REVIEW

Computational rationality: A converging paradigm for intelligence in brains, minds, and machines

Samuel J. Gershman,[1]* Eric J. Horvitz,[2]* Joshua B. Tenenbaum[3]*

After growing up together, and mostly growing apart in the second half of the 20th century, the fields of artificial intelligence (AI), cognitive science, and neuroscience are reconverging on a shared view of the computational foundations of intelligence that promotes valuable cross-disciplinary exchanges on questions, methods, and results. We chart advances over the past several decades that address challenges of perception and action under uncertainty through the lens of computation. Advances include the development of representations and inferential procedures for large-scale probabilistic inference and machinery for enabling reflection and decisions about tradeoffs in effort, precision, and timeliness of computations. These tools are deployed toward the goal of computational rationality: identifying decisions with highest expected utility, while taking into consideration the costs of computation in complex real-world problems in which most relevant calculations can only be approximated. We highlight key concepts with examples that show the potential for interchange between computer science, cognitive science, and neuroscience.

Imagine driving down the highway on your way to give an important presentation, when suddenly you see a traffic jam looming ahead. In the next few seconds, you have to decide whether to stay on your current route or take the upcoming exit—the last one for several miles—all while your head is swimming with thoughts about your forthcoming event. In one sense, this problem is simple: Choose the path with the highest probability of getting you to your event on time. However, at best you can implement this solution only approximately: Evaluating the full branching tree of possible futures with high uncertainty about what lies ahead is likely to be infeasible, and you may consider only a few of the vast space of possibilities, given the urgency of the decision and your divided attention. How best to make this calculation? Should you make a snap decision on the basis of what you see right now, or explicitly try to imagine the next several miles of each route? Perhaps you should stop thinking about your presentation to focus more on this choice, or maybe even pull over so you can think without having to worry about your driving? The decision about whether to exit has spawned a set of internal decision problems: how much to think, how far should you plan ahead, and even what to think about.

This example highlights several central themes in the study of intelligence. First, maximizing some measure of expected utility provides a general-purpose ideal for decision-making under uncertainty. Second, maximizing expected utility is nontrivial for most real-world problems, necessitating the use of approximations. Third, the choice of how best to approximate may itself be a decision subject to the expected utility calculus—thinking is costly in time and other resources, and sometimes intelligence comes most in knowing how best to allocate these scarce resources.

The broad acceptance of guiding action with expected utility, the complexity of formulating and solving decision problems, and the rise of approximate methods for multiple aspects of decision-making under uncertainty has motivated artificial intelligence (AI) researchers to take a fresh look at probability through the lens of computation. This examination has led to the development of computational representations and procedures for performing large-scale probabilistic inference; methods for identifying best actions, given inferred probabilities; and machinery for enabling reflection and decision-making about tradeoffs in effort, precision, and timeliness of computations under bounded resources. Analogous ideas have come to be increasingly important in how cognitive scientists and neuroscientists think about intelligence in human minds and brains, often being explicitly influenced by AI researchers and sometimes influencing them back. In this Review, we chart this convergence of ideas around the view of intelligence as computational rationality: computing with representations, algorithms, and architectures designed to approximate decisions with the highest expected utility, while taking into account the costs of computation. We share our reflections about this perspective on intelligence, how it encompasses interdisciplinary goals and insights, and why we think it will be increasingly useful as a shared perspective.

[1]Department of Psychology and Center for Brain Science, Harvard University, Cambridge, MA 02138, USA. [2]Microsoft Research, Redmond, WA 98052, USA. [3]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139, USA.
*Corresponding author. E-mail: gershman@fas.harvard.edu (S.J.G.); horvitz@microsoft.com (E.J.H.); jbt@mit.edu (J.B.T.)

Models of computational rationality are built on a base of inferential processes for perceiving, predicting, learning, and reasoning under uncertainty (*1*–*3*). Such inferential processes operate on representations that encode probabilistic dependencies among variables capturing the likelihoods of relevant states in the world. In light of incoming streams of perceptual data, Bayesian updating procedures or approximations are used to propagate information and to compute and revise probability distributions over states of variables. Beyond base processes for evaluating probabilities, models of computational rationality require mechanisms for reasoning about the feasibility and implications of actions. Deliberation about the best action to take hinges on an ability to make predictions about how different actions will influence likelihoods of outcomes and a consideration of the value or utilities of the outcomes (*4*). Learning procedures make changes to parameters of probabilistic models so as to better explain perceptual data and provide more accurate inferences about likelihoods to guide actions in the world.

