The Parts of an Imperfect Agent[1]

Sara Aronowitz

Abstract

Formal representations drawn from rational choice theory have been used in a variety of ways to fruitfully model the way in which actual agents are approximately rational. This analysis requires bridging between ideal normative theory, in which the mechanisms, representations, and other such internal parts are in an important sense interchangeable, and descriptive psychological theory, in which understanding the internal workings of the agent is often the main goal of the entire inquiry. In this paper, I raise a problem brought on by this gap: for almost every theory of approximate rationality, there will be an empirically indistinguishable alternative that individuates the parts of the agent in a significantly different way.

Introduction

Understanding when we meet and fail to meet the standards of rationality is an important part of understanding human cognition. To this end, formal theories of ideal rationality — such as decision theory — hold promise as ways of making precise what these standards might be. These models have, of course, a wide domain of application in laying a standard of rationality for communities, artificial agents, and the process of science. But our use of such idealized accounts often goes beyond creating a measuring-stick for human progress. We sometimes take these models to provide insight into the representations and computations with which we produce beliefs and behaviors. In doing so, I'll argue, we sometimes misstep in an interesting way: taking the ideal model too literally, we rely on it to adjudicate fine-grained debates about inner workings. Revealing these kinds of mistakes opens up a gap. As we move away from perfectly rational agents, questions emerge that, on the one hand, cannot be settled by the ideal model, and on the other hand, have deep implications for whether and how imperfect agents might be rational.

A wide variety of recent work, such as books by Julia Staffel (2020) and Richard Bradley (2017), has focused on how to relax the standards of ideal rationality to *evaluate* less perfect agents. In this paper, I'll instead take a step back. Before we can face these problems of evaluation, we need to first apply concepts such as credences, preferences, and inferences.

[1] This is a penultimate draft of a paper forthcoming in *Oxford Studies in Philosophy of Mind*.

Already, as I'll argue, difficulties arise in assessing which part, if any, of an imperfect agent should be brought under these labels. This question starts with one way in which philosophy of mind has relied on philosophy of science: to lend idealized formal models. We end up with a lesson that philosophy of mind might in turn share: to attend closely to the connection between standards of evaluation and the nature of the states which are being evaluated.

In Section 1, I introduce and motivate the project of "approximate rationality", a mix of normative and descriptive inquiry which will be the target of this paper. Section 2 presents two approximate rationality models of decision-making that use the same formal machinery and yet make substantially different commitments about inner workings, and consequently rationality. Could this problem be solved by picking the model closest to the ideal? I argue that this impasse is genuinely hard to solve, even in principle, in Section 3. Section 4 aims to extend this impasse to related debates.

1 Approximate rationality: what and why

The project I'll call "approximate rationality"[2] is an attempt to combine two forms of theorizing: descriptive psychological work on how we function, and normative work on what rational functioning amounts to. In combining these two, the approximate rationality project does not merely aim to conduct two lines of inquiry simultaneously, but to allow each to inform the other in creating a theory that is in some sense a united whole. Within this umbrella, we find several significantly different forms of combination.

One version (arguably what was suggested by John Anderson (1990) under the heading of 'rational analysis') might start from the empirical observation that humans are surprisingly successful in various domains, and then ask how this might be possible given psychological constraints and principles of rationality. This project may ask about success across any environment consistent with our evidence, or instead understand our success as built upon the particular environments we tend to encounter (the latter being a version of Herbert Simon's ecological rationality (1955)). We might instead accept that humans are not on the whole rational

[2] I use this term instead of ones like "bounded rationality" (Simon, 1990; Gigerenzer & Selton, 2002) or "rational analysis" (Anderson, 1990; Lewis, Howes & Singh, 2013) in order to signal that I mean a very broad set of approaches, including those that focus on explaining how we fail to be fully rational.

or irrational, but place greater scientific value on explanations of how we are rational, when we are, than explanations of irrationality or a-rationality.

Alternately, we might use descriptive psychological findings to provide insight into a question about rationality, such as: how is it possible to learn new theories selectively, that is without brute-force enumerative search? Along these lines, Dedre Gentner has built a theory of the rationality of analogical reasoning based on historical (Gentner et al 1997) and lab-based (e.g Markman & Gentner 1994; Gentner 2010) observations about how humans come up with new ideas. For instance, she uses the example of Kepler's development of an analogy between light and "the motive power" to refine and illustrate a theory of conceptual change where we map structures from one domain onto a novel hypothesis in a target domain, and then adjust the comparison along with our understanding of both domains. Here, findings about how we think emerge alongside analysis of the rationality of discovery of new concepts, both supporting one another: the more it appears a way of thinking is characteristic of actual human discoveries, the more it would seem to be a candidate for a rational mode of discovery, and vice versa. Her resulting theory, the structure-mapping theory, is a theory of how we solve a problem (inventing a new theory) that the most of ideal of agents never need to solve; with an unlimited ability to construct possible hypotheses, it is no longer necessary to be economical in adding just the new theories that can constructed most handily out of current ones through analogy and other forms of bootstrapping.

Gentner's work tells us a lot about human reasoning, of course, but it also illuminates something about the problem of new hypotheses itself, and consequently has been extended to artificial contexts where conformity to actual human reasoning is not a major aim (Falkenhainer et al 1989). That is, both Anderson's approach and Gentner's aim to understand approximate rationality through psychology as well as psychology through approximate rationality, though I separate out these two directions to make the conceptual point that each direction is a distinguishable form of inquiry.

There are many approximate rationality projects beyond the ones I've sketched here. But for the purposes of this paper, these differences are for the most part irrelevant. All of these projects aim to carry out a kind of combination and integration of two very different forms of theorizing, and the challenge I'll raise picks up on that feature. Before presenting the challenge, however, it's worth saying something about why this form of theorizing is attractive and important.

First, imagine that we can derive a theory of ideal rationality on purely a priori grounds. Even so, this may not be enough to determine a hierarchy of approximations without descriptive input. This could simply be because of our limited scientific imagination. But in some cases, determining closeness to the ideal is underdetermined without first specifying at least some features of the environment, such as an ecological hierarchy that evaluates success in the actual world and its near modal neighbors as closer to rationality than success in an equally sized range of distantly possible environments. In either case, without such a hierarchy, and provided that humans are never fully ideally rational, we would be unable to ever classify our behavior as more or less rational. Presumably this would fall far short of a central explanatory aim of both philosophy and psychology.

