Toward Boundedly Rational Analysis

Thomas Icard (icard@cmu.edu)
Department of Philosophy, Baker Hall 135
Carnegie Mellon, Pittsburgh, PA 15213-3890 USA

Abstract

The Bayesian program in cognitive science has been subject to criticism, due in part to puzzles about the role of rationality and approximation. While somewhat sympathetic with these concerns, I propose that a thoroughgoing *boundedly rational analysis* strategy can answer to some of them. Through simulation results I illustrate the method by showing how one can retrodict recently reported results about particle filter models of categorization (Sanborn et al., 2010). I also introduce new obstacles that surface once we take bounded rationality seriously. Specifically, again through simulation, I show that the analysis of optimal sampling from Vul et al. (2014) is interestingly complicated by the introduction of agents capable of metareasoning. Under broad conditions, such agents outperform all uniform k-sampling agents. This motivates the computational study of boundedly rational metareasoning in its own right.

Keywords: rational analysis, bounded rationality, algorithmic level, sampling, categorization, metareasoning.

The Rational Analysis Strategy

The program of rational analysis, pioneered by Marr and Poggio (1976), and greatly extended by Anderson (1990), seeks to understand cognition in terms of rational solutions to underlying problems the mind is assumed to be solving. Many cognitive phenomena can be characterized as inference problems under uncertainty, where some latent state must be inferred on the basis of observed information. For such problems Bayesian methods provide a robust and well understood notion of optimality or rationality (DeGroot, 2004). Bayesian models of cognition have become increasingly popular in recent years, and have been applied to phenomena as diverse as vision, causal learning, language understanding, and intuitive physics (see Griffiths et al. 2008; Tenenbaum et al. 2011). This work typically understands inference as conditionalization on a probability distribution assumed to capture the subject's 'intuitive model' of the situation or domain.

Rational analysis, and the Bayesian instantiation thereof, is sometimes used to show in what sense a given behavior can be understood as rational. Even when we have a mechanistic understanding of how some cognitive function works, a rational analysis can shed light on why it works the way it does, often generating new testable predictions (e.g., Movellan and McClelland 2001). A more ambitious use of the method is in guiding our search for mechanisms in the first place. Indeed, one of the motivations behind Marr and Poggio's and Anderson's proposals was to narrow down the search space of cognitive models, by assuming that the right model must be one that at least approximately solves the underlying (e.g., inference) problem. In that way, progress on the *computational level* problem (Marr, 1982)—in addition to being worthwhile in its own right—may also promise progress in the search for more mechanistic, biologically detailed, models of cognition.

Anderson (1990) proposed his often-rehearsed six steps comprising the rational analysis strategy:

1. Precisely specify the goals of the cognitive system.
2. Develop a formal model of the environment to which the system is adapted.
3. Make minimal assumptions on computational limitations.
4. Derive the optimal behavior given items 1 through 3.
5. Examine the empirical literature to see if the predictions of the behavioral function are confirmed.
6. If the predictions are off, iterate.

He then illustrated the methodology with four example domains: memory retrieval, category learning, judging causal strength, and problem solving (i.e., decision making). In each case, he was able to show a good fit to a wide range of data, demonstrating that the strategy can be successful.

Anderson's Rational Model of Categorization

As a running example, consider Anderson's (1990; 1991) analysis of categorization. One may object to some of the assumptions behind the model, but categorization *per se* is not our focus—it is merely intended as a useful illustration.

A subject is assumed to observe a sequence of objects with various combinations of features; upon observing a new object, the subject needs to make an inference about an unobserved feature. Let Y_i be the value of the feature of interest for the ith observed object, and let X_i be the ith vector of values for the remaining features. Let us furthermore abbreviate the conjunction of the first n values, $Y_1,\ldots,Y_n$ and $X_1,\ldots,X_n$, as $\mathbf{Y}_N$ and $\mathbf{X}_N$, respectively. The problem facing the subject is to infer the value of Y_n, given observations of $\mathbf{X}_N$ and $\mathbf{Y}_{N-1}$, i.e., to find the value of Y_n that maximizes the posterior probability $P(Y_n \mid \mathbf{X}_N, \mathbf{Y}_{N-1})$ of Y_n given $\mathbf{X}_N$ and $\mathbf{Y}_{N-1}$.