Last, systems with bounded computational power must consider important tradeoffs in the precision and timeliness of action in the world. Thus, models of computational rationality may include policies or deliberative machinery that make inferences and decisions at the "metalevel" in order to regulate base-level inferences. These decisions rely on reflection about computational effort, accuracy, and delay associated with the invocation of different base-level algorithms in different settings. Such metalevel decision-making, or "metareasoning," can be performed via real-time reflection or as policies computed during offline optimizations. Either way, the goal is to identify configurations and uses of base-level processes with the goal of maximizing the expected value of actions taken in the world. These computational considerations become increasingly important when we consider richer representations (graphs, grammars, and programs) that support signature features of human intelligence, such as recursion and compositionality (*5*).

Key advances in AI on computational machinery for performing inference, identifying ideal actions, and deliberating about the end-to-end operation of systems have synergies and resonances with human cognition. After a brief history of developments in AI, we consider links between computational rationality and findings in cognitive psychology and neuroscience.

Foundations and early history

AI research has its roots in the theory of computability developed in the 1930s. Efforts then highlighted the power of a basic computing system (the Turing Machine) to support the real-world mechanization of any feasible computation (*6*). The promise of such general computation and the fast-paced rise of electronic computers fueled the imagination of early computer scientists about the prospect of developing computing systems that might one day both explain and replicate aspects of human intelligence (*7*, *8*).

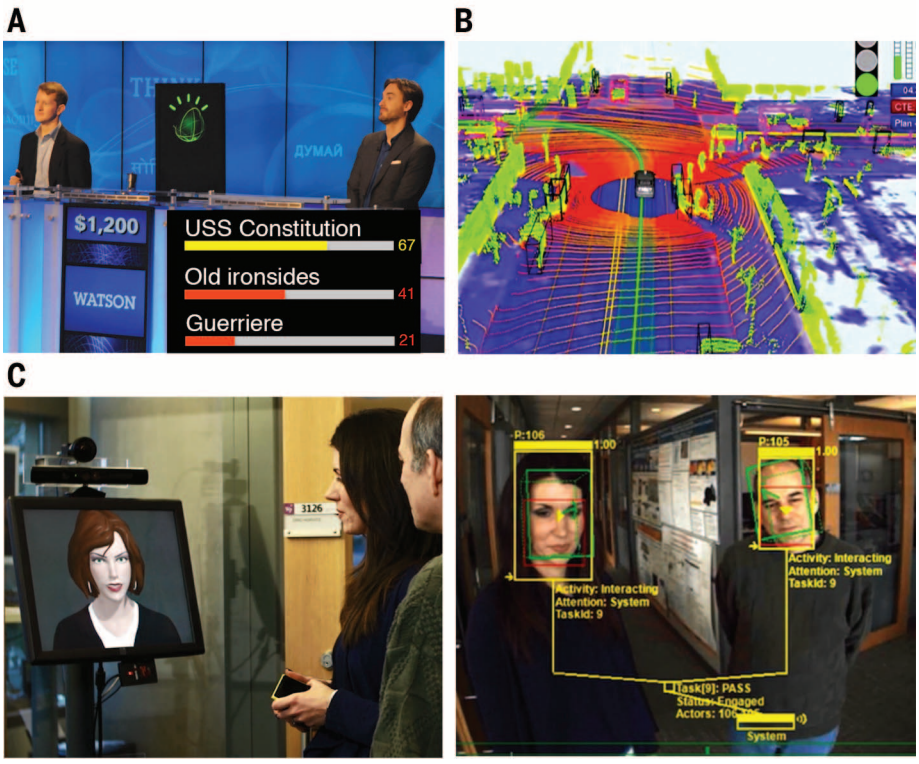

Fig. 1. Examples of modern AI systems that use approximate inference and decision-making. These systems cannot rely on exhaustive enumeration of all relevant utilities and probabilities. Instead, they must allocate computational resources (including time and energy) to optimize approximations for inferring probabilities and identifying best actions. (**A**) The internal state of IBM Watson as it plays *Jeopardy*!, representing a few high-probability hypotheses. [Photo by permission of IBM News Room] (**B**) The internal state of the Google self-driving car, which represents those aspects of the world that are potentially most valuable or costly for the agent in the foreseeable future, such as the positions and velocities of the self-driving car, other cars and pedestrians, and the state of traffic signals. [Reprinted with permission from Google] (**C**) The Assistant (left), an interactive automated secretary fielded at Microsoft Research, recognizes multiple people in its proximity (right); deliberates about their current and future goals, attention, and utterances; and engages in natural dialog under uncertainty. [Permission from Microsoft]