Second, we may not even be able to understand ideal rationality without leaning on descriptive findings. This might be a contingent fact about human inquiry, or a deeper fact about conceptual priority. In the first case, looking at descriptive data on human successes might be important in getting inspiration for even purely normative inquiry — after all, we might have left out key possibilities for rational optimization that might only emerge when presented through actual behavior. In the second case, some meta-epistemological views support the dependence of the ideal on the non-ideal, such that what defines the ideal thinker is dependent on empirical facts (Kornblith 2002). On this view, the normative theory is an idealization that might differ depending on the empirical starting point and be underdetermined in lieu of a connection to empirical inquiry.

Finally, purely descriptive psychology may likewise be hard to understand on its own, without drawing to some degree on theories of rationality. Following Gibbard (2012), we might take the idea of meaning, including the cognitive significance we appeal to in psychology, as essentially normative. Or drawing from Dennett (1988) or Davidson (1980), it may be that minds, intentions, beliefs and so on only emerge once we're in the business of interpreting behavior as oriented toward goals. On these projects, we don't just fit any model of beliefs and desires to an agent's actions, but our understanding is constrained by a (defeasible) preference for models that make *sense* of the agent's behavior by aligning actions with mental states in a normative way. That is, we might be able to imagine a world where the only way we study humans is using the same descriptive orientation we currently take towards studying volcanoes

— but in that world, there might be no psychology nor any science of the mind. At a minimum, it would be a very different kind of science.

Approximate rationality comes in a variety of different forms. In all cases, it is committed to a genuine synthesis of normative and descriptive inputs to produce a theory that explores rationality in realistic agents. While I will now raise a challenge for this synthesis, it's important to note that approximate rationality is well-motivated on both psychological and philosophical grounds and, far from being an optional add-on, may even be essential for "pure" normative and descriptive projects.

2 Two Competing Approximate Rationality Models

2.1 The Structural Model

We'll start with an example: I've modified two approximate rationality models in the literature to be mathematically identical. However, they disagree about how parts of their approximate agent map on to the ideal decision-theoretic agent, and thereby have a substantial disagreement about in what sense the approximate agent is rational. The first of these, which I'll call the Structural Model, turns up in many recent projects in various forms, but I'll present it through the account given by Howes et al (2016). Like many approximate rationality proponents in the Rational Analysis tradition, the authors present their framework as making rational sense of a common behavior previously understood as irrational. In this case, preference reversals. To illustrate this phenomenon, let's take an example. Suppose you are going to a restaurant which has a menu of two items, Xi'an noodles and eggplant. You find yourself more or less indifferent between these options, but on reflection, the eggplant sounds a little more tasty. Before you go to order, however, you realize you were mistaken. The menu actually contains three options, Xi'an noodles, Lanzhou noodles, and eggplant. You've always thought Xi'an noodles were tastier than Lanzhou — they have more interesting spices. Considering these three options, you find yourself drifting towards the Xi'an noodles.

This shift in choices should not seem outlandish: it's been well documented in various forms of behavioral experiments. In particular, it's not just that people sometimes change their minds when presented with a third (inferior) option. The key feature of this phenomenon is that the third, unchosen, option is a "decoy" (also called phantom or phantom decoy) — it's somehow

similar to but worse than one of the original options, and the presentation of the decoy tends to shift preference towards that similar, but better, option.

This form of contextual preference reversal is on its face irrational. The option of Lanzhou noodles, given that you're not going to choose them, should not change your preference among the other two options. Unless we get very creative with the way to understand the choice problem, an agent who exhibits preference reversals will violate the Independence of Irrelevant Alternatives (IIA). I'll define IIA as follows, where "$x > y$" stands for a strict preference for x over y:

IIA If A and B are two options such that A>B in choice set {A, B}, then it cannot be the case that B>A in any larger set that contains A and B.[3]

IIA can be thought of intuitively as guaranteeing that no independent third option can reverse the relationship between A and B. It's also worth noting that if we treat the value of every option as fixed and absolute (and apply a suitably straightforward decision-making rule), IIA follows automatically. Conversely, violations of IIA force us away from frameworks where we represent your desire for each option as a fixed value, and your behavior as flowing consistently from those values. Note that IIA itself, as I've defined it, is a constraint on preferences. If we allow our agent to act in a way that doesn't reflect her preferences, then a preference reversal does not necessarily imply a violation of IIA. So while preference reversals are prima facie evidence of IIA violations, they do not necessarily imply such a violation.

Howes et al. develop a psychological theory that rationalizes preference reversals, showing how they are consistent with IIA. Their model takes a standard decision-theoretic idealized agent and breaks the informational connection between her decision algorithm and her underlying preferences and utilities. This approximately rational agent possesses a classical, well-behaved utility distribution that respects IIA, but she only has partial access to this distribution through a noisy sampling process. The key here is that since she's uncertain about her own preferences,

[3] As has become standard outside of voting theory, this property is not actually Arrow's (1951) original IIA, but Sen's (1970) principles α and β (see Eels & Harper, 1991). For present purposes, the difference between these principles is not important, since in any case, the principle is incompatible with a standard reading of preference reversals where the addition of the decoy causes the person to prefer the initially dispreferred option. For that reason, I have formulated a weak version of IIA such that violations of this principle will necessarily be violations of the stronger versions.

features of the decision problem itself can give her information pertinent to estimating those preferences. This generates a sensitivity to context in decision behavior without sensitivity to context in the preferences or credences themselves. And so in the case of preference reversals, the rough idea is that the presentation of the Lanzhou noodles, and your subsequent feeling that the Xi'an noodles were similar but much better, gives you a bit of information about the absolute value of the Xi'an noodles. It suggests they must be pretty good, if a (randomly-generated) alternative is definitely worse. And thus the presentation of the decoy shifts your choice behavior, not because you've changed what you value, but because you've gotten a hint about what you value.

Howes et al. provide a model that is intended to capture human behavior, as well as show that the occurrence of preference reversals in certain contexts is rational (figure 1). The details of their model are mostly not pertinent to the aim of this paper. However, several features will be relevant: first, their model uses a sampling process to determine its own utilities and probabilities. Second, the underlying distributions from which the samples are taken are identified with the agent's actual utilities, probabilities, and preferences. Finally, their model actually samples both an expected utility value, and ordinal rankings over both utilities and probabilities.

