Deliberation Costs*

Brian Weatherson

2020

Our theory of rational choice should be sensitive to deliberation costs. It is irrational to take into account minor differences between goods, if the cost of taking those differences into account is greater than the expected gain from doing so. It has often been held in economics that this line of reasoning will lead to an infinite regress. I argue that the regress can be stopped if we take the rational chooser to be skilled at attending to the right information. On the appropriate model of skill, the rational agent will attend to the right information without reasoning about whether this is the right information to attend to.

Humans making decisions face two big limitations. First, we are informationally limited. We don't know everything and sometimes we don't know what we need to know in order to make the optimal decision. Second, we are computationally limited. We can't process all of the information that we have available to us before a decision needs to be made. Or at least, we can't do this in a costless manner.

Orthodox decision theory treats these two limitations very differently. To a first approximation, the whole point of orthodox decision theory is to handle the question of how to make decisions without full information. But on the other hand, orthodox decision theory simply assumes away the computational limitations. Orthodox decision theory is a theory of rational choice, and rationality is here understood to involve not being subject to these pesky computational limitations.

I think this is a serious mistake. In particular, I think there are several cases where our theory of rational choice can only give us the correct verdict if we allow it to be sensitive to both kinds of limitation. In this paper, I will discuss three such kinds of cases, and describe how rational choice theory might be revised so as to handle them.

I'm far from the first to notice this asymmetry in how orthodox decision theory handles the two limitations. For approximately as long as decision theory has existed, there have been people who have noted the oddity of ignoring computational limitations. But there has always been a powerful argument against taking computational limitations seriously. It is long been thought that attempt to do this would lead to a nasty kind of regress. It isn't entirely clear how the regress argument here is supposed to run; the argument is more often alluded to than carefully stated. But it is a major challenge and I will have something to say about it.

*Unpublished draft. Thanks to audiences at Michigan, Toronto, and the Ranch Metaphysics Workshop for valuable feedback, not all of which I've yet incorporated.

The short version of what I'm going to say is that while we should take both kinds of limitations seriously, we should treat them differently in our final theory. We should, as orthodox decision theory says, take a broadly evidentialist approach to informational limitations. That is, good decision makers should have credence distributions over the possibilities left open by their evidence, those credences should be sensitive to the evidence they have, and the choices they make should maximize expected value given those credences. On the other hand, we should take a broadly reliabilist approach to computational limitations. Good decision makers will adopt procedures for managing their own limitations that reliably produce good outcomes. There is no requirement that they adopt the procedures that are best supported by their evidence. The reason there is no requirement they do that is that figuring out what those reliable procedures might be is even more computationally taxing then the problem of deciding what to do. And if we're going to respect the fact that people can't always complete difficult computational tasks, we shouldn't expect them to perform the incredibly difficult task of figuring out how to adjust their decision procedures in light of the evidence about their own limitations.

You might think that the reason orthodox theory treats computational limitations this way is that it is simply trying to provide a theory of ideal decision making. There is a separate question, to be sure, of how non-ideal agents should make decisions. But the thought, or at least the hope, is that clearly stating what the ideal looks like will help the non-ideal agents in this task. I think there is a little reason to believe that this hope will be realized. In general, knowing what the ideal looks like provides us with very little guidance as to how to get better. Knowing that any ideal outcome has a certain attribute does not provide a reason, even a defeasible reason, for trying to to acquire that attribute (Lipsey and Lancaster 1956).

We can see this by simply thinking about the one limitation that orthodox theory does take seriously. A good decision maker without full information will in general behave nothing like a good decision maker with full information. For example, if you put the informationally limited agent in a casino they will do the exact opposite of what an informationally unlimited agent will do. The informationally unlimited agent will play every game and do quite well at them. The informationally limited agent, on the other hand, will play none of the games because they all have negative expected returns. I think is the general case. It's a bad idea to emulate the ideal agent, because us non ideal agents often have to act so as to minimize the damage that have other limitations can do. Everyone agrees that is true in the case of informational limitations, and I am going to try and argue that it's also true for computational limitations.

So here's the plan for the paper. First, in sections 1-2, I will introduce the three kinds of cases but I think motivate taking computational limitations seriously. Then, in sections 3-4, I will introduce the regress argument that is alleged to show that any attempt to do this will end badly. In sections 5-6, I will show how the broadly reliabilist approach to handling computational limitations that I favor can be motivated, and can avoid the regress. Sections 7 and 8 are contingent speculations about how non-ideal agents might choose reliably, and observations on how these debates connect to other

philosophical debates

Before I start on this, it's helpful to get clear on exactly what I am taking my orthodox opponent to be committed to. I take them to endorse the following three constraints on a theory of rational choice.

1. Rational agents have credences, and these credences are responsive to evidence.
2. These credences also respect the probability calculus.
3. Rational agents take actions that maximize their expected utility given these credences.

There are a lot of questions that I do not take my orthodox opponent to have a settled view on, though of course many orthodox theorists will have one view or another on one or other of these questions. These questions include

- Whether rationality puts any constraints on what can be valued;
- Whether our theory of rationality divides up failures to make rational choices into epistemic failures, axiological failures, and practical failures, and if it does make such a division, exactly how it should be made;
- Whether rationality requires that agent be self-aware, i.e., whether they know what their own credences and utilities are; and
- Exactly what evidence is, or what it means for credences to be responsive to evidence.

My hope is that I can provide an objection to orthodoxy that is insensitive to how orthodox theorists answered these questions. That's a rather ambitious project, since the answers one gives to these questions will help provide responses to some of the objections I shall offer. But I'm not going to try to anticipate every possible response the orthodox theorist could make. Indeed, I don't think that I've got anything like a knock down watertight argument against all possible versions of orthodoxy. What I think I do have is a set of reasons to consider an alternative, and an outline of what that alternative may look like.

1 Three Puzzles

1.1 Puzzle One - Close Calls

Let's start with an example from a great thinker. It will require a little exegesis, but that's not unusual when using classic texts.

> Well Frankie Lee and Judas Priest
> They were the best of friends
> So when Frankie Lee needed money one day
> Judas quickly pulled out a roll of tens
> And placed them on the footstool
> Just above the potted plain

> Saying "Take your pick, Frankie boy,
> My loss will be your gain."
> ("The Ballad of Frankie Lee and Judas Priest", 1968.
> Lyrics from Bob Dylan (2016) 225)

On a common reading of this, Judas Priest isn't just asking Frankie Lee how much money he wants to take, but which invididual notes. Let's simplify, and say that it is common ground that Frankie should only take $10, so the choice Frankie Lee has is which of the individual notes he will take. This will be enough to set up the puzzle.

Assume something else that isn't in the text, but which isn't an implausible addition to the story. The world Frankie Lee and Judas Priest live in is not completely free of counterfeit notes. And it would be bad for Frankie Lee to take a counterfeit note. It won't matter just how common these notes are, or how bad it would be. But our puzzle will be most vivid if each of these are relatively small quantities. So there aren't that many counterfeit notes in circulation, and the (expected) disutility to Frankie Lee of having one of them is not great. There is some chance that he will get in trouble, but the chance isn't high, and the trouble isn't any worse than he's suffered before. Still, other things exactly equal, Frankie Lee would prefer a genuine note to a counterfeit one.

Now for some terminology to help us state the problem Frankie Lee is in. Assume there are k notes on the footstool. Call them n_1, ..., n_k. Let c_i be the proposition that note n_i is counterfeit, and its negation g_i be that it is genuine. And let t_i be the act of taking note n_i. Let U be Frankie Lee's utility function, and Cr his credence function.

In our first version of the example, we'll make two more assumptions. Apart from the issue of whether the note is real or counterfeit, Frankie Lee is indifferent between the notes, so for some h, l, $U(t_i \mid g_i) = h$ and $U(t_i \mid c_i) = l$ for all i, with of course $h > l$. If we add an extra assumption that Frankie Lee thinks the probability that each of the notes is genuine is the same, we get the intuitive result back that he is indifferent between the banknotes.

But is that really a plausible move? Here is one way to start worrying about it. Change the example so that the country Frankie Lee and Judas Priest live in is very slowly modernising its currency. It is getting rid of old fashioned, and somewhat easy to counterfeit, paper money, and joining the civilised countries that use plastic money. Moreover, plastic bank notes are, for all intents and purposes, impossible to counterfeit. (At least, no one has yet figured out how to do it, and Frankie Lee knows this.)

Some of the notes Judas Priest offers are the new plastic notes, and some are the old paper notes. Now it seems clear that Frankie Lee should take one of the new notes, and not merely on aesthetic grounds. Rather, the fact that the plastic notes are less likely to be counterfeit is a reason to prefer to take them. And this is true no matter how unlikely it is that the paper notes are counterfeit, as long as this likelihood is non-zero.