Early pioneers in AI reflected about uses of probability and Bayesian updating in learning, reasoning, and action. Analyses by Cox, Jaynes, and others had provided foundational arguments for probability as a sufficient measure for assessing and revising the plausibility of events in light of perceptual data. In influential work, von Neumann and Morgenstern published results on utility theory that defined ideal, or "rational," actions for a decision-making agent (*4*). They presented an axiomatic formalization of preferences and derived the "principle of maximum expected of utility" (MEU). Specifically, they showed that accepting a compact and compelling set of desiderata about preference orderings implies that ideal decisions are those actions that maximize an agent's expected utility, which is computed for each action as the average utility of the action when considering the probability of states of the world.

The use of probability and MEU decision-making soon pervaded multiple disciplines, including some areas of AI research, such as projects in robotics. However, the methods did not gain a large following in studies of AI until the late 1980s. For decades after the work by von Neumann and Morgenstern, probabilistic and decision-theoretic methods were deemed by many in the AI research community to be too inflexible, simplistic, and intractable for use in understanding and constructing sophisticated intelligent systems. Alternative models were explored, including logical theorem-proving and various heuristic procedures.

In the face of the combinatorial complexity of formulating and solving real-world decision-making, a school of research on heuristic models of bounded rationality blossomed in the later 1950s. Studies within this paradigm include the influential work of Simon and colleagues, who explored the value of informal, heuristic strategies that might be used by people—as well as by computer-based reasoning systems—to cut through the complexity of probabilistic inference and decision-making (*9*). The perspective of such heuristic notions of bounded rationality came to dominate a large swath of AI research.

Computational lens on probability

In the late 1980s, a probabilistic renaissance swept through mainstream AI research, fueled in part by pressures for performing sound inference about likelihoods of outcomes in applications of machine reasoning to such high-stakes domains as medicine. Attempts to mechanize probability for solving challenges with inference and learning led to new insights about probability and stimulated thinking about the role of related representations and inference strategies in human cognition. Perhaps most influentially, advances in AI led to the formulation of rich network-based representations, such as Bayesian networks, broadly referred to as probabilistic graphical models (PGMs) (*1*, *2*). Belief updating procedures were developed that use parallel and distributed computation to update constellations of random variables in the networks.

The study of PGMs has developed in numerous directions since these initial advances: efficient approximate inference methods; structure search over combinatorial spaces of network structures; hierarchical models for capturing shared structure across data sets; active learning to guide the collection of data; and probabilistic programming tools that can specify rich, context-sensitive models via compact, high-level programs. Such developments have put the notions of probabilistic inference and MEU decision-making at the heart of many contemporary AI approaches (*3*) and, together with ever-increasing computational power and data set availability, have been responsible for dramatic AI successes in recent years (such as IBM's Watson, Google's self-driving car, and Microsoft's automated assistant). These developments also raise new computational and theoretical challenges: How can we move from the classical view of a rational agent who maximizes expected utility over an exhaustively enumerable state-action space to a theory of the decisions faced by resource-bounded AI systems deployed in the real world (Fig. 1), which place severe demands on real-time computation over complex probabilistic models?

Rational decisions under bounded computational resources

Perception and decision-making incur computational costs. Such costs may be characterized in different ways, including losses that come with delayed action in time-critical settings, interference among multiple inferential components, and measures of effort invested. Work in AI has explored the value of deliberating at the metalevel about the nature and extent of perception and inference. Metalevel analyses have been aimed at endowing computational systems with the ability to make expected utility decisions about the ideal balance between effort or delay and the quality of actions taken in the world. The use of such rational metareasoning plays a central role in decision-theoretic models of bounded rationality (*10*–*14*).