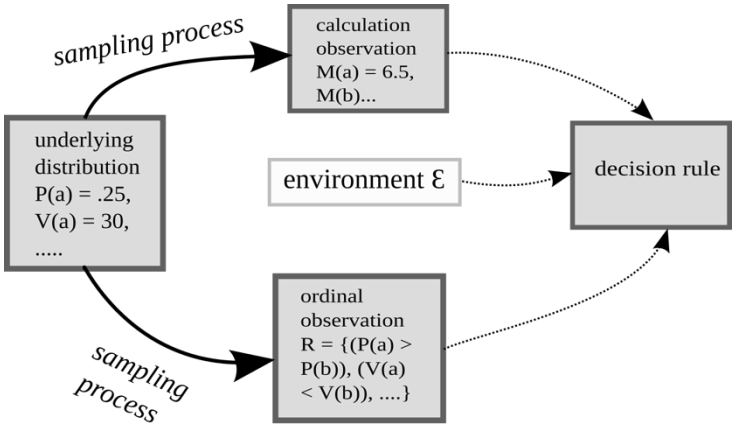

Figure 1: The model agent developed by Howes et al. Their model makes two kinds of observations of its own underlying state (left). The calculation observation is a noisy sample of the expected utility of an act, and the ordinal observation is a noisy sample of the value ordering over outcomes in the decision set, along with a second ordering representing the relative probability of these outcomes. The agent uses these two observations as inputs to a decision rule along with an expectation of the environment E. (Figure based on the original paper.)

The ordinal observation is crucial for generating preference reversals, as can be seen in the informal gloss given in the preceding paragraph.

Howes et al. take their model, if accurate, to vindicate human rationality: "In summary, our analysis of the effect of the phantom decoy suggests that the average behavior of the participants studied by Soltani et al. (2012) was rational; by making ordinal observations relative to an unavailable option suppresses the selection of options dominated by a phantom, these participants were behaving in a way that is consistent with an observer that seeks to maximize the expected value of selected gambles given noise." Here, they take the way in which their agent is rational to be that she is doing what even the best possible observer would suggest given her *evidential limitations*. But these are not standard evidential limitations. They are limits in knowing one's own values and beliefs — internal informational barriers, rather than limited access to underlying facts in the environment. In fact, this is one of two ways their agent might be deemed rational. She's also rational insofar as her preferences satisfy IIA.[4]

To generalize, structural models preserve rationality in the face of limitations by inserting informational barriers inside the agent. These barriers allow the structural agent to preserve two forms of rationality: first, she is what I'll call intra-internally rational. There's a sort of built-in mini agent, in our case the decision-maker, who is classically rational in the internalist sense, but merely lacks access to pertinent information on the other side of the barrier. Second, the structural agent is rational in the sense that her buried representations, on the far side of the barrier, may themselves be well-behaved, coherent, or otherwise classical. In this case, the underlying preferences satisfy IIA, and in fact presumably satisfy other constraints. On the other hand, there are ways in which the structural agent fails to be rational, which will be brought out via a contrast with our second model.

2.2 **The Dispositional Model**

I'll now discuss an account which is in some sense a rival to the Structural Model. This account is essentially the one presented by Icard (2016) — however, for dialectical purposes, I'll present a version of Icard's picture that adheres as closely as possible to the details of the Howes et al. model. In fact, this involves making just one simple change. Instead of thinking of the utilities

[4] see Rulli & Worsnip (2016) for a discussion of the place of IIA in rationality

and probabilities of the agent as the underlying distribution from which samples are drawn, as Howes et al. do, this dispositional model takes the probabilities and utilities of the agent to be her dispositions to sample from those distributions. In the previous section, our agent really took the value of choosing the Xi'an noodle to be always greater than choosing the eggplant, but her behavior varied due to her lack of consistent access to those preferences. On the dispositional model, we would treat an agent expressing these preferences as having a more complex preference ranking. Icard argues for an understanding of credences as dispositions to sample, and so our modified understanding will be as of preferences as dispositions to sample.

First, a technical qualification. Because the Howes et al. model involved a few different noisy observations, we could in principle discuss a few different forms of the dispositional model. That is, using exactly the same model, we have three sampling processes, and so we can talk about a disposition to sample from the expected utility calculation, the probability ranking, and the value ranking. However, since the vindication relies on the ordinal rankings, and in fact Howes et al. demonstrate that many other agents who use noisy ordinal sampling will exhibit the same behavior, it seems most pertinent to focus on the ordinal observations. And since the preference reversal decision problems do not involve much or any uncertainty about which outcome will occur given which act is chosen, it's simpler to just focus on the ordinal observation of utilities as an expression of preference.

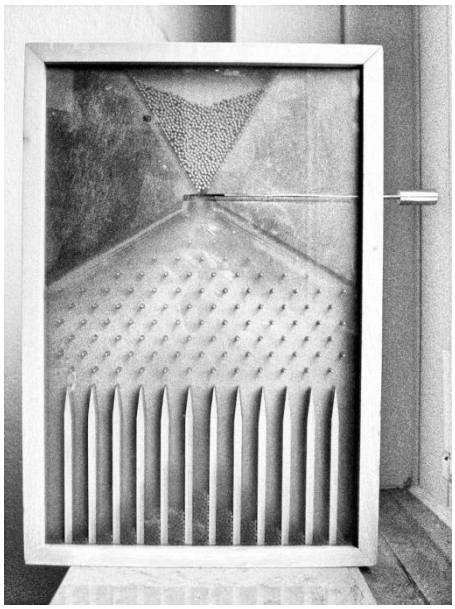

Figure 2: In the Galton box, balls are dropped from the top, and each ball, once it comes to rest in the slots at the bottom, acts as a sample from an approximately Gaussian distribution. Image source: https://en.wikipedia.org/wiki/Bean_machine

Icard's idea, in brief, is that treating credences as these dispositions sidesteps a lot of difficult issues about how credence distributions, which are massively complex, could be represented in the brain, as well as reflected in our sometimes quite inconsistent behaviors. It might initially appear odd to talk about a disposition to sample without a representation of the underlying distribution from which samples are taken. But Icard brings out that this oddness relies on an incorrect understanding of how sampling works. Not only do many sampling algorithms operate without such an explicit representation, but this even applies to simple sampling machines. In his example, the Galton box (figure 2) is a device where balls are dropped over a set of evenly-distributed pegs through an opening which tapers outwards. Balls dropped through the Galton box land in the bottom in a pattern which approximates a Gaussian curve. Since this is a random process (more or less), each dropped ball is a sample from the Gaussian distribution. But of course nowhere is there a representation or explicit encoding of that distribution, there is just a bunch of pegs spaced evenly on a board. The lesson of this example is that there are cases where the sample is the 'real' thing, and the underlying distribution is the abstraction.