Anderson argued that the ideal way to determine such probabilities would be to consider all possible clusterings of the first n objects, figuring out the probability of each, and using the clusterings to determine the probability of Y_n given X_n. In other words, one must determine the value of a latent variable Z_n, which corresponds to a clustering of the n objects observed so far. Then we can determine the posterior probabilities by summing over the possible clusterings:

$$P(Y_n \mid \mathbf{X}_N, \mathbf{Y}_{N-1}) \;=\; \sum_{Z_n} P(Y_n \mid Z_n)\, P(Z_n \mid \mathbf{X}_N, \mathbf{Y}_{N-1})\,. \quad (1)$$

$P(Z_n \mid \mathbf{X}_N, \mathbf{Y}_{N-1})$ is given in terms of Bayes Rule:

$$P(Z_n \mid \mathbf{X}_N, \mathbf{Y}_{N-1}) \;\propto\; P(\mathbf{X}_N, \mathbf{Y}_{N-1} \mid Z_n)\, P(Z_n)\,. \quad (2)$$

The first term is just the likelihood of the features given a clustering, which we also need to compute $P(Y_n \mid Z_n)$ in Eq. (1). It is given by a beta distribution (features are assumed to be independent, conditional on a clustering):

$$P(Y_k = v|Z_n) = \frac{\#_v + \beta}{\# + 2\beta}\,.$$

Here # is the number of objects Z_n clusters together with the kth object; and $\#_v$ is the number of those objects in the same cluster that have v as their Y-value.

The prior term for $P(Z_n)$ has one free parameter c, the *coupling parameter*, which determines how likely an object is to belong to a new clustering. The explicit form of the prior is rather complicated (see Anderson 1990, 1991 or Sanborn et al. 2010); it is easiest to understand as resulting from a sequential process so that clustering Z_{n+1} extends Z_n with distribution $P(Z_{n+1} = j|Z_n)$, given by cases:

$$\begin{cases} \frac{c \cdot M_j}{(1-c)+c \cdot n} & \text{if } j \text{ assigns the new object to an old cluster} \\ \frac{1-c}{(1-c)+c \cdot n} & \text{if } j \text{ assigns the new object to a new cluster} \end{cases}$$

M_j is the number of objects already in the cluster to which j assigns the new object, according to Z_n. With this prior, the more often objects are categorized as part of a particular cluster, the more likely new objects are to fall under that cluster.

The computations required by Eq. (1) are intractable. As Anderson pointed out, the number of clusterings Z_n grows exponentially. For $n = 10$, there are already $115,975$ possible clusterings, making the sum in Eq. (1) prohibitive in all but the simplest of cases. This is not an atypical feature of 'ideally rational' Bayesian models. Step 3 of the methodology above says to make minimal assumptions on such limitations. In addition to the constraint that the required computations should be tractable, Anderson also assumed that at any given time, a subject ought to have settled on a particular clustering of objects seen so far, so that as new objects are observed, the only question is how to extend that partition to include the new object. This led him to the following proposal:

> LOCAL MAP ALGORITHM: Upon observation of a new object with features X_n, let Z_n^* be the extension of the current partition Z_{n-1} that maximizes $P(Z_n \mid \mathbf{X}_N, \mathbf{Y}_{N-1})$. One can then estimate $P(Y_n \mid \mathbf{X}_N, \mathbf{Y}_{N-1})$ by calculating:
>
> $$\tilde{P}(Y_n \mid \mathbf{X}_N, \mathbf{Y}_{N-1}) \;=\; P(Y_n \mid Z_n^*)\, P(Z_n^* \mid \mathbf{X}_N, \mathbf{Y}_{N-1})\,. \qquad (3)$$

That is, instead of summing over all partitions every time one needs to make a prediction, Anderson's local algorithm has the subject deterministically choosing the maximum *a posteriori* (MAP) partition following each new data point. Eq. (3) is supposed to be a tractable version of Eq. (1).