Rational metareasoning has been explored in multiple problem areas, including guiding computation in probabilistic inference and decision-making

(*11*, *13*, *14*), controlling theorem proving (*15*), handling proactive inference in light of incoming streams of problems (*16*), guiding heuristic search (*13*, *17*), and optimizing sequences of action (*18*–*20*). Beyond real-time metareasoning, efforts have explored offline analysis to learn and optimize policies for guiding real-time metareasoning and for enhancing real-time inference via such methods as precomputing and caching portions of inference problems into fast-response reflexes (*21*).

The value of metalevel reflection in computational rationality is underscored by the complexity of probabilistic inference in Bayesian networks, which has been shown to be in the nondeterministic polynomial-time (NP)–hard complexity class (*22*). Such worst-case complexity highlights the importance of developing approximations that exploit the structure of real-world problems. A tapestry of approximate inferential methods have been developed, including procedures that use Monte Carlo simulation, bounding methods, and methods that decompose problems into simpler sets of subproblems (*1*). Some methods allow a system to trade off computation time for accuracy. For example, sampling procedures can tighten the bounds on probabilities of interest with additional computation time. Characterizations of the tradeoffs can be uncertain in themselves. Other approaches to approximation consider tradeoffs incurred with modulating the complexity of models, such as changing the size of models and the level of abstraction of evidence, actions, and outcomes considered (*11*, *21*).

A high-level view of the interplay between the value and cost of inference at different levels of precision is captured schematically in Fig. 2A. Here, the value of computing with additional precision on final actions and cost of delay for computation are measured in the same units of utility. A net value of action is derived as the difference between the expected value of action based on a current analysis and the cost of computation required to attain the level of analysis. In the situation portrayed, costs increase in a linear manner with a delay for additional computation, while the value of action increases with decreasing marginal returns. We see the attainment of an optimal stopping time, in which attempts to compute additional precision come at a net loss in the value of action. As portrayed in the figure, increasing the cost of computation would lead to an earlier ideal stopping time. In reality, we rarely have such a simple economics of the cost and benefits of computation. We are often uncertain about the costs and the expected value of continuing to compute and so must solve a more sophisticated analysis of the expected value of computation. A metalevel reasoner considers the current uncertainties, the time-critical losses with continuing computation, and the expected gains in precision of reasoning with additional computation.

As an example, consider a reasoning system that was implemented to study computational rationality for making inferences and providing recommendations for action in time-critical medical situations. The system needs to consider the losses incurred with increasing amounts of delay with action that stems from the time required for inference about the best decision to take in a setting. The expected value of the best decision may diminish as a system deliberates about a patient's symptoms and makes inferences about physiology. A trace of a reasoning session guided by rational metareasoning of a time-critical respiratory situation in emergency medicine is shown in Fig. 2B (*14*). An inference algorithm (named Bounded Conditioning) continues to tighten the upper and lower bounds on a critical variable representing the patient's physiology, using a Bayesian network to analyze evidence. The system is uncertain about the patient's state, and each state is associated with a different time criticality and ideal action. The system continues to deliberate at the metalevel about the value of continuing to further tighten the bounds. It monitors this value via computation of the expected value of computation. When the inferred expected value of computation goes to zero, the metalevel analysis directs the base-level system to stop and take the current best inferred base-level action possible.

Computational rationality in mind and brain

In parallel with developments in AI, the study of human intelligence has charted a similar progression toward computational rationality. Beginning in the 1950s, psychologists proposed that humans are "intuitive statisticians," using Bayesian decision theory to model intuitive choices under uncertainty (*23*). In the 1970s and 1980s, this hypothesis met with resistance from researchers who uncovered systematic fallacies in probabilistic reasoning and decision-making (*24*), leading some to adopt models based on informal heuristics and biases rather than normative principles of probability and utility theory (*25*). The broad success of probabilistic and decision-theoretic approaches in AI over the past two decades, however, has helped to return these ideas to the center of cognitive modeling (*5*, *26*–*28*). The development of methods for approximate Bayesian updating via distributed message passing over large networks of variables suggests that similar procedures might be used for large-scale probabilistic inference in the brain (*29*). At the same time, researchers studying human judgment and decision-making continue to uncover ways in which people's cognitive instincts appear far from the MEU ideals that economists and policymakers might have hoped for.