But now go back to the base case, where all the money is paper. A small change in probability of being counterfeit seems to be enough to give Frankie Lee a reason to prefer some of them to the others. Indeed, the only way for him to be indifferent between the notes is if the probability of any one being counterfeit is exactly the same

as the probability of any other being counterfeit. But that two of the notes have exactly the same probability of being counterfeit is a measure zero event. It isn't happening. So Frankie Lee shouldn't be indifferent between the notes.

Of course, if the notes look exactly the same, then the probability that each is counterfeit is exactly the same. But that's only because that probability is one. In that case Frankie Lee should run away as fast as possible. That's not the realistic case.

The realistic case is that the notes look a little different to each other in ever so many respects. (Including, one hopes, their serial numbers.) Some will be a little more faded, or a little more torn, or a little more smudged or crumpled, than the others. It is overwhelmingly likely that these fades, tears, smudges, spills etc are the result of the normal wear and tear on the currency - wear and tear that paper notes tend to wear on their face. But every imperfection in every note is some evidence, very very marginal evidence but still evidence, that the note is counterfeit. And since Frankie Lee's evidence, on any extant theory of evidence, includes visible things like the tears, smudges etc on the notes, they are pieces of evidence that affect the evidential expected utility of taking each note. So if Frankie Lee wants to maximize evidential expected utility, there is precisely one note he should take. Though it probably won't be obvious to him which one it is, so rationality requires Frankie Lee to spend some time thinking about which note is best.

This is intuitively the wrong result. (Though it is what happens in the song.) Frankie Lee should just make a choice more or less arbitrarily. Since expected utility theory does not say this, expected utility theory is wrong.

The Frankie Lee and Judas Priest case is weird. Who offers someone money, then asks them to pick which note to take? And intuitions about such weird cases cases are sometimes deprecated. Perhaps the contrivance doesn't reveal deep problems with a philosophical theory, but merely a quirk of our intuitions. I am not going to take a stand on any big questions about the epistemology of intuitions here. Rather, I'm going to note that cases with the same structure as the story of Frankie Lee and Judas Priest are incredibly common in the real world. Thinking about the real world examples can both show us how pressing the problems are, and eventually show us a way out of those problems.

So let's leave Frankie Lee for now, just above the potted plain, and think about a new character. We will call this one David, and he is buying a few groceries on the way home from work. In particular, he has to buy a can of chickpeas, a bottle of milk, and a carton of eggs. To make life easy, we'll assume each of these cost the same amount: five dollars.[1] None of these purchases is entirely risk free. Canned goods are pretty safe, but sometimes they go bad. Milk is normally removed from sale when it goes sour, but not always. And eggs can crack, either in transit or just on the shelf. In David's world, just like ours, each of these risks is greater than the one that came before.

David has a favorite brand of chickpeas, of milk, and of eggs. And he knows where in the store they are located. So his shopping is pretty easy. But it isn't completely

[1] If that sounds implausible to you, make the can/bottle/carton a different size, or change the currency to some other dollars than the one you're instinctively using. But I think this examples works tolerably well when understand as involving, for example, East Carribean dollars.

straightforward. First he gets the chickpeas. And that's simple; he grabs the nearest can, and unless it is badly dented, or leaking, he puts in in his basket. Next he goes onto the milk. The milk bottles have sell-by dates printed in big letters on the front. And David checks that he isn't picking up one that is about to expire. His store has been known to have adjacent bottles of milk with sell-by dates 10 days apart, so it's worth checking. But as long as the date is far enough in the future, he takes it and moves on. Finally, he comes to the eggs. (Nothing so alike as eggs, he always thinks to himself.) Here he has to do a little more work. He takes the first carton, opens it to see there are no cracks on the top of the eggs, and, finding none, puts that in his basket too. He knows some of his friends do more than this; flipping the carton over to check for cracks underneath. But the one time he tried that, the eggs ended up on the floor. And he knows some of his friends do less; just picking up the carton by the underside, and only checking for cracks if the underside is sticky where the eggs have leaked. He thinks that makes sense too, but he is a little paranoid, and likes visual confirmation of what he's getting. All done, he heads to the checkout, pays his $15, and goes home.

The choice David faces when getting the chickpeas is like the choice Frankie Lee faces. He has to choose from among a bunch of very similar seeming options. In at least the chickpeas example, he should just pick arbitrarily. But for very similar reasons to Frankie Lee, expected utility theory won't say that.

The standard model of practical rationality that we use in philosophy is that of expected utility maximization. But there are both theoretical and experimental reasons to think that this is not the right model for choices such as that faced by Frankie or David. maximizing expected utility is resource intensive, especially in contexts like a modern supermarket, and the returns on this resource expenditure are unimpressive. What people mostly do, and what they should do, is choose in a way that is sensitive to the costs of adopting one or other way.

There are two annoying terminological issues around here that I mostly want to set aside, but need to briefly address in order to forestall confusion.

I'm going to assume maximizing expected utility means taking the option with the highest expected utility given facts that are readily available. So if one simply doesn't process a relevant but observationally obvious fact, that can lead to an irrational choice. I might alternatively have said that the choice was rational (given the facts the chooser was aware of), but the observational process was irrational. But I suspect that terminology would just add needless complication.

I'm going to spend more time on another point that is partially terminological, but primarily substantive. That's whether we should identify the choice consequentialists recommend in virtue of the fact that it maximizes expected utility with one of the options (in the ordinary sense of option), or something antecedent. I'm going to stipulate (more or less) that it is consistent with consequentialism that the choice can be something antecedent - it can be something like a choice procedure. And I'm going to argue that this is what the rational consequentialist should choose.

I'm going to call any search procedure that is sensitive to resource considerations a satisficing procedure. This isn't an uncommon usage. Charles Manski (2017) uses the

term this way, and notes that it has rarely been defined more precisely than that. But it isn't the only way that it is used. Mauro Papi (2013) uses the term to exclusively mean that the chooser has a 'reservation level', and they choose the first option that crosses it. This kind of meaning will be something that becomes important again in a bit. And Chris Tucker (2016), following a long tradition in philosophy of religion, uses it to mean any choice procedure that does not optimize. Elena Reutskaja et al -Reutskaja et al. (2011) contrast a 'hybrid' model that is sensitive to resource constraints with a 'satisficing' model that has a fixed reservation level. They end up offering reasons to think ordinary people do (and perhaps should) adopt this hybrid model. So though they don't call this a satisficing approach, it just is a version of what Manski calls satisficing. Andrew Caplin et al (2011), on the other hand, describe a very similar model to Reutskaja et al's hybrid model - one where agents try to find something above a reservation level but the reservation level is sensitive to search costs - as a form of satisficing. So the terminology around here is a mess. I propose to use Manski's terminology: agents satisfice if they choose in a way that is sensitive to resource constraints.

Ideally they would maximize, subject to constraints, but saying anything more precise than this brings back the regress problem that we started with. Let's set it aside just a little longer, and go back to David and the chickpeas.

When David is facing the shelf of chickpeas, he can rationally take any one of them - apart perhaps from ones that are seriously damaged. How can expected utility theory capture that fact? I think if it identifies David's choices with the cans on the shelf, and not with a procedure for choosing cans, then it cannot.

It says that more than one choice is permissible only if the choices are equal in expected utility. So the different cans are equal in expected utility. But on reflection, this is an implausible claim. Some of the cans are ever so slightly easier to reach. Some of the cans will have ever so slight damage - a tiny dint here, a small tear in the label there - that just might indicate a more serious flaw. Of course, these small damages are almost always irrelevant, but as long as the probability that they indicate damage is positive, it breaks the equality of the expected utility of the cans. Even if there is no visible damage, some of the labels will be ever so slightly more faded, which indicates that the cans are older, which ever so slightly increases the probability that the goods will go bad before David gets to use them. Of course in reality this won't matter more than one time in a million, but one in a million chances matter if you are asking whether two expected utilities are strictly equal.

The common thread to the last paragraph is that these objects on the shelves are almost duplicates, but the most careful quality control doesn't produce consumer goods that are actual duplicates. There are always some differences. It is unlikely that these differences make precisely zero difference to the expected utility of each choice. And even if they do, discovering that is hard work.

So it seems likely that, according to the expected utility model, it isn't true that David could permissibly take any can of chickpeas that is easily reachable and not obviously flawed. Even if that is true, it is extremely unlikely that David could know it to be true. But one thing we know about situations like David's is that any one of the (easily

reached, not clearly flawed) cans can be permissibly chosen, and David can easily know that. So the expected utility model, as I've so far described it, is false.