Our version of Icard's picture, then, takes exactly the same formal machinery as the structural model of Howes et al. But instead of identifying preference with the underlying, noisily observed, ordering, we identify the preference with the disposition or propensity to sample from that ordering. This change has a few significant ramifications.

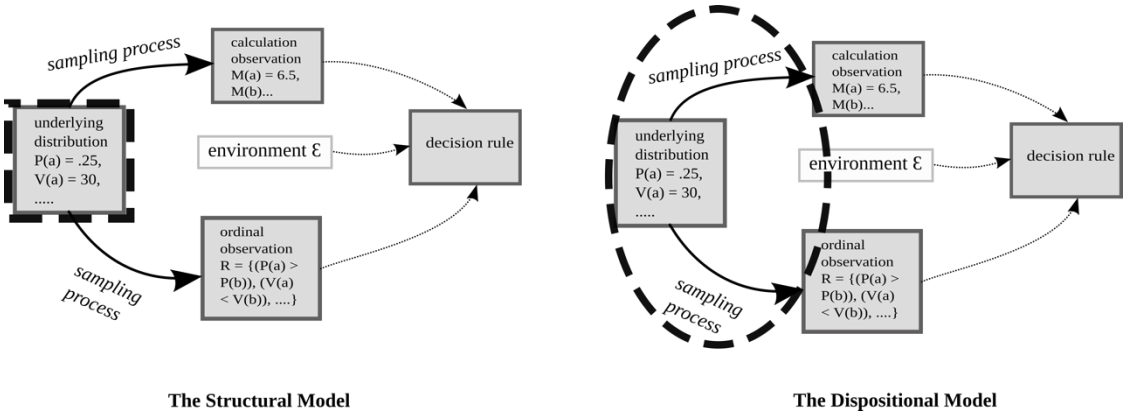

Figure 3: The structural and dispositional models. Heavy dashed lines locate the representation of credences and utilities. On the structural model, these are identical to the underlying distribution. In the dispositional model, utilities and credences are instead represented by a disposition to sample.

The structural model satisfied IIA, and the dispositional model does not. It might be more accurate to say that the dispositional model is not evaluable with respect to IIA, because treating the preferences as a disposition rather than an ordering means it is something of a category mistake to apply the criterion. But this is not merely a syntactic problem. Consider the fact that many sampling methods are predictably biased. An agent's disposition, and hence her actual preferences, will take into account her method of sampling. So assuming a relatively small number of samples, an agent with a propensity to use biased sampling methods will have preferences that are not particularly well described by the underlying distribution. In fact, this is the move Icard relies on to accommodate inconsistent behaviors. So even if we are allowed to find a unique ordering that reflects the disposition, that ordering would not be identical with the ordering given by the underlying distribution. Therefore, the dispositional agent does not typically meet IIA, even on a very liberal interpretation.

The structural model was intra-internally rational, in that the agent was doing its best given informational limitations. The dispositional model doesn't make use of a barrier beyond which the agent cannot "see". Instead, the underlying distribution is conceived of as implicit. So while it was true for the structural agent that some observer could do no better than the agent herself when it came to making decisions based on available information, this will not be true of the dispositional agent. Presumably, some ideal observer could extract the underlying distribution, given that it is implicit. Then, this observer could in some sense do better than the dispositional agent herself. In what sense could the observer do better? She could produce a consistent set of responses that reflect the performance of the dispositional agent if she was able to take *infinitely many samples*. There are some thorny issues about whether this counts as doing better by the agent's own lights, but I will tentatively suggest that it does — at least internal to the notion of sampling, if more samples in the limit does not result in "better" behavior, something seems to have gone very wrong.

So the dispositional agent seems less rational than the structural agent in terms of consistency constraints like IIA and intra-internal rationality. But the dispositional agent has a form of rationality that the structural agent lacks. The dispositional agent always acts (or at least chooses) in a way that reflects her preferences. This is clearly false of the structural agent: the barrier trick which allowed her to preserve rationality of the other sort is her undoing here.

3. The Individuation Problem

This brings us to the problem. We have two agents, who have all of the same formal machinery. They input the same information and output the same choices according to the same mathematical operations. But we have moved the labels of 'utilities', 'credences', and 'preferences', and each agent comes along with an interpretation of what is accessible and inaccessible (this is in effect just another label, the 'barrier' that distinguishes the inner agent from the outer agent). This seemingly trivial adjustment had significant consequences for how we understand the rationality of the two agents. The classical ideal agent satisfies IIA, she is intra-internally rational, and her choices always directly reflect her preferences. The structural agent is only rational in the first two ways, and the dispositional agent only in the last way.

The problem is this. Approximate rationality accounts must answer questions like: in what way are people rational? Otherwise, they do not make good on their promises to bridge between ideal and descriptive. But in this example, we've seen how a pair of such accounts differ dramatically in how they would answer this question. So there must be a way to determine which account is correct, at least in principle. Otherwise, the question cannot be answered.

However, I'll now argue that there are serious difficulties in even imagining how we could decide between the structural and dispositional accounts in this case. In the next section, I show that this problem is not unique to this context, but generalizes within approximate rationality projects and perhaps to other domains which aim to bridge the ideal and descriptive.

I will not offer an exhaustive and conclusive argument to the effect that adjudicating between the structural and dispositional accounts is impossible: indeed, I hope that it is possible. But I will consider three avenues of evidence, and suggest some difficulties with each: normative theorizing, descriptive evidence, and theoretical virtues.

3.1 Drawing on the purely normative

Can we appeal to considerations in the ideal version of decision theory to decide between the two? One obvious way to do so would be if decision theory itself had an opinion about what kind of object preferences (or credences) are, such that we could compare the ideal mental objects to the non-ideal, even absent differences in behavior.

I'll now argue that ideal agents as understood in decision theory, do not have inner workings. My argument will have two parts. First, the considerations that determine the features of an ideal agent are not such that, even in principle, they could have anything substantial to say about one

set of inner workings over another. Second, since an ideal agent is merely an idealization and not an actual creature, the agent doesn't strictly speaking possess any more precise characteristics than can be specified by their theoretical role.