Computational rationality offers a framework for reconciling these contradictory pictures of human intelligence. If the brain is adapted to compute rationally with bounded resources, then "fallacies" may arise as a natural consequence of this optimization (*30*). For example, a generic strategy for approximating Bayesian inference is by sampling hypotheses, with the sample-based approximation converging to the true posterior as more hypotheses are

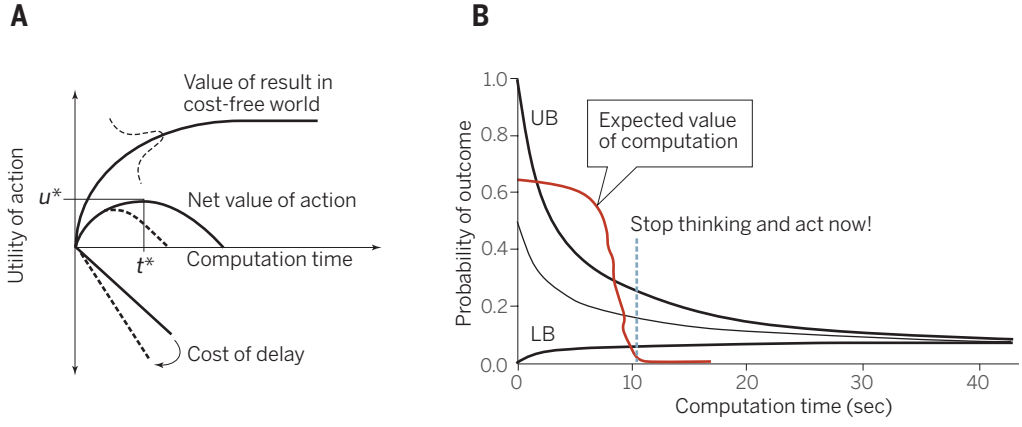

Fig. 2. Economics of thinking in computational rationality. (**A**) Systems must consider the expected value and cost of computation. Flexible computational procedures allow for decisions about ideal procedures and stopping times (t^*) in order to optimize the net value of action (u^*). In the general case, the cost-free value associated with obtaining a computed result at increasing degrees of precision and the cost of delay with computation are uncertain (indicated by the bell curve representing a probability distribution). Thus, the time at which further refinement of the inference should stop and action should be taken in the world are guided by computation of the expected value of computation. [Adapted from (*16*) with permission] (**B**) Trace of rational metareasoning in a time-critical medical setting. [Adapted from (*14*) with permission] A bounding algorithm continues to tighten the upper bound (UB) and lower bound (LB) on an important variable representing a patient's physiology. When a continually computed measure of the value of additional computation (red line) goes to zero, the base-level model is instructed to make a recommendation for immediate action.