1.2 Puzzle Two - Psychic Costs of Bias

In all but a vanishingly small class of cases, the different cans will not have the same expected utility. But figuring out which can has the highest expected utility is going to be work. It's possible in principle, I suppose, that someone could be skilled at it, in the sense that they could instinctively pick out the can whose shape, label fading, etc., reveal it to have the highest expected utility. Such a skill seems likely to be rare - though I'll come back to this point below when considering some other skills that are probably less rare. For most people, maximizing expected utility will not be something that can be done through skill alone; it will take effort. And this effort will be costly, and almost certainly not worth it. Although one of the cans will be ever so fractionally higher in expected utility than the others, the cost of finding out which can this is will be greater than the difference in expected utility of the cans. So aiming to maximize expected utility will have the perverse effect of reducing one's overall utility, in a predictable way.

The costs of trying to maximize expected utility go beyond the costs of engaging in search and computation. There is evidence that people who employ maximizing strategies in consumer search end up worse off than those who don't. Schwartz et al. (2002) reported that consumers could be divided in 'satisficers' and 'maximizers'. And once this division is made, it turns out that the maximizers are less happy with individual choices, and with their life in general. This finding has been extended to work on career choice (Iyengar, Wells, and Schwartz 2006), where the maximizers end up with higher salaries but less job satisfaction, and to friend choice (Newman et al. 2018), where again the maximizers seem to end up less satisfied.

There are two things that can go wrong when you try to maximize. Maximising requires considering the strengths and weaknesses of each of the choices. That means, it requires giving at least some consideration to the negative attributes of what you end up choosing. And these can cause you to be less happy with the actual choice when those negative attributes are realized. And it also means giving consideration to the positive attributes of the choices not made. And this could lead to regret when you have to adopt a choice that lacks those positive attributes. So there are two very natural paths by which the attempt to maximize could backfire, any incurs costs that wouldn't have been incurred by the person who simply makes an arbitrary choice.

There is evidence here that both these paths are realised, and that maximisers do indeed end up psychically worse off than satisficers. Now to be sure, there are both empirical and theoretical reasons to be cautious about accepting these results at face value. Whether the second path, from consideration of positive attributes of the non-chosen option to felt regret, is psychologically significant seems to be tied up with the 'paradox of choice' (Schwartz 2004), the idea that sometimes giving people even more choices makes them less happy with their outcome, because they are more prone to regret. But it is unclear whether such a paradox exists. One meta-analysis (Scheibehenne,

Greifeneder, and Todd 2010) did not show the effect existing at all, though a later meta-analysis finds a significant mediated effect (Chernev, Böckenholt, and Goodman 2015). But it could also be that the result is a feature of an idiosyncratic way of carving up the maximizers from the satisficers. Another way of dividing them up produces no effect at all (Diab, Gillespie, and Highhouse 2008).

The theoretical reasons relate to Newcomb's problem. Even if we knew that maximizers were less satisfied with how things are going than satisficers, it isn't obvious that any one person would be better off switching to satisficing. They might be like a two-boxer who would get nothing if they took one-box. There is a little evidence in Iyengar, Wells, and Schwartz (2006) that this isn't quite what is happening, but the overall situation is unclear.

But the philosophical questions here are a bit simpler than the psychological questions. Whether maximisers in general are subject to these two kinds of costs is a tricky empirical question. Whether there could be one maximiser who is subject to them, and who knows that they are, is a much easier question. Of course someone could be like that. Indeed, it seems beyond dispute that many real people are subject to these costs. The only empirical question is whether these people are a significant minority or a significant majority.

And all it takes for the philosophical question to be pressing is that some choosers are, and know that they are, disposed to incur these psychological costs if they consciously try to maximise expected value. Our theory of choice should have something to say to them, and orthodox theory is silent. Especially for choices that are intended to produce happiness, the happiness effects of the choice procedure itself should be taken into account. But orthodox theory ignores it.

1.3 Puzzle Three - Mathematical Challenges

For a final case, let's consider Kyla, a student taking a mathematics exam. It's getting towards the end of the exam and she's facing quite a bit a time pressure. She comes to a true false question, and she knows that she knows how to solve questions like it. But she also knows that there are other kinds of questions that she is better at solving under time pressure. And while this is just a true false question, the exam is set up so that she gets a large negative score if she gets the question wrong. The expected return of simply guessing is strongly negative.

The rational thing for Kyla to do is to go on to other questions and come back to this one if she has time. But orthodox theory doesn't allow for this. The probability of any mathematical truth is one. And it's part of orthodoxy that credences are supposed to be probability functions. So whatever the correct answer is, offering it will have positive expected utility given Kyla's credences, assuming those credences are rational.

So orthodoxy gets this choice, and all other choices that turn on mathematical ignorance, badly wrong. The case where Kyla simply has to decide whether to answer the question now or come back to it later is in some ways a relatively easy case. The really hard decisions are about how much time to allocate to solving various mathematical

problems, when there are both costs to spending time, and rewards to solving as many problems as possible. These can often be important decisions, and ones that our theory should have something to say about. But orthodoxy does not have anything to say. It's time to look for something else.

2 Dialectical Interludes

2.1 Interlude One - The Obvious Answer

By this time you might be expecting a relatively simple answer to this question. The problem is that the orthodox theorist was focussing on the wrong choice. We shouldn't focus on the choice to take this can of chickpeas or that one, or to answer true or false to this question. Rather, we should focus on the choice to choose one procedure or another. And the rational chooser will choose the procedure that is on average best.

That solves our cases quite well. The best procedure for Frankie Lee or David to adopt is to choose arbitrarily. Any other procedure will take time, and it's not going to be time well spent. The best procedure for the person wracked by regret at choices they didn't make is also to choose somewhat arbitrarily, before the regrets have time to embed. Conversely, the best procedure for Kyla is to skip any questions that she can't do quickly, and come back to them if it turns out she has time.

Given some very weak assumptions, *Maximise expected utility* will not be an optimal procedure in this sense. Actually it's ambiguous what it means to say someone should adopt the procedure *Maximise expected utility*, but however you disambiguate that, it's wrong. The procedure *Calculate what maximises expected utility then choose it* is not optimal, because the calculations may not be worth the effort. The procedure *Instinctively choose what maximises expected utility* is a very efficient procedure if it is available, but for most agents it isn't available. We should no more criticise agents who don't adopt it than we criticise agents who don't get to work by apparating.

I'm going to adopt a version of the view that the rational choice is the outcome of an optimal procedure. But I'm not going to adopt the most obvious version of this obvious answer. In particular, I'm not going to say that agents should adopt the procedure such that adopting that procedure maximises expected value. Rather, I think, they should adopt the procedure that maximises something like average value. We'll return to this in a bit, but first I want to clear up some other dialectical points.

2.2 Dialectical Point Two - Possible Orthodox Solutions

There are ways of tinkering with orthodoxy to avoid some of the problems that I raised in the previous section. For example, dropping the constraint that credences are probabilities would avoid giving the wrong answer in Kyla's case. And maybe, just maybe, there is a theory of evidence, or of evidential support, such that the evidential expected utility of each of Frankie Lee's choices are not distinct. I'm certainly not going to try to go through every possible theory of evidence, or of evidential support, to show that this isn't the case.

But I do want to note three constraints on an orthodox solution to the problems that I have raised.

First, the solution must handle all the cases. This is not a completely trivial point. The reason orthodoxy fails in the three cases is a little different in each case. There is not, at least as far as I can tell, a simple way to handle all of them simultaneously while staying roughly within orthodoxy.

Second, the solution must not introduce any more complications of its own. For example, you could try to solve some of the problems by saying that the decision maker's evidence includes just those facts that are immediately available to her. Perhaps there is some sense of 'immediacy' in which this provides the start of a solution to the first two puzzles. (I think the third puzzle won't be solved this way, but the first two might.) But this solution introduces problems of its own. For example, a decision can be irrational in virtue of the fact that a moment's thought would've revealed to the decision maker that it will lead to disaster. If we restrict evidence to what is available at less than a moment's thought, then we get this case wrong. If we don't put such a restriction in place, then we're back to having problems with the first two puzzles I mentioned above. I don't want to clean there is nothing for the orthodox theorist to do here, but I do think it will be tricky to handle the puzzles without licencing irrational thoughtlessness.

Third, any orthodox solution should be just as simple and as well motivated as the obvious answer I just discussed. Saying that we should focus on procedures and not on the choices they lead to on an occasion resolves all of these puzzles in a simple and natural way. Even if an orthodox solution can be found to all three puzzles, if it requires three different changes to the orthodox view, it's hard to believe that it will be preferable to a simple solution in terms of procedures.