What do I mean by "inner workings"? The difference in the two models in section 2 highlights a disagreement between understanding the sampling process as part of the agent's preferences or merely a reflection of deeper preferences – a disagreement about where the preferences are in the model, rather than anything about the inputs or outputs of the agent (since these, in our example, are completely identical). More generally, descriptions of inner workings answer questions such as: how many parts does the agent have? Does she represent possibilities by numbers, symbols, or something else? Does she have a single representation of utilities, or more than one? By saying that ideal decision-theoretic agents have no inner workings, I mean that such questions have no answers or are poorly posed with respect to these agents.

To start, it's worth noting that this lack of inner workings is implied by the way ideal agents are discussed generally, and particularly in early writing on decision theory. For example, Savage (1972) introduces the decision-theoretic agent as a means to determine "criteria for deciding among possible courses of action". That is, the discussion of the agent is merely in service of determining standards for rational action. Ramsey (1931) likewise contrasts a notion of beliefs as real mental states knowable through introspection with one on which beliefs are merely the causes of action, concluding "the kind of measurement of belief with which probability is concerned is not this kind [i.e. the introspectable kind] but is a measurement of belief qua basis of action"(171). Humans, then, presumably have some actual internal workings that correspond to causes of action.

But how to go from causes to reasons? In humans, these things can be far apart. For instance, Railton (2017) provides a complex set of criteria to separate "causal-explanatory reasons" (a subset of causes) from "putative-rationale reasons". Railton explains how these come apart when we act without engaging our capacity to act *for* a reason and are thereby merely caused to act, which sets up a puzzle about a naturalistic account of what that capacity might be. Railton's discussion is in contrast to our ideal agent. For instance, the ideal agent never chooses to gamble just because they are feeling stressed, or opens the fridge out of habit. This creature whose actions are always determined rationally has no further sense of internal workings as causal-explanatory reasons beyond whatever provides reasons to act, since there are no merely caused

actions for such an agent. The ideal decision-theoretic agent, thus, only acts when they have reason to, and only from that reason (as opposed to by accident, from pure habit, and so on) and thereby can't be said to have causal-explanatory reasons as distinct from putative-rationale reasons[5].

Of course, in the variety of literature on decision-theory, constructs such as preferences and credences are taken literally (that is, taken to designate inner workings) to various degrees. I am not aiming to argue that theorists never take the more literal stance. Instead, we should ask: do the reasons why we take decision-theoretic agents to be ideal have anything to do with inner workings? If the arguments for these agents as models of the ideal, or for certain properties being ascribed to the agents on ideal grounds do not depend on inner workings at all, we can then conclude that these agents *as ideals* do not have inner workings. That is, if they would do their job in the ideal theory just as well without any particular inner workings, we can consider the inner workings dispensable. Let's turn to some of these arguments.

First, Dutch book arguments are a family of arguments that aim to discredit a particular version of ideal agency. The targeted agents are shown to be potential victims of a series of bad bets, such that the agent herself will agree to the bets, and that once she takes these bets, she's guaranteed to lose money. While interesting differences exist between arguments in this family, they all implicitly target any internal operations that would lead to the bad bets.

But these arguments target the downstream (hypothetical) consequences of a decision-making algorithm: what about arguments that point to an intrinsic problem? One example of such an argument comes from Joyce (2005) who sets up an example of an urn which could spit out coins that have any possible bias, and you have no information about the distribution of such biases among the coins. Joyce suggests that there is something different about your evidential situation with respect to a coin from that mysterious urn as compared to an ordinary coin, and uses this intuition to support the claim that a good theory of credal states ought to treat the two situations as requiring different epistemic responses. Unlike the Dutch book argument, this argument does not target the distal consequences of internal states but instead places a constraint

[5] Of course, an ideal agent can be caused to behave, where behaving designates non-intentional activity, but when they act (where action is understood to require intention), their causal-explanatory reasons are inseparable from putative-rationale reasons.

on credal representations. So it should be a favorable case for someone who disagrees with my claim and holds that decision theory does actually concern itself with inner workings.

However, looking more closely, we see that what Joyce is targeting is the degree of information contained in precise credences. The problem with various precise ways of accommodating the urn example is that they represent the situation as if the agent had far more evidence than she actually had: even if the principle of indifference could give us the least informative precise credences to adopt, Joyce argues, such credences would still be far too informative. This critique, then, does not reach as far into the inner workings of the agent as it might have appeared. Informativeness of a mental representation, for an ideal agent, is *transparent* — her credences are informative with respect to distinguishing situation x from y if and only if they allow her (or someone with those credences) to in principle act differently in x than in y.

Of course, we often talk as if the inner workings of these agents are real. We say, for instance, decision-theoretic agents have two distributions, one of utilities and the other of probabilities. And especially when discussing rational credences, we abstract away from Ramsey's idea that beliefs (and degrees thereof) are just the things that make the right kind of causal contribution to action. But given that the kind of arguments we rely on for such theories are almost always indifferent to inner workings, it would be uncharitable to take this talk too literally. Instead, we should accept that the notions of utilities, probabilities, decision algorithms and so on are in the ideal case merely *functional* at the agent level: they individuate only as finely as a function from possible environmental inputs to possible behavioral outputs. This means that any finer-grained question about two kinds of inner working that in the ideal case would amount to the same function are not answerable.

None of this should be particularly surprising, given that, for instance, we also don't think there are real differences between Turing machines and Abacus machines — two idealizations that are capable of computing exactly the same set of functions. This analogy, however, also suggests a warning with respect to the point I've made in this section. Consider the comparison between standard, deterministic Turing machines and non-deterministic Turing machines. The latter can be simulated on the former, such that the set of computable functions will be the same for both. However, this deterministic simulation massively increases the number of steps that the machine would take to compute some functions relative to the non-deterministic machine.

Since these steps are really just the same steps, it seems reasonable to talk about this difference in terms of time, since any assignment of time to steps would result in the non-deterministic machine being faster. The moral of this comparison, I take it, is that even the slightest de-idealization (in this case, allowing in a very generic notion of time costs such that the same steps take the same amount of time) starts to make differences in inner workings consequential, and hence real.