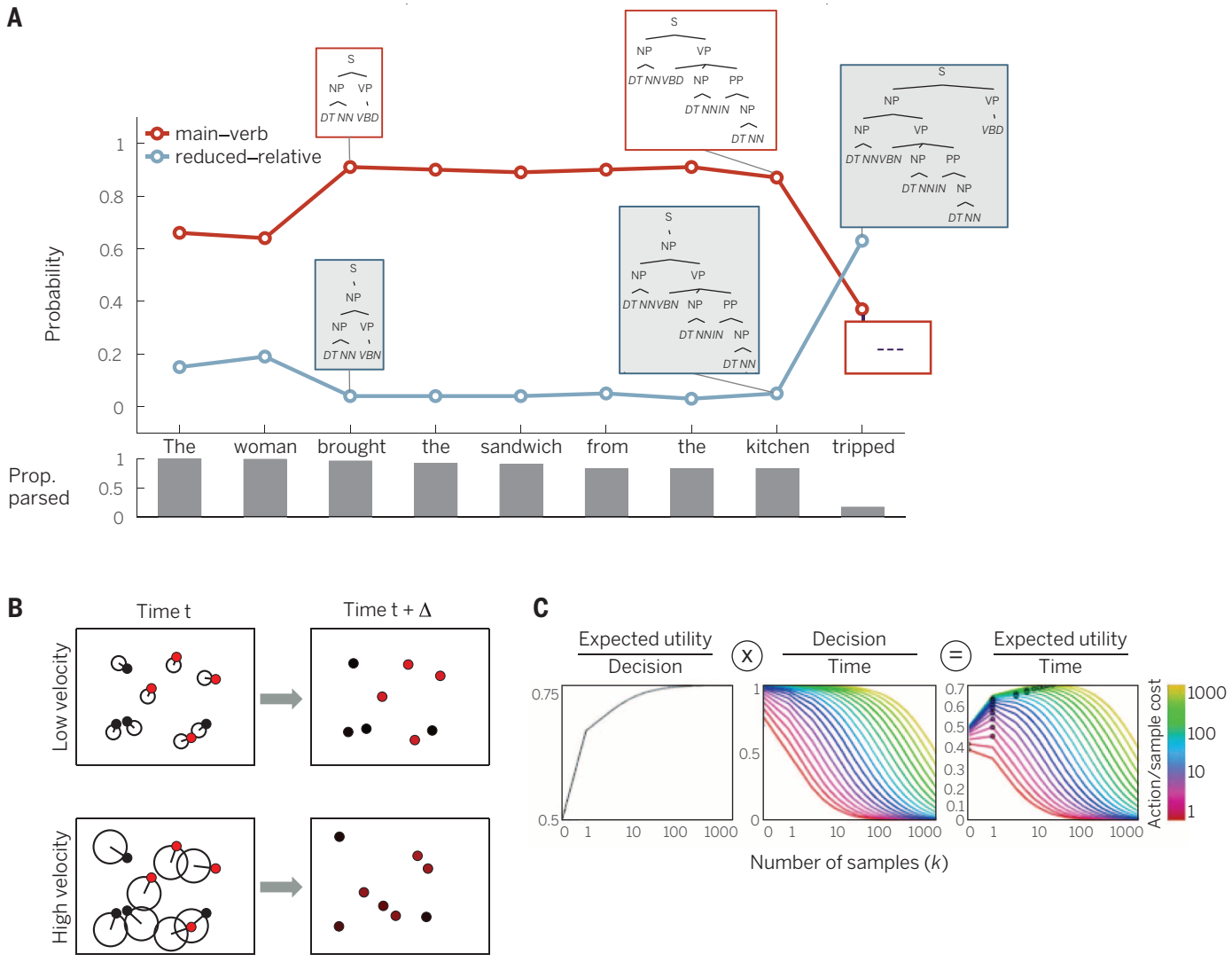

Fig. 3. Resource-constrained sampling in human cognition. (**A**) Incremental parsing of a garden-path sentence. [Adapted from (*36*)] (Top) Evolution of the posterior probability for two different syntactic parses (shown in boxes). The initially favored parse is disfavored by the end of the sentence. (Bottom) A resource-constrained probabilistic parser (based on a particle filter with few particles) may eliminate the initially unlikely parse and therefore fail to correctly parse the sentence by the end, as shown by the proportion of particle filters with 20 particles that successfully parse the sentence up to each word. (**B**) Sample-based inference for multiple-object tracking. In this task, subjects are asked to track a subset of dots over time, marked initially in red. Lines denote velocity, and circles denote uncertainty about spatial transitions. In the second frame, red shading indicates the strength of belief that a dot is in the tracked set (More red = higher confidence) after some time interval Δ, for a particle-filter object tracker. Uncertainty scales with velocity, explaining why high-velocity objects are harder to track (*32*). (**C**) For a sampling-based approximate Bayesian decision-maker, facing a sequence of binary choices, expected utility per decision, and number of decisions per unit time can be combined to compute the expected utility per unit time as a function of the number of posterior samples and action/sample cost ratios. Circles in the rightmost graph indicate the optimal number of samples at a particular action/sample cost ratio. For many decisions, the optimal choice tradeoff between accuracy and computation time suggests deciding after only one sample. [Reprinted from (*38*) with permission]

sampled (*1*). Evidence suggests that humans use this strategy across several domains, including causal reasoning (*31*), perception (*32*, *33*), and category learning (*34*). Sampling algorithms can also be implemented in biologically plausible neural circuits (*35*), providing a rational explanation for the intrinsic stochasticity of neurons.