2.3 Dialectical Point Three - Terminology

At this point, some people might want to simply stipulate that the word "rational" picks out the choice that a computationally ideal actor would take. Even if it's good in some sense for David to choose arbitrarily, there is still an ideal can to choose, and he only deserves the honorific rational if he chooses it.

I am not going to get into a fight over terminology here. If people want to continue inquiring into what David would ideally do, then I'm not going to get in their way. But I found this inquiry unmotivated for three reasons. First, if we're going to consider what David would ideally do, then I'm more interested in what he'd do if he were really ideal, and knew everything. I don't see the appeal of investigating what he would do given one, but only one, kind of idealisation. Second, I don't think the ideal is a particularly good guide. Knowing what the gods do doesn't help the mortals, for mortals just get burned if they try to be like gods.

But the biggest reason concerns a purpose that I think is a central function of the concept of rationality. We have a need to make the people around us intelligible and predictable. And the best way we have to do this is to understand the constraints and the motivations of people around us, and feed those into a theory of rational choice

that outputs a decision given constraints and motivations. It doesn't always work, especially if you are trying to make predictions. But it beats most of the alternatives by a comfortable margin.

If that's the reason for having a theory of rational choice, then the orthodox theory is not fit for purpose. The person who stands in the grocery store aisle deliberating over which can to get is neither intelligible nor predictable. The theory that says rational agents adopt procedures that do well on average, given their constraints and motivations, does make the ordinary behavior of supermarket shoppers intelligible and predictable.

When I say 'we' need to make folks around us intelligible and predictable, I mean this to work at two different levels. From a very early age, we do this kind of reasoning about particular individuals to learn about the world (Scott and Baillargeon 2013). If a child sees a competent seeming adult use a particular method to solve a problem, and the adult does not seem to have any constraints that the child is free of, the child will copy what the adult does (Levy and Alfano 2020). This makes perfect sense; the adult is rational (and better informed than the child), so probably the adult's procedure is optimal for the child. If we know that children do this, we can exploit it to trick them. For example, we can demonstrate sub-optimal procedures, and children will mimic them for a surprisingly long time. But this isn't because the child is a fool; it's because they have a clever way of learning about the world that can misfire when people set out to confound it.

But I also mean this to work at the level of social analysis. The whole point of game theoretic explanations of social phenomena is that we can make a pattern of behavior intelligible by simply presenting the constraints and motivations of the choosers, and then showing how rational behavior on everyone's part produces the outcome. The research program this paper is a part of is motivated by the hope, and it is a hope more than a theory, that the same theory of rationality can serve both the child who is selectively imitating those around them, and the social scientist with their game theoretic models. Whether that's true in general or not, I think both the child and the theorist are better served by a theory of rational choice that is sensitive to computational limitations and deliberation costs. And it's their perspectives that I'm most interested in when theorising about rationality.

There is one other terminological dispute that I have no interest in entering into, but I need to make explicit in order to set aside. Some philosophers use 'decision theory' to refer to the study of purely procedural aspects of rationality. On this picture, there are three parts to rational choice: epistemology, axiology and decision theory. A rational agent will comply with all three. Compliance with the first is manifest in rational credences. Compliance with the second is manifest in rational values. And compliance with the third is manifest in choices that are rational given the first two. I don't much care for this highly factorised model of rational choice theory. Imagine we see someone punching themselves in the head, and ask why they are doing this. If they say, "I want to bring about world peace, and I believe this is the best way to do it", we don't reply, "Well, I guess two out of three isn't bad". We just think they are irrational. But for cur-

rent purposes I don't want to debate this. This paper is about the theory of rational choice. If you think that encompasses more than decision theory, that it also includes epistemology and axiology, then this isn't a paper in decision theory strictly speaking. But even someone who thinks the theory of rational choice can be factorised in this way still thinks there is a theory of rational choice. And my plan here is to offer a rival theory. Whether what I offer is a rival theory of *decision* turns on terminological questions about 'decision theory' that I'm hereby setting aside.

3 History and Regresses

The idea that rational people are sensitive to their own computational limitations has a long history. It is often traced back to a footnote of Frank Knight's. Here is the text that provides the context for the note.

> 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. What he does, in so far as he deliberates between the alternatives at all*, is to consider together with reference to successive amounts of his "commodity," the utility of each increment against its "cost in effort," and evaluate the net result as either positive or negative (Knight 1921, 66–67)

And the footnote attached to 'at all' says this

> Which, to be sure, is not very far. Nor is this any criticism of the boy. Quite the contrary! It is evident that the rational thing to do is to be irrational, where deliberation and estimation cost more than they are worth. That this is very often true, and that men still oftener (perhaps) behave as if it were, does not vitiate economic reasoning to the extent that might be supposed. For these irrationalities (whether rational or irrational!) tend to offset each other. (Knight 1921, 67fn1)

Knight doesn't really give an argument for the claim that these effects will offset. And as John Conlisk (1996) shows in his fantastic survey of the late 20th century literature on bounded rationality, it very often isn't true. Especially in game theoretic contexts, the thought that other players might think that "deliberation and estimation cost more than they are worth" can have striking consequences. But our aim here is not to think about economic theorising, but about the nature of rationality.

There is something paradoxical, almost incoherent, about Knight's formulation. If it is "rational to be irrational", then being "irrational" can't really be irrational. There are two natural ways to get out of this paradox. One, loosely following David Christensen (2007), would be to say that "Murphy's Law" applies here. Whatever one does will be irrational in some sense. But still some actions are less irrational than others, and the least irrational will be to decline to engage in deliberation that costs more than it

is worth. I suspect what Knight had in mind though was something different (if not obviously better). He is using 'rational' as more or less a rigid designator of the the property of choosing as a Marshallian maximizer does. And what he means here is that the disposition to not choose in that way will be, in the long run, the disposition with maximal returns.

This latter idea is what motivates the thought that rational agents will take what John Conlisk calls "deliberation costs" into account. And Conlisk thinks that this is what rational agents will do. But he also raises a problem for this view, and indeed offers one of the clearest (and most widely cited) statements of this problem.

> However, we quickly collide with a perplexing obstacle. Suppose that we first formulate a decision problem as a conventional optimization based on the assumption of unbounded rationality and thus on the assumption of zero deliberation cost. Suppose we then recognize that deliberation cost is positive; so we fold this further cost into the original problem. The difficulty is that the augmented optimization problem will itself be costly to analyze; and this new deliberation cost will be neglected. We can then formulate a third problem which includes the cost of solving the second, and then a fourth problem, and so on. We quickly find ourselves in an infinite and seemingly intractable regress. In rough notation, let P denote the initial problem, and let $F(.)$ denote the operation of folding deliberation cost into a problem. Then the regress of problems is P, $F(P)$, $F^2(P)$, ... (Conlisk 1996, 687)

Conlisk's own solution to this problem is not particularly satisfying. He notes that once we get to F^3 and F^4, the problems are 'overly convoluted' and seem to be safely ignored. This isn't enough for two reasons. First, even a problem that is convoluted to state can have serious consequences when we think about solving it. (What would *Econometrica* publish if this weren't true?) Second, as is often noted, $F^2(P)$ might be a harder problem to solve than P, so simply stopping the regress there and treating the rational agent as solving this problem seems to be an unmotivated choice.

As Conlisk notes, this problem has a long history, and is often used to dismiss the idea that folding deliberation costs into our model of the optimising agent is a good idea. I use 'dismiss' advisedly here. As he also notes, there is very little *discussion* of this infinite regress problem in the literature before 1996. The same remains true after 1996. What is done is that instead people appeal to the regress in a sentence or two to set aside approaches that incorporate deliberation cost in the way that Conlisk suggests.

Up to around the time of Conlisk's article, the infinite regress problem was often appealed to by people arguing that we should, in effect, ignore deliberation costs. After his article, the appeals to the regress comes from a different direction. It is usually from theorists arguing that deliberation costs are real, but the regress means it will be impossible to consistently incorporate them into a model of an optimizing agent. So we should instead rely on experimental techniques to see how people actually handle deliberation costs; the theory of optimization has reached its limit. This kind of move is

found in writers as diverse as Gigerenzer and Selten (2001), Odell (2002), Pingle (2006), Mangan, Hughes, and Slack (2010), Ogaki and Tanaka (2017) and Chakravarti (2017). And proponents of taking deliberation costs seriously within broadly optimizing approaches, like Miles Kimball (2015), say that solving the regress problem is the biggest barrier to having such an approach taken seriously by economists. So let's turn to how we might solve it.

