations. The right idealisation does not describe what we should be like, but rather what it is helpful to model us as being like.

5.3 Non-Ideal Theory

If actual decision theory is a kind of ideal theory, that means there is a space for a non-ideal theory. And there are a bunch of interesting philosophical questions about it. I think the right non-ideal theory will be some kind of reliabilism. Even if that's right, it hardly settles matters. There are, after all, many different kinds of reliabilism, and we'd need to have answers to the decision theoretic equivalents of the generality problem, and the new evil demon problem. Still, these feel like answerable questions, and there are interesting projects to work on here.

5.4 Reconciliation

If two types of theory exist, both ideal and non-ideal, some reconciliation possibilities open up. Perhaps the right thing to say about Newcomb's Problem is that there is a sense in which one should take one box, and there is a sense in which one should take two boxes. One way to get this result would be to endorse the following three claims.

1. The right ideal decision theory is some broadly causal decision theory.
2. The right non-ideal decision theory is some kind of reliabilism.
3. In Newcomb's Problem, one boxers and two boxers are in the same reference class, so the right thing to do is the thing that, on average, produces the best results in that large class.

I don't want to endorse all these; I'm particularly sceptical of 3. The point is just that when we distinguish ideal from non-ideal theories we open up some new options in what might seem like stale debates.

5.5 Other Idealisations

The most interesting question that opens up from this way of thinking about decision theory is whether we could develop any other idealisations that are explanatorily powerful. As Weisberg notes, an important kind of idealisation involves developing many models that help explain different phenomena. Here that might involve changing what information the ideal agent has, or what computational powers they have.

One natural further idealisation is to model deciders as having not just rational credences, but true beliefs, about certain domains. In practice we do this a bit. Standard models of consumer choice assume consumers know the prices of different goods, and know their own preference structure, rather than merely assuming they have rational beliefs about these things. Standard practice in decision theory is to assume that the decider knows what options are available. If we try to have that fall out of a general practice of assuming they are rational, we end up with

difficult choices about what counts as an option (Hedden 2012). It's simpler to treat knowledge of options as a distinct, but useful, idealisation.

Could we weaken the assumption of costless and perfect computation? In economics there has been some interesting projects along these lines. One that's particularly relevant here is the development of cursed equilibrium models (Eyster and Rabin (2005)). In cursed equilibrium models, agents maximise expected utility with respect to some information, but not the information they actually have. In particular, they don't always take fully into account what they can figure out about other people's information from observing the acts other people perform. It's a bit more complicated than this in practice, but roughly it's as if people ignore what other people are doing.

These models are relevant here for two reasons. One is that the main example I used of decision theory working, Akerlof's model for used cars, involved people making just the kind of inference that they do not make in cursed equilibrium models. The reason the used car market settles at such a discount, in an Akerlof model, is that buyers reason from the fact that sellers are choosing to sell that sellers have private information. That inference, from observed behaviour to conclusions about the private information the other person has, is exactly what agents do not make in cursed equilibrium models. This matters because in a bunch of experimental settings, cursed equilibrium models make more accurate predictions than rational choice models.

This doesn't on its own show the Akerlof explanation is wrong. It might just show that explanation was incomplete. To complete the explanation we could simply add the premises that cars are expensive, and that people act more carefully when making expensive purchases. The first premise is clearly true, cars are indeed expensive, and there is some evidence for the second. Still, thinking about cursed equilibrium models, which are still incredibly idealised, helps both explain new phenomena, and appreciate more fully the explanations that rational choice models make.

Cursed equilibrium models have not been developed nearly as fully as rational choice models; it's only very recently that fully dynamic versions of them have been put forward (Cohen and Li

2023; Fong, Lin, and Palfrey forthcoming). I certainly don't want to say this is the only way to modify the idealisations in standard decision theory, or even the best such way. What I do want to say is that thinking about decision theory as the project of building good simplified models suggests that the project of building multiple models of decision could have some value.

Alt Text for Diagrams

- Figure 1: "A map of the 48 contiguous states of the USA, with dots at the locations of 257 major cities."
- Figure 2: "Figure 1 plus a jagged path through the 257 cities."
- Figure 3: "A similar drawing to Figure 2 with a smoother path."
- Figure 4: "Another path through the 257 cities with a different path, the shortest one the author knows."

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.

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

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

Fong, Meng-Jhang, Po-Hsuan Lin, and Thomas R. Palfrey. forthcoming. "Cursed Sequential

Equilibrium." *American Economic Review*, forthcoming.

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.org/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.

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.