Anderson showed that his local MAP algorithm was able to account for a wide array of empirical phenomena collected from over two decades of work on categorization, including order effects, prototype effects, the relative ease of learning different categories of Boolean concepts, and several more (see Anderson 1990, 1991 for discussion).

Rationality and Approximation: Criticisms

Despite its success in modeling diverse cognitive phenomena, the Bayesian program as a whole has come under criticism recently (Jones and Love, 2011; Eberhardt and Danks, 2011; Bowers and Davis, 2012; Marcus and Davis, 2013). My focus here will be on two recurrent themes of this criticism.

In light of the intractability of computations like that in (1) above, one might wonder what role these 'ideal Bayesian' models are supposed to play. As the categorization example demonstrates, so far as concrete mechanisms are concerned, the ideal model can at best help focus our search for tractable models, as approximations to that ideal. Once we give up on the ideal as a model for the cognitive mechanism, however, one might reasonably worry that the link to rationality is severed. If people are approximating Bayesian solutions, then in what sense is their behavior really Bayesian? More broadly, in what sense is a Bayesian approximation rational?

This worry is coupled with a related, empirical criticism. In much of the experimental data used to support Bayesian models, the distribution of responses is shown to match the posterior distribution for the proposed model. On the face of it, this looks like a disconfirmation that people's individual behavior is Bayesian. Assuming MAP inference is the ideal, it would appear that most individual subjects are behaving irrationally. This raises the challenge of specifying when a given Bayesian analysis is vindicated by the data, and when people's behavior has been genuinely rationalized. If the hypothesis that people are (in an appropriate sense) Bayesian is to be falsifiable, it must be possible to find instances where a Bayesian analysis would be inappropriate.

The Sampling Hypothesis

These criticisms have been partially addressed by a line of work proposing that people do not explicitly calculate posterior distributions, but rather *sample* from the appropriate posteriors. This *Sampling Hypothesis* (e.g., Vul et al. 2014) accounts for posterior matching under the assumption that each subject in an experiment is drawing relatively few, e.g., one or two, samples from the normative posterior.

Moreover, this behavior can be rationalized in a certain sense. Assuming additional samples from a distribution come at a cost, under certain further assumptions about the utilities and probabilities, Vul et al. (2014) showed that it can be optimal to draw only a single sample before making a decision.

In addition to the generic posterior matching phenomenon, concrete sampling algorithms, e.g., based on Markov chain Monte Carlo, have been used to explain more specific cognitive phenomena (for many references, see Griffiths et al. 2008; Tenenbaum et al. 2011), including in categorization.

Sanborn et al. (2010), for example, showed that the fit of Anderson's model to the data on categorization could be improved by replacing his Local MAP Algorithm with one based on the *particle filter*. Instead of choosing the MAP clustering at each stage, the subject maintains at any given time a set of R 'particles', each corresponding to a clustering,

and bases inferences on the whole set:

> PARTICLE FILTER ALGORITHM (SANBORN ET AL.): Upon observation of a new object with features X_n, draw samples $Z_n^{(1)}, \ldots, Z_n^{(R)}$ from $P(Z_n \mid \mathbf{X}_N, \mathbf{Y}_{N-1})$. One can then approximate $P(Y_n \mid \mathbf{X}_N, \mathbf{Y}_{N-1})$ by calculating:

$$\tilde{P}(Y_n \mid \mathbf{X}_N, \mathbf{Y}_{N-1}) = \sum_{r=1}^{R} P(Y_n \mid Z_n^{(r)})\, P(Z_n^{(r)} \mid \mathbf{X}_N, \mathbf{Y}_{N-1}). \quad (4)$$

They compared the MAP Algorithm with the cases of $R = 1$ and $R = 100$ particles. For several empirical findings, all three provided a good fit. For order effects, the single-particle-filter and the MAP algorithm were closer to the human data than the 100-particle-filter. However, one characteristic of a particle filter with few particles—not possessed by the MAP algorithm—is its ability to predict individual variation. In line with the posterior matching behavior described above, when two clusterings Z and Z' have roughly equal probability, but that of Z is marginally higher, the MAP algorithm will always settle on Z, while the R-particle-filter will noisily choose between Z and Z'. This behavior is in fact borne out in the experiments that Anderson himself described. Thus, the particle filter algorithm—also touted as being *more rational* than the MAP algorithm, as it is an approximation to the ideal model in a precise sense—models the human data at least as well as the MAP algorithm, and in some ways even better.