4 Four Non-Solutions

My solution, as I've mentioned a couple of times, is a form of reliabilism. The rational choice is the one that would be produced by using the procedure that does best on average. That procedure will just be maximising expected utility when computational costs are zero, and will involve appeal to expected utility maximisation in many other cases. But it won't, in general, simply be expected utility maximisation.

To get clearer on what the reliabilist solution is, and how it is motivated, I want to first go through three other solutions that I don't think work.

First, we could just say that the rational choice is simply the choice that produces the best actual result. This gets some cases intutively wrong; it says that it is never rational to leave a casino without gambling. And it eliminates a type of choice that we think is real: the lucky guess. We want lucky guesses to be cases that produce good outcomes, but are not rational. If the rational choice just is the best choice, this is impossible. Since lucky guesses are not impossible, this theory can't be right.

Second, we could say that the rational choice is the choice that maximises expected value. But I've already gone over why that is wrong. There are really two things we could mean by saying the rational choice is the one that maximises expected value, and both of them are wrong. We could say that the rational choice is to compute what has the highest expected value, and then choose it. But this gives the wrong result in all the cases that I discussed at the beginning. Or we could say that the rational choice is to insinctively pick the one with the highest expected value. But there is no more reason to think this is something that choosers can do than there is to think that choosers can instinctively pick the choice with the highest actual value. So this is unrealistic.

Third, we could say the choice is the output of the procedure such that adopting that procedure maximises expected value, given one's evidence about the world and about the nature of procedures. Here, I think, the regress has bite because the same arguments from the previous paragraph still apply. We really need to distinguish two possible things we might mean by saying that one should adopt the procedure such that adopting it maximises expected value. First, we could mean that choosers should compute which procedure will maximise expected value, and then adopt it. But this will get the wrong result in Frankie Lee's case, and in David's. They shouldn't be doing any computation at all. So perhaps instead we could say that the rational chooser will instinctively choose the procedure with the highest expected value. But there is no more reason to think that choosers could always do that there is to think that they can simply choose the first-order option with the highest actual or expected value. So this idea fails,

and it fails for just the same reasons as the suggestion of the previous paragraph. That doesn't prove a regress is looming, but it doesn't look good.

The fourth option I'll discuss is designed to avoid this problem, and it is going to look somewhat more promising. Maybe we can't all at once choose the best procedure. But we can do it piecemeal.

In general, here's a way to adopt a complicated procedure. When faced with a certain class of problems, adopt the simplest procedure that agree with the complicated procedure over the range of choices you face. Then, as the problems expand, start either complicating the procedure, or adopt a meta-procedure for choosing which simple procedure to adopt on an occasion. Over time, if all goes well, you'll eventually adopt something like the complicated procedure, and do it without having to solve impossibly hard calculations about procedural effectiveness, or having miraculously good instincts.

One appeal of this approach is that it blocks the regress. If one selects a procedure piecemeal in this way, there is a good sense in which $F(P) = F^2(P) = F^3(P) = ...$ After all, there won't be a difference between adopting a procedure, and adopting a procedure for adopting that procedure. Both of them will just involve making the choices you have to make on a given day, and looking for the opportunity to integrate those choices into a larger and more systematic theory. By adopting a first-order procedure piecemeal, you also adopt a second-order procedure piecemeal. And if $F(P) = F^2(P) = F^3(P) = ...$, then the regress doesn't get going.

The problem is that this is too demanding. We want choosers to maximise. We don't expect them to be able to maximise over every possible choice situation, just over the one in front of them. If I'm buying chickpeas, and I arbitrarily choose one of the cans, that's all to the good. It's a rational choice. And, crucially, it stays being a rational choice even if I have dispositions to choose badly in other choice situations. But on the 'piecemeal' model being considered here, those dispositions to choose badly are partially constitutive of my choice procedure. And rational choice is a matter of choosing in virtue of adopting the correct choice procedure. So someone who is irrational somewhere is, it turns out, irrational everywhere. This is a bad result. There is something right about the idea that the rational chooser will just choose what's in front of them, and do so in a sensible way. But we shouldn't go on to say that rational choice requires that the global procedure one thereby implicitly adopts is the right one; that's too high a bar.

5 Skilled Choice

The way to see what's right about the last proposal, and to see our way to the correct solution, is to somewhat reconceptualise rational choice. We shoudl conceive of the rational chooser as a skilled chooser. And we should think skills are a matter of reliably doing well across realistic situations.

The justification for conceiving of rational choice as skilled choice is largely pragmatic. Thinking of rationality that ways results in a plausible theory of rationality, and other ways of thinking about rationality resulted in implausible theories. So rather than

argue for the conception of rationality as skill, I'm going to more or less assume it, and hope to justify this assumption by its fruitfulness. What I will argue for is the idea that skill involves reliably succeeding across realistic situations.[2]

Think for a bit about skilled athletes, or skilled players of chess or other games. Part of being skilled is succeeding. But it isn't just about success. Some people win due to luck. The skilled player won't always win, but they will reliably win across a range of situations.

Which situations are those? They are the situations that are normal enough for the kind of activity being engaged in. These might be dependent on highly contingent features of the activity. A chess player who wins international tournaments must be very skilled. We wouldn't retract that assessment if it turned out they only played well in quiet environments, and frequently lost chess games in noisy pubs. High level chess is played in quiet environments, so that's what matters.

A football player whose instincts only go right when there is no wind around is not particularly skilled. Someone who doesn't know how to adjust their passes when the wind changes is not skilled; it is lucky that they get ever connect on a pass. Conversely, a football player whose instincts are finely calibrated to the actual gravitational field strength around here could be highly skilled. It's not part of footballing skill that one is able to adjust to changes in gravitational field strength. Some kinds of flexibility, such as ability to adjust to wind conditions, matter, while others, such as ability to adjust to a different gravitational field, do not. There are intermediate cases where the importance of the ability to adjust is dependent on contingent attributes of the activity. Top level Australian Rules Football is almost always played at sea level. An Australian Rules Footballer whose instincts are calibrated for play at sea level, and who has no ability to adjust to changes in altitude, might still be highly skilled. But in a sporting competition where top flight games are frequently played in Mexico City or Quito, an inability to adjust to changing altitudes is a substantial limitation on one's skill. It is luck, not skill, that causes one to succeed in contests at one's favoured altitude. But it isn't luck that the Australian Rules Footballer is playing at sea level; that's a stable generalisation about the sport.

The same kind of story holds true for the skilled chooser. They have to do well, and not by chance. But that can involve having instincts that are calibrated to the environment one is actually in, and which would misfire in other environments. A skilled supermarket shopper need not be applying procedures that would do well in a medieval market. But they must be applying procedures that will keep working if the shelving of various items is changed.

That's to say, the skilled chooser will adopt a procedure that will, on average, produce the best results in circumstances like the ones they are in.There is an implicit notion of probability in that definition. But it isn't the notion of credence, or even of rational credence. Rather, it is the notion of how likely it is, or how frequent it is, that different circumstances obtain. That's the sense in which the theory is reliabilist.

When I say 'produce the best results', I mean the best results of the available pro-

[2]This is very similar to the modal understanding of skill in Beddor and Pavese (2020).

cedures. Just like we don't require rational commuters to apparate, we don't require rational choosers to instinctively maximise utility, or expected utility. They (just) have to do the best they can.

Skilled action frequently involves doing things where one has no evidence for the utility of such performances. It can even involve doing things where one has evidence against the utility of what one is doing. To see this, imagine a junior athlete who is thriving against competitors their own age with an unusual technique. They are told, by seemingly trustworthy coaches, that to thrive at higher levels, they have to adopt a more orthodox technique. But though they have reason to believe these coaches, they keep instinctively lapsing back into their unusual techniques. And, amazingly, the coaches are wrong, and what looked like a technique for winning against kids in parks ends up working at international level competition. (This isn't entirely unlike the story of Australian cricketer Steve Smith.) Such an athlete may be highly skilled. And their skill consists in, among other things, their instincts to do things that they have (misleading) evidence will not work. Their skill, that is, involves deploying a procedure that is actually reliable, even after they get evidence it is unreliable. I think the same is true of skilled choice. Sometimes, the skilled chooser will deploy a technique that they think is defective, and even one that they think is defective on reasonable grounds. As long as it works, it can still be the basis for skilled, and hence rational, choosing.

6 Regress Blocking

With all that in place, let's return to the regress problem, and in particular to Conlisk's statement of it. Why should we think the rational agent solves $F(P)$, and not $F^n(P)$ for some $n > 1$? I want to say that's just what rational choice is; it's skillfully managing one's own computational and informational limitations. And skill in this sense involves getting it right, and doing so reliably, not necessarily thinking through the problem. This suggests two questions.