Even if I've shown that ideal decision theory does not directly mandate inner workings, it might seem as though we could, for instance, measure the 'distance' between each of the two and the ideal version, and take a shorter distance to be a mark in favor. The issue with this idea is that the three elements of rationality are not obviously comparable. Is it more rational to have preferences that are consistent with each other, or to have choices that are consistent with one's preferences? These are both consistency requirements, of a sort, and it seems quite arbitrary to establish precise weights to accord them.

Further, even if such a comparison were possible, a second problem would follow: if there were pairs of rational approximation accounts, such that they differed significantly in their degree of rationality, would it really be acceptable to use this as a deciding factor or even a tiebreaker? This would entail a form of bias towards rationality that, depending on the degree of difference, could be quite substantial. There's a difficult line to walk here: the smaller the degree of difference, the less such a difference should be used to decide between theories, but the larger the degree of difference, the more problematic a rationality bias would be.

One might make the case that a rationality bias is relatively innocent. Along these lines, I once heard a proponent of this approach justify it by saying: "is a psychological theory even an explanation if it doesn't explain why a behavior is rational?". That is, in this domain, we might think a good explanation (often) just must be a rational explanation. In other domains, we find a preference for explanatory over non-explanatory accounts to be benign, and even optimal. So we might seek to assimilate a preference for theories on which subjects are more rational to this more general category of preference for explanatory theories. But scientific theories do not in general seek to determine *whether* things in nature should have an explanation, but *how* they can be explained. Were these theories to attempt to discover the boundaries of what should be explained, it would indeed be illegitimate to prefer explanatory theories. On the other hand, a common debate among approximate rationality accounts is which behaviors of ours are rational

and which are irrational. In the background lurks a more general disagreement about whether on the whole we usually meet the standards of rationality. So it seems like positioning rational explanations as especially explanatory does not justify a preference for rationality when engaged in at least one substantial debate within approximate rationality. Thus even if we could assess whether the structural or dispositional model is more rational, we would still face obstacles in using that difference in favor of one or the other theory.

4.2 Drawing on descriptive data

The use of descriptive data to adjudicate between models seems more promising. Could we not, for example, identify a neural system that resembles an underlying distribution, or one that is closer to a Galton box? After all, the two accounts differ in what they claim is explicitly represented, even though they use all the same numerical machinery.

I think there is something both right and wrong with this line of thought. The correct part is that the plausibility of the dispositional model rests on empirical findings: it cannot be that the underlying distribution, a set of well-behaved relations among outcomes individuated in some suitable way, is genuinely explicitly neurally represented, since if that were the case, there would be something overly complex about relying on the disposition to explain behavior over time when we could instead talk about this tidy neural representation.

However, the structural account need not be committed to any particular explicit representation. Why not? The idea that we would find a list of outcomes with an ordering relation somehow written in the brain takes decision theory far too literally. That is, we've already seen that ideal theory is not concerned with inner workings. This means that the particular way we describe utility distributions is one out of a vast set of internal workings that function the same. There is nothing special about the one we have chosen. From this we can conclude two things: first, it would be quite surprising if we found anything like the particular conventional representation in the brain, and second, since the criteria for identifying an 'explicit' representation of preference in the brain cannot be any finer in grain than the notion of what a preference is, there is really no support from ideal theory for thinking a so-called explicit representation is more psychologically real or even more explicit than a so-called "implicit" one.

Thus, there is no ready account of what kind of neural finding could favor one of these accounts over the other. The fundamental problem is that these accounts deal in concepts taken from the ideal theory, and yet accord them a kind of representational commitment that cannot

come along with the concept. This is not to say that these representational questions are vacuous. It's just that however much we learn about rational choice, and however much we learn about brains, we will not amass a theory of whether preferences, utilities or credences are represented in the brain, ceteris paribus. What we need is a theory of preference-representations, utility-representations, and so on that *add* to the ideal concepts further constraints on inner workings. Empirical evidence alone cannot do this.

4.3 Drawing on theoretical virtues

Finally, could theoretical virtues such as simplicity, depth, and so on close this gap? Let's consider an extreme form of the structural account. We'll assume there is so much variability in human behavior that the best structural explanation is that the barrier is very opaque, and the ability to sample both limited and inconsistent. On this model, we've put so much information beyond the barrier, and allowed so little of it to be accessed, that our inner agent has become quite "small": she has little information, and her influence on decision-making is dwarfed by that of the variability in the input she receives from across the barrier. Call this the tiny agent model. It strikes me that some of the theoretical virtues count against the tiny agent model. It seems less explanatory than the original structural model, less deep, and perhaps even less parsimonious.

Conversely, imagine a version of the dispositional model that is fit to human behavior that is very consistent, and representable with a neat and unique preference ordering. This dispositional model still identifies preferences with a propensity to sample as opposed to an explictly represented distribution, but of course the disposition now explains very little beyond what can be explained by saying that the preference ordering is represented implicitly. Presumably the best fitting dispositional model would take this agent to be using massive numbers of samples, so I'll call this the massive sampling model. Just like the tiny agent model, the massive sampling model seems unexplanatory: the variance in context that dispositions were meant to explain has been nearly eliminated. This makes the sampling algorithm itself look overly complex and unparsimonious.

The point of these two examples is this: as we move from more consistent, classical behaviors to more sporadic, variable behaviors, we see that at the extremes, theoretical virtues do make a difference. These considerations seem to favor the structural model over the dispositional at the classical and consistent end, and the dispositional over the structural at the

sporadic and variable end (ceteris paribus). But the difference between the extremes mentioned above and the case of preference reversal shows that in this case we're somewhere in the middle of the spectrum: the documentation of preference reversals in many domains of decision-making is evidence enough that our behavior is not deeply classical, but the optimality of human choice in general would suggest we are not fully sporadic either. We can also raise an epistemic obstacle to the application of theoretical virtues. Given persistent disagreement between psychologists who think of humans as highly sporadic and those who model us as nearly perfectly rational (and everything in between), we may even be unable to locate where humans are on the spectrum of consistency, whether in the case of preference reversals or more generally. This would suggest that theoretical virtues are unlikely to provide a neat fix for our problem.

In summary, in the case of fitting ideal rational choice models to understand actual human behavior, we find that the structural model and the dispositional model are not just two theories waiting on a decisive piece of neural or psychological evidence. Because this debate rests on the representational encoding of concepts that are merely functional in their original ideal formulations, more ideal decision theory cannot help determine these representational nuances. And while some empirical evidence might help make one or both of these theories look overly complex or otherwise off, there's a wide swath of empirical possibilities that seem consistent with both. In fact, this difficulty seems to also originate in under-specification from the ideal level: the concepts we're trying to apply are just not fine-grained enough to make decisive predictions.