Does Sampling Answer to the Criticisms?

The Sampling Hypothesis goes some way toward answering the criticisms described above. However, it leaves some important questions unanswered. First, the result from Vul et al. (2014)—that it may be optimal to draw only a few samples—assumes from the start that an agent will make its decision on the basis of some number of samples from a given model. In particular, the analysis does not compare drawing samples from the model with any other non-sampling algorithms. Unless we have some independent reason for assuming sampling is the only candidate algorithm, this strategy does not properly follow step 4 of Anderson's program.

The same worry applies to both the MAP and particle filter approximations to the rational model of categorization. Anderson (1991) made some informal remarks about why the MAP algorithm is rational, perhaps even optimal given certain constraints (412); and Sanborn et al.'s (2010) model is pitched as a 'more rational approximation' because (4) asymptotically converges to the ideal (1). However, neither has given a convincing argument for why one or the other is rational in any sense we would care about. Why would an agent using an approximation to the model in (1) be well adapted to its environment? In particular, why would such an agent be better adapted than one who uses some other algorithm that does not approximate the ideally rational model?

Boundedly Rational Analysis

It is commonly assumed that a computational level analysis constrains the algorithmic level analysis. This is not always reasonable, however. Sometimes, once computational costs are properly taken into account, the optimal algorithm looks nothing like the ideal model or any straightforward approximation thereto (examples to follow).

One of the primary messages of this paper is that, once we take costs seriously, we should no longer think of the 'problem being solved' as being one of pure inference, inherited from the computational level; instead we should think of the algorithmic problem to be solved as one of constraint optimization: make the best guess subject to memory, time, energy, and other cost constraints. This idea is of course familiar from early work by Simon (1957), and emphasized by many since (e.g., Gigerenzer and Goldstein 1996). The ideal model in (1) epitomizes what Simon referred to as a *substantively rational* solution. What we want, as part of a rational analysis, is a *boundedly rational*, or *procedurally rational*, solution.

Sketch of a Theory of Bounded Rationality

Let us model an agent's environment using a prior probability distribution $P(H)$ over latent states of the world H, together with a likelihood function $P(D_1, \ldots, D_n|H)$ for sequences of n observations. Thus, upon making the first n observations $\mathbf{D} = D_1, \ldots, D_n$, the posterior probability for H is given by Bayes Rule: $P(H|\mathbf{D}) \propto P(\mathbf{D}|H)\, P(H)$. Our agent will face a decision problem, with some set $\mathcal{A} = \{A_1, \ldots, A_m, \ldots\}$ of possible actions, and a utility $u(A_i, H) \in \mathbb{R}$ for all $A_i \in \mathcal{A}$ and each value of H. Call the initial distribution, a sequence of observations, and a decision problem together a *scenario*.

Making no assumptions about the agent's computational limitations, we can define an *agent function* α to be a mapping from observations $\mathbf{D}$ to a distribution $\alpha(\mathbf{D})$ over $\mathcal{A}$. That is, $\alpha(\mathbf{D})$ assigns a probability to each $A_i \in \mathcal{A}$. The *fitness* ϕ of an agent function α is given by:

$$\phi(\alpha) = \sum_H P(H) \cdot \sum_{\mathbf{D}} P(\mathbf{D}|H) \cdot \sum_j \alpha(\mathbf{D})(A_j) \cdot u(A_j, H) \quad (5)$$

Nature chooses a state H and generates some observations $\mathbf{D}$ based on H; then the agent must take an action A_i; the payoff is the weighted sum of utility for each of the actions it might take. When it exists, an optimal (highest fitness) agent function α^* is one that never chooses an action A_i whose expected utility under the posterior distribution (conditioned on $\mathbf{D}$) is dominated by another action A_j:

> OPTIMAL AGENT FUNCTIONS: α^* is optimal if for all A_i and $\mathbf{D}$: $\alpha^*(\mathbf{D})(A_i) = 0$, whenever there is $A_j \in \mathcal{A}$, such that $\sum_H P(H \mid \mathbf{D})\, u(A_j, H) > \sum_H P(H \mid \mathbf{D})\, u(A_i, H)$.