1. Why should we allow this kind of unreflective rule-following in our solution to the regress?
2. Why should we think that $F(P)$ is the point where this consideration kicks in, as opposed to P, or anything else?

There are a few ways to answer 1. One motivation traces back to the work by the artificial intelligence researcher Stuart Russell (1997). (Although really it starts with the philosophers Russell cites as inspiration, such as Cherniak (1986) and Harman (1973).) He stresses that we should think about the problem from the outside, as it were, not from inside the agent's perspective. How would we program a machine that we knew would have to face the world with various limitations? We will give it rules to follow, but we won't necessarily give it the desire (or even the capacity) to follow those rules self-consciously. That might be useful some of the time - though really what's more useful is knowing the limitations of the rules. And that can be done without following the

rules as such. It just requires good dispositions to complicate the rules one is following in cases where such complication will be justified.

Another motivation is right there in the quote from Knight that set this literature going. Most writers quote the footnote, where Knight suggests it might be rational to be irrational. But look back at what he's saying in the text. The point is that it can be perfectly rational to use considerations other than drawing curves and making utility scales. What one has to do is follow internal rules that (non-accidentally) track what one would do if one was a self-consciously perfect Marshallian agent. That's what I'm saying too, though I'm saying it one level up.

Finally, there is the simple point that on pain of regress any set of rules whatsoever must say that there are some rules that are simply followed. This is one of the less controversial conclusions of the debates about rule-following that were started by Wittgenstein (1953). That we must at some stage simply follow rules, not follow them in virtue of following another rule, say the rule to compute how to follow the first rule and act accordingly, is an inevitable consequence of thinking that finite creatures can be rule followers.

So question 1 is not really a big problem. But question 2 is more serious. Why $F(P)$, and why not something else? The short answer will be that any reason to think that rational actors maximize *expected* utility, as opposed to *actual* utility, will also be a reason to think that they solve $F(P)$ and not P.

Start by stepping back and thinking about why we cared about expected utility instead of actual utility in the first place. Why not just say that the best thing to do is to produce the best outcome, and be done with it? Well, we don't say that because we take it as a fixed point of our inquiry that agents are informationally limited, and that the best thing to do is what is best given that limitation. Given some plausible assumptions, the best thing for the informationally limited agent to do would be to maximize expected utility. This is a second-best option, but the best is unavailable given the limitations that we are treating as unavoidable.

But agents are not just informationally limited, they are computationally limited too. And we could have instead treated that as the core limitation to be modelled. As Conlisk says, it is "entertaining to imagine" theorists who worked in just this way, taking the agents in their models to have computational but not informational limitations (Conlisk 1996, 691). Let's imagine that when we meet the Martian economists, that's how they reason. Conlisk notes a few things that the Martian economists might do. They might disparage their colleagues who take informational limitations seriously as introducing ad hoc stipulations into their theory. They might argue that informational limitations are bound to cancel out, or be eliminated by competition. They might argue that apparent informational limitations are really just computational ones, or at least can be modelled as computational ones. And so on.

What he doesn't add is that they might suggest that there is a regress worry for any attempt to add informational constraints. Let Q be the initial problem as the Martians see it. That is, Q is the problem of finding the best outcome given full knowledge of the situation, but the actual computational limitations of the agent. Then we suggest

that we should also account for the informational limitations. Let's see if this will work, they say. Let $I()$ be the function that transforms a problem into one that is sensitive to the informational limitations of the agent. But if we're really sensitive to informational limitations, we should note that $I(Q)$ is also a problem the agent has to solve under conditions of less than full information.[3] So the informationally challenged agent will have to solve not just $I(Q)$, but $I^2(Q)$, and $I^3(Q)$ and so on.[4]

Orthodox defenders of (human versions of) rational choice theory have to think this is a bad argument. And I think most of them will agree with roughly the solution I'm adopting. The right problem to solve is $I(Q)$, on a model where Q is in fact the problem of choosing the objectively best option. If one doesn't know precisely what one's knowledge is, then one has to maximize expected utility somewhat speculatively. But that doesn't mean that one shouldn't maximize expected utility.

But the bigger thing to say is that neither we nor the Martians really started with the right original problem. The original problem, O, is the problem of choosing the objectively best option. The humans start by considering the problem $I(O)$, i.e., P, and then debate whether we should stick with that problem, or move to $F(I(O))$. The Martians start by considering the problem $F(O)$, i.e., Q, then debate whether we should stick with that or move to $I(\mathrm{F}(O))$. And the answer in both cases is that we should move.

Given the plausible commutativity principle, that introducing two limitations to theorising has the same effect whichever order we introduce them, $I(F(O)) = F(I(O))$. That is, $F(P) = I(Q)$. And that's the problem that we should think the rational agent is solving.

But why solve that, rather than something more or less close to O? Well, think about what we say about an agent in a Jackson case who tries to solve O not $I(O)$. (A Jackson case, in this sense, is a case where the choice with highest expected value is known to not have the highest objective value. So trying to get the highest objective value will mean definitely not maximizing expected value.) We think it will be sheer luck if they succeed. We think in the long run they will almost certainly do worse than if they tried to solve $I(O)$. And in the rare case where they do better, we think it isn't a credit to them, but to their luck. In cases where the well-being of others is involved, we think aiming for the solution to O involves needless, and often immoral, risk-taking.

The Martians can quite rightly say the same things about why $F(O)$ is a more theoretically interesting problem than O. Assume we are in a situation where $F(O)$ is known to differ from O, such as the case Kyla was in. Or, for a different example, imagine the decision maker will get a reward if they announce the correct answer to whether a particular sentence is a truth-functional tautology, and they are allowed to pay a small fee to use a computer that can decide whether any given sentence is a tautology. The solution to

[3] At this point the Martians might note that while they are grateful that Williamson (2000) has highlighted problems with the KK principle, and these problems show some of the reasons for wanting to idealise away from informational limitations, they aren't in fact relying on Williamson's work. All they need is that agents do not exactly what they know. And that will be true as long as the correct epistemic logic is weaker than S5. And that will be true as long as someone somewhere has a false belief. And it would just be weird, they think, to care about informational limitations but want to idealise away from the existence of false beliefs.

[4] At this point, some of the Martians note that the existence of Elster (1979) restored their faith in humanity.

O is to announce the correct answer, whatever it is. The solution to $F(O)$ is to pay to use the computer. And the Martians might point out that in the long run, solving $F(O)$ will yield better results. That if the agent does solve problems like O correctly, even in the long run, this will just mean they were lucky not rational. That if the reward is that a third party does not suffer, then it is immorally reckless to not solve $F(O)$, i.e., to not consult the computer. And in general, whatever we can say that motivated "Rational Choice Theory", as opposed to "Choose the Best Choice Theory", they can say too.

Both the human and the Martian arguments look good to me. We should add in both computational and informational limitations into our model of the ideal agent. And that's the solution to the regress. It is legitimate to think that there is a rule that rational creatures follow immediately, on pain of thinking that all theories of rationality imply regresses. And thinking about the contingency of how Rational Choice Theory got to be the way it is suggests that the solution to what Conlisk calls $F(P)$, or what I've called $F(I(O))$, will be that point.

7 The Nature of Good Procedures

Since this is meant to be a theory of rational choice for real people, it would be helpful to say a few words about what these reliable procedures that stop the regress might be. In principle they could be anything, but in practice I think three kinds of procedures are particularly important: instincts, planning, and modelling. I'll say a bit about each of these in turn.

Humans are surprisingly good at instinctively allocating reasonable amounts of cognitive resources to computational tasks. In artificial intelligence research, one of the big challenges is trying to make machines be as good as humans at figuring out which problems to allocate cognitive resources to. This is sometimes known as the frame problem. Here's a typical description of this from a recent survey article.

> And, more generally, how do we account for our apparent ability to make decisions on the basis only of what is relevant to an ongoing situation without having explicitly to consider all that is not relevant? (Shanahan 2016)

Note that this assumes is that humans are actually very good at this rather hard task - setting aside the irrelevant without first thinking that it is irrelevant. This has to be instinctive. We don't go around thinking about how much time to spend thinking on various subjects. That would be self-defeating. Obviously we are far from perfect at this, but it is striking how good we are at it.