4. Generalizing the Problem

Is this problem unique to the current debate about preferences, or to the slightly broader one about applying ideal decision theory to psychology? I'll now suggest that the structural/dispositional divide occurs all over in different sorts of approximate rationality debates, and that at a more general level, the problem of stretching ideal concepts beyond their intended commitments also recurs in other ideal/descriptive interfaces beyond rationality.

In the context of approximate rationality, the structural and dispositional theories can be juxtaposed in a few different versions: centering on preferences, as I've developed them here,

or credences, as in Icard's actual account, as well as various related permutations. In each of these cases, sampling models will have a structural and a dispositional interpretation. Further, these elements of a decision-theoretic agent are closely connected, and so it might be somewhat unnatural to adopt a structural interpretation for credences and a dispositional interpretation for utilities, for instance. Credences are also typically evaluated according to synchronic and diachronic coherence, and these properties are rationally significant. We can replace IIA, then, with a corresponding credal coherence principle: the structural model in this case explains the appearance of contextual credal change (or "reversal") by the inaccessibility of the underlying, coherent credence function, whereas a dispositional account explains the contextual shift by appeal to features of the disposition to sample. As before, these differences would lead to evaluatively significant consequences. To simplify a bit, the issue becomes: are we sadly unable to access our true, hidden, credence function or do we have somewhat complex credences that depend on context? And we might find similar reasons on either side as well: perhaps the structural account is too humuncular, or the dispositional account insufficiently representational.

Further, cases where theorists employ sampling models could in principle always be interpreted according to the dispositional or structural model. This is because there is nothing special about the application of the model in the context we started with: sampling can always be thought of as uncovering a "real" underlying distribution or a convenient expression of a disposition. However, the two interpretations will not always be similarly plausible: the balance can shift based on context, or on the mathematical implementation.

First, the context of application can shift the balance of plausibility away from a true impasse. For example, sampling models of various kinds have been applied to the problem of learning new theories, such that an agent is modeled as sampling from the space of possible theories during learning. In this application, the dispositional interpretation seems clearly better than the structural interpretation: it's more natural to think of the space of all possible states as implicit and merely instantiated in a disposition than to think of it as really encoded or otherwise there in the agent. This is in part because the space of possibilities is usually massively complex, but also because if it's not "in" the agent, this space still has an independent reality. The pressure to adopt the structural model is thus reduced. Subjective preferences, on the other hand, in a sense should be in the agent if they are to be anywhere.

Second, sampling processes can have different mathematical forms, and different degrees of noise and bias. For example, comparing slice sampling and Metropolis-Hastings sampling (two forms of Markov Chain Monte Carlo sampling[6]), we find that Metropolis, unlike slice, produces results sensitive to an initial choice of step size (i.e. how far apart steps are in the random walk process, a quantity that is fixed in Metropolis but dynamic in slice). In Metropolis, a greater share of the output will be explained by features of the algorithm rather than features of the underlying distribution. This difference makes the dispositional interpretation for Metropolis comparatively more attractive than for slice, ceteris paribus[7]. Likewise, the more biased an algorithm, the greater the distance between an explanation based solely on the underlying distribution and the best explanation. Noise is a bit different than bias, however, since noise will always reduce on repeated sampling whereas some biases may remain — the structural model is at home in cases where repeated sampling gets closer and closer to the underlying disposition, since the epistemic barrier between the agent and her underlying distribution is in the simplest case increasingly eroded by the acquisition of more information. Of course, many epistemic barriers induce bias in addition to noise, but when we have either an immense amount of bias or a bias that is complex, more bells and whistles will need to be added to the structural account. So not all uses of sampling processes in approximate rationality will give rise to a troubling impasse, though in principle the pair of interpretations will be available.

Another area where ideal frameworks stop short is in modeling agents over time. Here we see a re-capitulation of the individuation problem. Most people come to want different things over time. One way to describe this is epistemic: people come to learn what they really want over time. As in the structural model, this approach takes the underlying values to be constant, but separated from the decision-maker by some sort of barrier. This is the view Richard Pettigrew (2020) calls the One True Utility Solution to the problem of choice over time. In this debate, the dispositional model could take a few different forms: we might take the relevant disposition to

[6] Markov Chain Monte Carlo (MCMC) sampling is a way of taking a random walk through a potential situation and drawing an inference about the underlying features as you go. It uses linked samples that have the Markov property, that is it treats the likelihood of each next point in the walk as independent of the past conditional on the present. In these algorithms, we try to estimate the properties of an underlying distribution through many iterations of these walks. Each walk is divided into steps that go from the current location of sampling to a new one, and these two versions of MCMC differ on how these step sizes are determined.

[7] There are ways in which Metropolis differs from slice that might cut in the other direction, which I ignore here.

be so contextual so as to support talking about agents in vastly different contexts as having different values. Or we might adopt a view on which preferences are not uncovered over time, but are instead the same dispositions they always were, just leading to different behaviors in different contexts: what Pettigrew calls the Unchanging Utility Solution. Emphasizing the rational significance of this debate, the choice of what to identify as the utilities drives a choice in how to understand norms for choice in these temporal problems, the project of Pettigrew's book.

A second context where the structural/dispositional dichotomy appears is in the rational analysis of memory. Gershman (2021) lays out the memory problem as one of encoding traces that must subsequently be decoded. This is clearly a structural interpretation. The trace to be decoded is really there, it's not that we're merely disposed to come up with it in the right context. A dispositional interpretation here might be built around the constructive episodic simulation theory of memory (Addis et al (2008)). This account holds that rather than being encoding, episodic memories are constructions that are generated at recall. And yet, we presumably have some standing disposition to construct this or that simulation that exists over time, explaining individual constructive episodes. In this case, the two competing models can be thought of as structural and dispositional, but it would take some work to bring them under the same formalism as I have with the two models in the present paper.

Outside of individual rationality, other contexts in which the ideal meets the descriptive include political philosophy and evolutionary science. In the former, notions such as "justice" are sometimes formulated in a purely normative context, and then applied to actual circumstances. In the latter, an idealized model of how traits are selected for is sometimes stretched downwards to fit real natural history. It is unfortunately beyond the scope of this paper to tackle these contexts. However, I note the abstract similarity in structure as an invitation to the reader to consider whether similar impasses may arise in these disparate domains.