This notion of optimality captures the computational level problem. On Anderson's analysis of categorization, H is the state of the world, a specification of all the properties of all the objects in the world. The observations are sequences of objects; then upon viewing a new object, the agent must act appropriately, depending on an unobserved property of this new object. In many cases we can simply assume that the actions correlate one-to-one with the possible values of the

unobserved variable (e.g., poisonous : avoid, nutritious : consume), and the utility is positive for a correct guess, and zero or negative for an incorrect guess, for example.

Given that most problems of interest are hard, with the associated optimal agent functions intractable, we want to study not just abstract agent functions, but more concrete representations of agents and the actual computations they perform. Suppose we have fixed some class Π of programs in a given language. We can think of programs $\pi \in \Pi$ as reflecting the mental steps an agent goes through in the course of receiving data $\mathbf{D}$ and deciding which action A_i to perform. Following each new data point D_k, there is some distribution over $\mathcal{A}$ reflecting the agent's proclivities to perform various actions, at that point in time. In this way π refines a more abstract agent function α_π. Let us suppose, very abstractly, that we can associate with a given program π, under a certain scenario, an expected cost: $C_\pi(\mathbf{D})$. The *cost-adjusted fitness* of π is then:

$$\phi(\pi) \;=\; \phi(\alpha_\pi) - C_\pi \,, \tag{6}$$

where $C_\pi = \sum_H P(H) \cdot \sum_{\mathbf{D}} P(\mathbf{D} \mid H) \cdot C_\pi(\mathbf{D})$ is the overall expected cost. That is, we take the fitness of the program's associated agent function less expected costs. An agent is *boundedly rational* to the extent that the cost-adjusted fitness of its program is high (cf. Russell and Subramanian 1995).

In this setup, we can interpret Vul et al.'s (2014) results as showing that, if we take Π to include the k-samplers for all k, then for certain values of C and in certain decision problems, the 1-sampler is most boundedly rational.

Boundedly Rational Categorization

Finding a boundedly optimal algorithm can be difficult in general. However, it is often possible to compare the (cost-adjusted) fitness of algorithms in simulated environments. For illustration, we performed a comparison between Anderson's local MAP algorithm, several particle filter algorithms ($R = 1, 2, 5, 10$), and a baseline 'reflex agent', which makes random predictions on new observations, and maximizes with respect to count frequency on previously observed objects.

Specifically, we generated sequences of data according to the Dirichlet process described above, consisting of N objects varying along d binary dimensions, before generating a test object with some feature hidden. Each agent observes the N objects, updating its representation at each step, and then makes an inference about the hidden feature, receiving payoff 1 if correct, 0 otherwise. With parameter settings typical of the categorization experiments reported in Anderson (1990, 1991), the results are depicted in Table 1.

In all scenarios, the 10-particle-filter agent is optimal. The next highest performing agent is highlighted in bold. With very few observations, the 5-particle-filter outperforms the MAP algorithm, while we find the opposite result with more observations. With an intermediate number they are on a par.

If a rational analysis is to be guided toward the algorithm that appears most rational, and the empirical results suggest that the 1-particle-filter provides a better fit to human data than the MAP algorithm, then these simulation results may look discouraging. However, with a *boundedly* rational analysis, we would take costs into account.[1] Under the tentative assumption that it costs more to reduce noise than to tolerate some noise, it may be that the difference in expected fitness is made up for by this difference in cost. Furthermore, the 2- and 5-particle-filter agents are already competitive with the MAP algorithm. While Sanborn et al. did not explicitly study these agents (A. Sanborn, *p.c.*), it is easy to see that such algorithms would likely provide an equally good alternative. Recall that in many cases, the MAP, $R = 1$, and $R = 100$ algorithms all matched the data well. For the other cases, particle filters with 5 or fewer particles also exhibit order effects, and would predict individual variation. At the same time, if maintaining more particles comes at a cost, this cost would have to be very low for the increase from $R = 5$ to $R = 10$ particles, for example, to be worth the small gain in fitness.