Recent work on 'vigilance' has illustrated how good we are at one aspect of this problem (Sperber et al. 2010). Somehow, and I don't think it is clear how, we manage to keep track of our environment in a comprehensive enough away that it allows us to focus on those things that need focusing on. For example, when walking down a busy street, we don't make a model of the expected movements of each of the individuals around us. That would be too computationally taxing. But we do pay enough attention to each of those individuals for us to be able to focus on any one of them if they

seem to pose a particular challenge or threat. If one of them is weaving in a drunken manner, or carrying a sword, we are able to focus on them very quickly. To do this we must be paying at least background attention to every one of them. I think this turns out to be a common phenomenon. There are many situations where we don't have the ability to carefully consider everything that's going on, but we do manage to pick out the things around us that need close attention. And that requires monitoring of the entire environment, and doing some very quick and dirty processing of the resulting inputs. As I said, it's a bit of a mystery how we do this. But whatever we do, it's an amazing feat of insinctively solving a cognitive resource allocation problem.

When I say we do some of these things instinctively, I don't mean that our ability to do them is innate. We might pick them up by learning from those around us. This learning need not be conscious. It might happen by imitation. It is sometimes thought that humans' disposition to over imitate those around them is a kind of irrationality (Levy and Alfano 2020). But my guess is that it is part of what grounds our skill in solving these hard cognitive resource allocation problems.

But rather than speculate further about what future research will show about the range and limits of human instinct, let's turn to two ways of consciously adopting reliable procedures. In his discussion of the regress, Miles Kimball (2015) suggests a few options that might work. I want to focus on two of them: planning and modelling.

> Least transgressive are models in which an agent sits down once in a long while to think very carefully about how carefully to think about decisions of a frequently encountered type. For example, it is not impossible that someone might spend one afternoon considering how much time to spend on each of many grocery-shopping trips in comparison shopping. In this type of modelling, the infrequent computations of how carefully to think about repeated types of decisions could be approximated as if there were no computational cost, even though the context of the problem implies that those computational costs are strictly positive. (Kimball 2015, 174)

And that's obviously relevant to David in the supermarket. He could, in principle, spend one Saturday afternoon thinking about how carefully to check each of the items in the supermarket before putting it in his shopping cart. And then in future trips, he could just carry out this plan. In general, planning as a device for incurring computational costs at a time when those costs are less costly.

This isn't a terrible strategy, but I suspect it's rarely optimal. For one thing, there are much better things to do with Saturday afternoons. For another, it suggests we are back in the business of equating solving $F(P)$ with approximately solving P. And that's a mistake. Better to just say that David is rational if he just does the things that he would do were he to waste a Saturday afternoon this way, and then plan it out. And that thought leads to Kimball's more radical suggestion for how to avoid the regress,

> [M]odelling economic actors as doing constrained optimization in relation to a simpler economic model than the model treated as true in the

> analysis. This simpler economic model treated as true by the agent can be called a "folk theory" (Kimball 2015, 175)

It's this last idea I plan to explore in more detail. (It has some similarities to the discussion of small worlds in Joyce (1999) 70-77.) The short version is that David can, and should, have a little toy model of the supermarket in his head, and should optimize relative to that model. The model will be false, and David will know it is false. And that won't matter, as long as David treats the model the right way.

There are a lot of things that could have gone wrong with a can of chickpeas. They could have gone bad inside the can. They could have been contaminated, either deliberately or through carelessness. They could have been sitting around so long they have expired. All these things are, at least logically, possible.

But these possibilities, while serious, have two quite distinctive features. One is that they are very rare. In some cases they may have never happened. (I've never heard of someone deliberately contaminating canned chickpeas, though other grocery products like strawberries have been contamination targets.) The other is that there are few easy ways to tell whether they are actualised. You can scan each of the cans for an expiry date, but it is really uncommon that this is relevant, and it takes work since the expiry dates are normally written in such small type. If a can is really badly dented, I guess that weakens the metal and raises ever so slightly the prospect of unintentional contamination. But it's common to have shelves full of cans that have no dents, or at most very minor ones.

Given these two facts - the rarity of the problems and the difficulty in getting evidence that significantly shifts the probability that this is one of the (rare) problems - the rational thing to do is choose in a way that is insensitive to whether those problems are actualised. Or, perhaps more cautiously, one should be vigilant, in the sense of Sperber et al. (2010), to some of these problems, and ignore the rest. But being vigilant about a problem means, I take it, being willing to consider it if and only if you get evidence that it is worth considering. In the short run, you still ignore the potential problem.

And to ignore a potential problem is to choose in a way that is insensitive to evidence for the problem. That makes sense for both the banknotes and the chickpeas, because engaging in a choice procedure that is sensitive to the probability of the problem will, in the long run, make you worse off.

In Kimball's terms, the rational shopper will have a toy model of the supermarket in which all cans of chickpeas that aren't obviously damaged are safe to eat. This will be a defeasible model, but on a typical grocery trip, it won't be defeated. In Joyce's terms, the small worlds the shopper uses in setting up the decision problem they face will all be ones in which the chickpeas are safe.

So the suggestion is that very often, the way to be rational is to have right model in your head, and apply it correctly. A choice is the rational choice in your situation iff it is the recommendation of the right model. And the right model includes just as much information, and just as many complications, as the situation demands. The regress is blocked, on this picture, because you don't have to have computed, or even be in a position to compute, that the right model is the right model. Here I am following Knight. Rational agents don't have to have worked through Marshall's Principles; they

just have to think and act as if they had. But crucially, they don't have to even act as if they are applying the Principles to the world. They could apply them to a good model of the world, and that's good enough.

8 Three Philosophical Postscripts

8.1 Idealisation

The story I'm telling here about how rational agents use models is very similar, and indeed draws heavily on, the story that Michael Strevens (2008) tells about how scientists use idealisations. On that story, to use an idealisation is to set some messy value to a computationally more simple value (often 0 or 1), and to (implicitly) assert that the difference between the actual value and the computationally simpler value is irrelevant for current purposes.

One benefit Strevens gets from this is that he is spared saying that scientists use falsehoods in their reasoning. After all, it is often true that the difference between the messy value and the simple value is irrelevant for current purposes - and that's all that the scientist is committing to.

The same is true in this picture. Frankie Lee can't know that the banknotes are all equally likely to be genuine; because that's not strictly true. But he can know that the right model to use in his current situation is one that sets the probability of any note's genuineness to 1. That's both true - assuming that our picture of what makes a model right is one that takes deliberation costs seriously - and well supported by his evidence.

8.2 Epistemic Luck

On the story I'm telling, whether decision makers are rational or irrational will often be a matter of luck. This is as you should expect if rationality is a matter of successfully applying a skill. Most skills are not infallible. Being skilled at an activity means one usually succeeds, or at least one succeeds at a higher rate than is normal, but on any given day one could fail. The epistemic failures I call irrationality, even though the person in some sense does the same thing as they do in cases where they act rationally.

Here is one version of that. Recall the version of the Frankie Lee example where the country has just started modernising its financial system by introducing plastic banknotes. Frankie Lee knows that plastic banknotes are genuine - no one has figured out how to forge them yet. So if some of them are on offer, he should take one. But, let's imagine, he's temporarily forgotten this fact. So he takes one of the paper notes. This is irrational.

But it's also bad luck. It's not normally required that we scour our memories for any relevant information before making a decision like this. Normally, Frankie Lee could have put in this much cognitive effort, and ended up rational. But the world did not cooperate, and he ended up irrational.

I think any story that connects rationality to succeeding via skill will have the consequence that sometimes whether one is rational is in part a matter of luck. But the

possibility of epistemic luck shouldn't surprise us. Assume that what one should, rationally, do and believe is a function of what one knows. And assume that the right epistemic logic is weaker than S5. Then one won't always know what one knows. So one won't always know what one should do or believe. So if one believes what one should, or does what one should, this will be in some sense a matter of luck. And surely the right epistemic logic is weaker than S5. Even if you think the anti-luminosity arguments are bad, and the right epistemic logic is stronger than S4, you shouldn't think that people know what it is they don't know. (False beliefs, for example, are typically pieces of non-knowledge that are not known to be not knowledge.) So we shouldn't be surprised that there is epistemic luck.

8.3 Knowledge and Rational Choice

So my preferred picture of rational action in cases where there are deliberation costs is that the chooser has a model of the decision problem in their head, and they know it is a good model. That's a constraint on rationality, but it's also a constraint on knowledge. If the chooser knows that p, they can't be using a model where it might be that $\neg p$. That, I think, is the core way that practical factors encroach on knowledge - sometimes one is in a practical situation where the best model allows for the possibility of $\neg p$, and being in such a situation defeats any putative knowledge that p.

But I used to say something different about how practical factors affected knowledge. I used to say something like the following.

- One knows p only if the rational choice (or choices) conditional on p are the rational choice (or choices) unconditionally for any choice one is considering.
- The rational choice, either conditional or unconditional, is the one with the highest expected utility, or if there are ties, then all of them are rational choices.