The individuation problem I initially described can be formulated whenever we move from a classical representation of a distribution smoothly connected to decision-making and action, to a setting where a sampling process intercedes between the underlying distribution and the decision rule. Except at the extremes of minimal or maximal contextual variation in decision-making, both the structural and the dispositional model have claims to correctly describe the representational structure, and both will have distinct rational implications. A more general

version of the problem occurs where noise is inserted between the joints of canonical representations — in these cases, including representational debates in memory and preference change, we can recognize a related structural/dispositional standoff even without the device of the sampling process.

3 Conclusion

I've presented an impasse between two ways of interpreting exactly the same formalism. These two ways are not trivially different: they entail significant differences in understanding the rationality of the agent being modeled. I've argued that purely ideal considerations are too coarse-grained to adjudicate between these competing theories, and that without a suitable criterion for applying concepts like 'preferences', empirical data will not be of any help either.

This problem illustrates the double-edged sword of the breadth of ideal, formal methods. On the one hand, the lack of constraint on internal workings is what enables decision theory to be employed as a model of agents who are composed in varied ways, giving us a way to evaluate a biology lab determining which experiments to pursue and a traffic control system under the same umbrella. On the other hand, precisely this lack of specificity opens up a gap between the sparse nature of the categories and the work we want to put them to in determining which part, if any, of an imperfect agent is the best fit. An imperfect agent is one whose operations fall short of a rational standard. But along with this, an imperfect agent almost always differs from a perfect agent in terms of internal composition: the imperfect agent's inner workings are divided into pieces that are separated by noise, and biased approximations such as sampling methods create newly distinct categories, such as the disposition to sample. To understand these imperfect agents and even to hold them up to a rational standard requires determining how their parts correspond to the parts of an ideal agent, what counts as the utilities, preferences, and credences. The dispositional and structural models are two families of such an assignment that show up all over as rival descriptions of imperfect agent beyond the cases of sampling models.

Approximate rationality is an important and perhaps even essential project. But we need a further source of constraint in order to move forward without making the mistake of fetishizing the merely notational features of our mathematical models or being lost in a sea of incomparable but distinct theories. Where could further constraints come from? Not from purely ideal analysis, nor from descriptive science alone. Instead, I want to suggest that this problem can only be

solved by bringing in new considerations that are in part normative — but not fully ideally so. These normative constraints would in that sense be sui generis to approximate rationality. They would tell us something about efficiency, partial coherence, learning trajectories, and so on. Perhaps these constraints are even already part of our philosophical arsenal, but we have yet to recognize their character.

Acknowledgements

Thanks to audiences at York University and UT Austin, as well as Daniel Drucker, Uriah Kriegel, Emily Liquin, and Tania Lombrozo.

References

Addis, Donna Rose and Schacter, Daniel L. 2008. Constructive episodic simulation: Temporal distance and detail of past and future events modulate hippocampal engagement. *Hippocampus*, 18(2):227–237.

Anderson, John R. 1990. *The adaptive character of thought*. Hillsdale, NJ: Erlbaum.

Arrow, Kenneth J. 1951/2012. *Social Choice and Individual Values*. Yale University Press.

Bradley, Richard. 2017. Decision theory with a human face. Cambridge University Press.

Davidson, David. 1980/2001. *Essays on Actions and Events: Philosophical Essays Volume 1*. Clarendon Press.

Dennett, Daniel C. 1988. Précis of the intentional stance. *Behavioral and brain sciences*, 11(3):495–505.

Eells, Ellory and Harper, William L. 1991. Ratifiability, game theory, and the principle of independence of irrelevant alternatives. *Australasian Journal of Philosophy*, 69(1):1–19.

Falkenhainer, Brian, Kenneth D. Forbus, and Dedre Gentner. 1989. The structure-mapping engine: Algorithm and examples. *Artificial intelligence*, *41*(1), 1-63.

Gentner, Dedre, Sarah Brem, Ronald W. Ferguson, Arthur B. Markman, Bjorn B. Levidow, Phillip Wolff, and Kenneth D. Forbus. 1997. Analogical reasoning and conceptual change: A case study of Johannes Kepler. *The journal of the learning sciences*, *6*(1), 3-40.

Gershman, Samuel J. 2021. The rational analysis of memory. In *Oxford handbook of human memory*. Oxford University Press Oxford, UK.

Gibbard, Allan. 2012. *Meaning and normativity*. Oxford University Press.

Howes, Andrew, Paul A. Warren, George Farmer, Wael El-Deredy, and Richard L. Lewis. 2016. Why contextual preference reversals maximize expected value. *Psychological review*, 123(4):368.

Icard, Thomas. 2016. Subjective probability as sampling propensity. *Review of Philosophy and Psychology*, 7(4):863–903.

Joyce, James M. 2005. How probabilities reflect evidence. *Philosophical perspectives*, 19:153–178.

Kornblith, Hilary. 2002. *Knowledge and its Place in Nature*. Oxford University Press.

Markman, Andrew B., & Gentner, Dedre. 1993. Structural alignment during similarity comparisons. *Cognitive psychology*, *25*(4), 431-467.

Pettigrew, Richard. 2020. *Choosing for changing selves.* Oxford University Press, USA.

Railton, Peter. 2017. At the core of our capacity to act for a reason: The affective system and evaluative model-based learning and control. *Emotion Review*, *9*(4), 335-342.

Ramsey, Frank P. 1931. The foundations of mathematics and other logical essays. Number 214. K. Paul, Trench, Trubner & Company, Limited.

Rulli, Tina, and Alex Worsnip. 2016. IIA, rationality, and the individuation of options. *Philosophical Studies*, *173*(1), 205-221.

Savage, Leonard J. 1972. *The foundations of statistics.* Courier Corporation.

Sen, Amartya. 2017. *Collective choice and social welfare*. Harvard University Press.

Simon, Herbert A. 1955. A behavioral model of rational choice. *The quarterly journal of economics*, *69*(1), 99-118.

Soltani, Alireza, Benedetto De Martino, and Colin Camerer. 2012. A range-normalization model of context-dependent choice: a new model and evidence. *PLoS computational biology*, *8*(7), e1002607.

Staffel, Julia. 2020. *Unsettled thoughts: A theory of degrees of rationality*. Oxford University Press, USA.