	$d = 3, N = 3$	$d = 5, N = 8$	$d = 3, N = 30$
reflex	0.561	0.555	0.654
MAP	0.601	**0.626**	**0.656**
$R = 1$	0.573	0.603	0.625
$R = 2$	0.596	0.623	0.629
$R = 5$	**0.606**	**0.626**	0.642
$R = 10$	0.608	0.640	0.662

Table 1: Average payoffs for categorization agents, with d binary dimensions (i.e. features) and N observations. The coupling parameter is set to $c = 0.5$. In all cases, $\beta = 1$.

This small study is not conclusive, but it is quite suggestive. We can tentatively conclude that, under Anderson's own assumptions about the nature of the environment, reasonable assumptions about cost would result in a particle filter with between 1 and 5 particles being boundedly rational.

Calculation versus Look-up

Note that in Table 1, when $N = 30$ the reflex agent outperforms all but the MAP agent and the 10-particle-filter agent. Given the simplicity of the computations this agent performs, we would expect it to incur low costs, and thus to be quite boundedly rational in this scenario, according to the definition given above, perhaps the most boundedly rational of all six agents. In this particular categorization example, we might not want to assume the environment will be such that an agent will observe 30 objects before having to make a prediction. However, the example makes a more general point, that if we were to assume this did capture the structure of the environment, our boundedly rational analysis would not justify hypothesizing a more complicated agent type.

Consider a different example, inspired by a recent discussion in the vision literature (Maloney and Mamassian, 2009). Imagine a point estimation problem in which the underly-

[1] It is worth mentioning that with a higher coupling parameter, $c = 0.75$, the 1-particle-filter agent does outperform the local MAP agent, even ignoring costs. This is because the MAP agent is more often fooled by 'garden path' sequences, whereas the particle filter has some chance of escaping them (cf. Sanborn et al. 2010).

ing state of the world is drawn from a normal distribution $S \sim \mathcal{N}(\mu, \sigma_1^2)$, where μ is the mean and σ_1^2 is the variance. The agent obtains a noisy reading D of S, which is also described by a normal distribution around the true point S, i.e., $D \sim \mathcal{N}(S, \sigma_2^2)$, for some σ_2^2. With action space $\mathcal{A} = \mathbb{R}$, the utility function for making an estimate $\tilde{S}$ when the true value is S is given by the usual squared error, $U(\tilde{S}, S) = -(\tilde{S} - S)^2$. The optimal agent function is the one that maximizes fitness according to Eq. (5) as given above (minimizing expected error, making the necessary adjustments to Eq. (5) for the continuous setting). Once we consider agent *programs*, refining the more general agent function, several possibilities emerge. The agent could separately represent information about the state of the world and about the problem being solved and combine them in some appropriate way. Alternatively, it is possible for the agent to manifest the same behavior with a simpler method. Letting $\tau_1 = 1/\sigma_1^2$ and $\tau_2 = 1/\sigma_2^2$—the *precision* of S and D, respectively (DeGroot, 2004, 38)—the optimal agent function can also be described by the following:

$$\tilde{S} = \frac{\tau_1}{\tau_1 + \tau_2}\mu + \frac{\tau_2}{\tau_1 + \tau_2}D\,,$$

as a function only of the data point D. In other words, performing optimally in this task requires merely being able to apply a linear map of the form $x \mapsto a + bx$.

Simulating an agent that learns a and b through simple linear regression, with different settings of the learning parameter, we see how (boundedly) rational such an agent can be. When the learning parameter is as high as 0.1, it does reasonably well after only 10 trials but soon after levels off in performance, remaining suboptimal. If it is set lower, e.g., near 0.01, it takes much longer to perform well; but eventually, after about 100,000 trials, its performance is indistinguishable from the agent that straightforwardly computes Eq. (5) with known mean μ and variances σ_1^2 and σ_2^2. If the scenario in which we are assessing agent fitness is one where training time is cheap and trials are amply available, then this is another case where we would not be justified in assuming that an agent adapted to this setting will implement something directly approximating Bayesian calculations. There would be no reason for an agent facing this environment to represent separate information about the world, since it can apply a very simple 'look-up' rule to decide what to do.