And it turns out that combination of views is untenable. This was shown independently twice over, once by Alex Zweber (2016) and then, separately, by Charity Anderson and John Hawthorne -Anderson and Hawthorne (2019). They considered situations like the original Frankie Lee example, and noted that my view had the implausible consequence that Frankie Lee did not know, of each note, that it was genuine. After all, as it stands Frankie Lee should be indifferent between the notes, but conditional on one of the notes being genuine, he should prefer that one. And that's implausible. Both papers go on to note other implausibilities that purportedly follow, but already we should acknowledge this is a problem. (Whether my view was really committed to the other implausibilities is something I could argue about, but it doesn't matter because this is already a perfectly good counterexample.)

The solution, I now think, is to qualify the second bullet point above. What I should have said instead is

- The rational choice is the one with the highest expected utility on the model that the chooser is rationally using, or if there are ties, then all of them are rational choices.

And now the problem goes away. It is rational for Frankie Lee to use the model where all the notes are genuine - it isn't worth the cost of using a more complicated model. And on that model, conditionalising on the hypothesis that one of the notes is genuine doesn't change anything. So if Frankie Lee is using that model, he knows the notes are genuine. If he isn't using that model then he doesn't know the notes are genuine. But this isn't because of any pragmatic theory of knowledge - it's simply that to know *p* requires one actually take *p* as given, and Frankie Lee fails that criteria.

So cases like Frankie Lee, or David and the chickpeas, are perfectly good counterexamples to the version of epistemic pragmatic encroachment I used to endorse. But they don't show that pragmatic theories are false in general; they just show I got an important detail wrong. To get these details right we need a better theory of when people (rationally) ignore the details.

References

Anderson, Charity, and John Hawthorne. 2019. "Knowledge, Practical Adequacy, and Stakes." *Oxford Studies in Epistemology* 6: 234–57.

Beddor, Bob, and Carlotta Pavese. 2020. "Modal Virtue Epistemology." *Philosophy and Phenomenological Research* 101 (1): 61–79. doi: 10.1111/phpr.12562.

Caplin, Andrew, Mark Dean, and Daniel Martin. 2011. "Search and Satisficing." *American Economic Review* 101 (7): 2899–2922. doi: 10.1257/aer.101.7.2899.

Chakravarti, Ashok. 2017. "Imperfect Information and Opportunism." *Journal of Economic Issues* 51 (4): 1114–36. doi: 10.1080/00213624.2017.1391594.

Chernev, Alexander, Ulf Böckenholt, and Joseph Goodman. 2015. "Choice Overload: A Conceptual Review and Meta-Analysis." *Journal of Consumer Psychology* 25 (2): 333–58. doi: 10.1016/j.jcps.2014.08.002.

Cherniak, Christopher. 1986. *Minimal Rationality*. Cambridge, MA: MIT Press.

Christensen, David. 2007. "Does Murphy's Law Apply in Epistemology? Self-Doubt and Rational Ideals." *Oxford Studies in Epistemology* 2: 3–31.

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

Diab, Dalia L., Michael A. Gillespie, and Scott Highhouse. 2008. "Are Maximizers Really Unhappy? The Measurement of Maximizing Tendency." *Judgment and Decision Making* 3 (5): 364–70.

Dylan, Bob. 2016. *The Lyrics: 1961-2012*. New York: Simon & Schuster.

Elster, Jon. 1979. *Ulysses and the Sirens: Studies in Rationality and Irrationality*. Cambridge: Cambridge University Press.

Gigerenzer, Gerd, and Reinhard Selten. 2001. *Bounded Rationality: The Adaptive Toolbox*. Cambridge, MA: MIT Press.

Harman, Gilbert. 1973. *Thought*. Princeton: Princeton University Press.

Iyengar, Sheena S., Rachael E. Wells, and Barry Schwartz. 2006. "Doing Better but Feeling Worse: Looking for the 'Best' Job Undermines Satisfaction." *Psychological Science* 17 (2): 143–50. doi: 10.1111/j.1467-9280.2006.01677.x.

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

Kimball, Miles. 2015. "Cognitive Economics." *The Japanese Economic Review* 66 (2): 167–81. doi: 10.1111/jere.12070.

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

Levy, Neil, and Mark Alfano. 2020. "Knowledge from Vice: Deeply Social Epistemology." *Mind* 129 (515): 887–915. doi: 10.1093/mind/fzz017.

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

Mangan, Jean, Amanda Hughes, and Kim Slack. 2010. "Student Finance, Information and Decision Making." *Higher Education* 60 (5): 459–72. doi: 10.1007/s10734-010-9309-7.

Manski, Charles F. 2017. "Optimize, Satisfice, or Choose Without Deliberation? A Simple Minimax-Regret Assessment." *Theory and Decision* 83 (2): 155–73. doi: 10.1007/s11238-017-9592-1.

Newman, David B., Joanna Schug, Masaki Yuki, Junko Yamada, and John B. Nezlek. 2018. "The Negative Consequences of Maximizing in Friendship Selection." *Journal of Personality and Social Psychology* 114 (5): 804–24. doi: 10.1037/pspp0000141.

Odell, John S. 2002. "Bounded Rationality and World Political Economy." In *Governing the World's Money*, edited by David M. Andrews, C. Randall Henning, and Louis W. Pauly, 168–93. Ithaca: Cornell University Press.

Ogaki, Masao, and Saori C. Tanaka. 2017. *Behavioral Economics: Toward a New Economics by Integration with Traditional Economics*. Singapore: Springer.

Papi, Mario. 2013. "Satisficing and Maximizing Consumers in a Monopolistic Screening Model." *Mathematical Social Sciences* 66 (3): 385–89. doi: 10.1016/j.mathsocsci.2013.08.005.

Pingle, Mark. 2006. "Deliberation Cost as a Foundation for Behavioral Economics." In *In Handbook of Contemporary Behavioral Economics: Foundations and Developments*, edited by Morris Altman, 340–55. New York: Routledge.

Reutskaja, Elena, Rosemarie Nagel, Colin F. Camerer, and Antonio Rangel. 2011. "Search Dynamics in Consumer Choice Under Time Pressure: An Eye-Tracking Study." *American Economic Review* 101 (2): 900–926. doi: 10.1257/aer.101.2.900.

Russell, Stuart J. 1997. "Rationality and Intelligence." *Artificial Intelligence* 94 (1-2): 57–77. doi: 10.1016/S0004-3702(97)00026-X.

Scheibehenne, Benjamin, Rainer Greifeneder, and Peter M. Todd. 2010. "Can There Ever Be Too Many Options? A Meta-Analytic Review of Choice Overload." *Journal of Consumer Research* 37 (3): 409–25. doi: 10.1086/651235.

Schwartz, Barry. 2004. *The Paradox of Choice: Why More Is Less*. New York: Harper Collins.

Schwartz, Barry, Andrew Ward, John Monterosso, Sonja Lyubomirsky, Katherine

White, and Darrin R. Lehman. 2002. "Maximizing Versus Satisficing: Happiness Is a Matter of Choice." *Journal of Personality and Social Psychology* 83 (5): 1178–97. doi: 10.1037/0022-3514.83.5.1178.

Scott, Rose M., and Renée Baillargeon. 2013. "Do Infants Really Expect Agents to Act Efficiently? A Critical Test of the Rationality Principle." *Pscyhological Science* 24 (4): 466–74. doi: 10.1177/0956797612457395.

Shanahan, Murray. 2016. "The Frame Problem." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Spring 2016. Metaphysics Research Lab, Stanford University.

Sperber, Dan, Fabrice Clément, Christophe Heintz, Olivier Mascaro, Hugo Mercier, Gloria Origgi, and Deirdre Wilson. 2010. "Epistemic Vigilance." *Mind and Language* 25 (4): 359–93. doi: 10.1111/j.1468-0017.2010.01394.x.

Strevens, Michael. 2008. *Depth: An Account of Scientific Explanations*. Cambridge, MA: Harvard University Press.

Tucker, Chris. 2016. "Satisficing and Motivated Submaximization (in the Philosophy of Religion)." *Philosophy and Phenomenological Research* 93 (1): 127–43. doi: 10.1111/phpr.12191.

Williamson, Timothy. 2000. *Knowledge and its Limits*. Oxford University Press.

Wittgenstein, Ludwig. 1953. *Philosophical Investigations*. London: Macmillan.

Zweber, Adam. 2016. "Fallibilism, Closure, and Pragmatic Encroachment." *Philosophical Studies* 173 (10): 2745–57. doi: 10.1007/s11098-016-0631-5.

Unpublished. First posted in 2020.