Broader Scenarios

In order to justify the complexity required by (even merely approximately) Bayesian calculations, we need to consider more complex scenarios, in which there is either uncertainty about the problem to be faced, a sequence of problems to be faced, or both. For instance, if the categorizer may have to make decisions after 3, 8, 10, 30, etc. observations, then it may be important to make accurate predictions with relatively little data. Likewise, in the normal-normal example, the agent may not be able to go through 100,000 learning trials.

Other general features of Bayesian (and related) models, such as their capacity for generalization, abstraction, transfer, etc. (Tenenbaum et al., 2011), will also only surface by considering iterated scenarios, in which the results from one learning episode, for example, may benefit the learner in a later episode. Psychologists have been keenly aware of the need for 'inductive biases' to model empirical learning dynamics. But there is a normative aspect as well, in that agents exhibiting these general features will outperform agents lacking them in sufficiently broad scenarios. Importantly, approximations, such as sampling algorithms, inherit these general properties from the ideal models. A thoroughgoing boundedly rational analysis would invoke such considerations to show that one or another approximation is indeed more boundedly rational than the alternatives, if indeed it is.

Boundedly Rational Metareasoning

Once we introduce enough uncertainty over what problem the agent will face, it becomes substantially more difficult to determine what a boundedly rational solution will look like. In particular, we introduce the possibility that an agent may have the capacity to reason online about how to solve the problem it finds itself facing; that is, we introduce the possibility of boundedly rational metareasoning agents. To take an example, in the work described above by Vul et al. (2014), there is a single decision problem, and the analysis shows what the optimal sampling strategy is for that decision problem. If there is uncertainty over the decision problem, it turns out a simple metareasoner dominates all fixed k-samplers.

In computer simulations, we randomly drew parameters for a Bayes net, a state and an observation, and then randomly generated a decision problem that depended on some subset of the (five) variables, with utilities ranging between 0 and 100. A *sample cost* C was drawn from a normal distribution with $\sigma^2 = 1.0$ and $\mu \sim \text{Uniform}(0, 3)$. We then compared the performance of eight agents. The first seven drew fixed numbers k of (perfect) samples—1, 2, 3, 4, 5, 7, and 9—from the network, conditioned on the observation, and then made a decision using the obvious rule from Vul et al. (2014), incurring kC reduction in utility on account of the k samples.

The remaining metareasoning agent first applied a heuristic to determine how many samples to take. This heuristic χ was an extremely simple stepwise function depending only on the cost of a sample:

$$\chi(C) = \begin{cases} 1 & \text{if } 2.5 < C\,; \\ 2 & \text{if } 1.5 < C < 2.5\,; \\ 4 & \text{if } 1.0 < C < 1.5\,; \\ 9 & \text{if } 0.5 < C < 1.0\,; \\ 15 & \text{if } C < 0.5\,. \end{cases}$$

The intended interpretation is that χ captures the relative importance of the problem compared to the cost of sampling. Clearly, more sophisticated functions are conceivable, but this is already sufficient to make the point. The results of the simulations are given in Table 2.

Importantly, the metareasoning agent suffered the cost of each sample it decided to take, but it was also charged for the initial step of calculating $\chi(C)$. For the simulation results reported in Table 2, the cost of this step was assumed

	average utility
1-sampler	65.6
2-sampler	67.4
3-sampler	67.4
4-sampler	66.8
5-sampler	65.9
7-sampler	63.5
9-sampler	61.1
metareasoner	**68.5**

Table 2: Average payoffs, out of 100,000 runs.

to be equal to the cost of a single sample, which is arguably quite uncharitable for a simple step-function. Nonetheless, the metareasoner still performed significantly better. We also ran 100,000 iterations without charging for the preprocessing step, and in this case the metareasoner's average utility was 70.6. Thus, even if the cost of this simple preprocessing step is constantly 3.0—on average twice the cost of a sample (average sample cost is 1.5)—the metareasoner still outperforms the best constant samplers (the 2-sampler and 3-sampler).

In a broader scenario with multiple problems, if we include a metareasoner among the possible agents, it turns out to be optimal. Tellingly, Vul et al. (2014) found that people apparently do strategically adjust the number of samples they draw.

Conclusion

We have proposed a boundedly rational analysis strategy, as a way of making progress on the search for algorithmic-level models. This strategy promises answers to some common criticisms of Bayesian models in cognitive science. In the case of categorization, we have seen that such an analysis can retrodict the result observed by Sanborn et al. (2010) that a particle filter with few particles models the human data better than the 'ideal' model or Anderson's (1990; 1991) alternative.

The strategy requires taking a broad view of what bounded (instrumental) rationality means, and considering more extended scenarios in which agent performance is compared. This will not preclude (approximate) Bayesian analyses—on the contrary, I would conjecture—but it will not presuppose them either. While this is arguably necessary to respond fully to the criticisms, it also poses new challenges and avenues for research. Specifically, we saw the need to consider metareasoners in the space of possible agent algorithms. Does this open the door to an unmanageable search space, spoiling the apparent advantage of invoking rational analysis as a search strategy? To close, I would like to suggest that here, as elsewhere in cognitive science, we can divide and conquer: perhaps computational cognitive science ought to take this kind of metareasoning as an object of study in its own right.

Acknowledgements

This paper draws from my Ph.D. thesis (Icard, 2013). I would like to thank my committee, and David Danks, for helpful conversations, and the conference reviewers for comments.

References

Anderson, J. R. (1990). *The Adaptive Character of Thought*. Lawrence Earlbaum Associates, Inc.

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, 98(3):409–429.

Bowers, J. S. and Davis, C. J. (2012). Bayesian just-so stories in psychology and neuroscience. *Psychological Bulletin*, 138(3):389–414.

DeGroot, M. H. (2004). *Optimal Statistical Decisions*. John Wiley & Sons.

Eberhardt, F. and Danks, D. (2011). Confirmation in the cognitive sciences: The problematic case of Bayesian models. *Minds and Machines*, 21(3):389–410.

Gigerenzer, G. and Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review*, 103(4):650–699.

Griffiths, T. L., Kemp, C., and Tenenbaum, J. B. (2008). Bayesian models of cognition. In Sun, R., editor, *The Cambridge Handbook of Computational Cognitive Modeling*, pages 59–100. Cambridge University Press.

Icard, T. F. (2013). *The Algorithmic Mind: A Study of Inference in Action*. PhD thesis, Stanford University.

Jones, M. and Love, B. C. (2011). Bayesian Fundamentalism or Enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *Behavioral and Brain Sciences*, 34(4):169–231.

Maloney, L. T. and Mamassian, P. (2009). Bayesian decision theory as a model of human visual perception: Testing Bayesian transfer. *Visual Neuroscience*, 26:147–155.

Marcus, G. F. and Davis, E. (2013). How robust are probabilistic models of higher-level cognition? *Psychological Science*, 24(12):2351–2360.

Marr, D. (1982). *Vision*. W.H. Freeman and Company.

Marr, D. and Poggio, T. (1976). From understanding computation to understanding neural circuitry. MIT A.I. Memo 357.

Movellan, J. R. and McClelland, J. L. (2001). The Morton-Massaro Law of Information Integration: Implications for models of perception. *Psychological Review*, 108:113–148.

Russell, S. and Subramanian, D. (1995). Provably bounded-optimal agents. *Journal of Artificial Intelligence Research*, 2:1–36.

Sanborn, A. N., Griffiths, T. L., and Navarro, D. J. (2010). Rational approximations to rational models: Alternative algorithms for category learning. *Psychological Review*, 117(4):1144–1167.

Simon, H. A. (1957). *Models of Man*. Wiley.

Tenenbaum, J. T., Kemp, C., Griffiths, T. L., and Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, 331:1279–1285.

Vul, E., Goodman, N. D., Griffiths, T. L., and Tenenbaum, J. B. (2014). One and done? Optimal decisions from very few samples. *Cognitive Science*. forthcoming.