We see a correspondence between the sampling algorithms humans appear to use and those used in state-of-the-art AI systems. For example, particle filters—sequential sampling algorithms for tracking multiple objects moving in a dynamic uncertain environment (*36*)—are at the heart of the Google self-driving car's picture of its surroundings (Fig. 1B) and also may describe how humans track multiple objects (Fig. 3B) (*32*). When only a small number of hypotheses are sampled, various biases emerge that are consistent with human behavior. For instance, "garden path" effects in sentence processing, in which humans perseverate on initially promising hypotheses that are disconfirmed by subsequent data, can be explained by particle filters for approximate online parsing in probabilistic grammars (Fig. 3A) (*37*). These biases may in fact be rational under the assumption that sampling is costly and most gains or losses are small, as in many everyday tasks; then, utility can be maximized by sampling as few as one or a few high-posterior probability hypotheses for each decision (Fig. 3C) (*38*).

This argument rests crucially on the assertion that the brain is equipped with metareasoning

mechanisms sensitive to the costs of cognition. Some such mechanisms may take the form of heuristic policies hardwired by evolutionary mechanisms; we call these "heuristic" because they would be metarational only for the range of situations that evolution has anticipated. There is also evidence that humans have more adaptive metareasoning mechanisms sensitive to the costs of cognition in online computation. In recent work with the "demand selection" task (*39*–*41*), participants are allowed to choose between two cognitive tasks that differ in cognitive demand and potential gains. Behavioral findings show that humans trade off reward and cognitive effort rationally according to a joint utility function (*40*). Brain imaging of the demand selection task has shown that activity in the lateral prefrontal cortex, a region implicated in the regulation of cognitive control, correlates with subjective reports of cognitive effort and individual differences in effort avoidance (*41*).

Several recent studies have provided support for rational metareasoning in human cognition when computational cost and reward tradeoffs are less obvious (*42*, *43*). As an example, humans have been found to consistently choose list-sorting strategies that rationally trade time and accuracy for a particular list type (*42*). This study joins earlier work that has demonstrated adaptive strategy selection in humans (*44*, *45*) but goes beyond them by explicitly modeling strategy selection using a measure of the value of computation. In another study (*46*), humans were found to differentially overestimate the frequency of highly stressful life events (such as lethal accidents and suicide). This "fallacy" can be viewed as rational under the assumption that only a small number of hypotheses can be sampled: Expected utility is maximized by a policy of utility-weighted sampling.

Computational tradeoffs in sequential decision-making

Computational rationality has played an important role in linking models of biological intelligence at the cognitive and neural levels in ways that can be seen most clearly in studies of sequential decision-making. Humans and other animals appear to make use of different kinds of systems for sequential decision-making: "model-based" systems that use a rich model of the environment to form plans, and a less complex "model-free" system that uses cached values to make decisions (*47*). Although both converge to the same behavior with enough experience, the two kinds of systems exhibit different tradeoffs in computational complexity and flexibility. Whereas model-based systems tend to be more flexible than the lighter-weight model-free systems (because they can quickly adapt to changes in environment structure), they rely on more expensive analyses (for example, tree-search or dynamic programming algorithms for computing values). In contrast, the model-free systems use inexpensive, but less flexible, look-up tables or function approximators. These efforts have conceptual links to efforts in AI that have sought to reduce effort and to speed up responses in real time by optimizing caches of inferences via offline precomputation (*21*).

Studies provide evidence that model-based and model-free systems are used in animal cognition and that they are supported by distinct regions of the prefrontal cortex (*48*) and striatum (*49*). Evidence further suggests that the brain achieves a balance between computational tradeoffs by using an adaptive arbitration between the two kinds of systems (*50*, *51*). One way to implement such an arbitration mechanism is to view the invocation of the model-based system as a meta-action whose value is estimated by the model-free system (*51*).

Early during learning to solve a task, when the model-free value estimates are relatively inaccurate, the benefits of using the model-based system outweigh its cognitive costs. Thus, moderately trained animals will be sensitive to changes in the causal structure of the environment (for example, the devaluation of a food reinforcer by pairing it with illness). After extensive training, the model-free values are sufficiently accurate to attain a superior cost-benefit tradeoff (*46*). This increasing reliance on the model-free system manifests behaviorally in the form of "habits"—computationally cheap but inflexible policies. For example, extensively trained animals will continue pursuing a policy that leads to previously devalued reinforcers (*52*).

The arbitration mechanism described above appears to adhere to the principles of computational rationality: The model-based system is invoked when deemed computationally advantageous through metareasoning (Fig. 4A). For

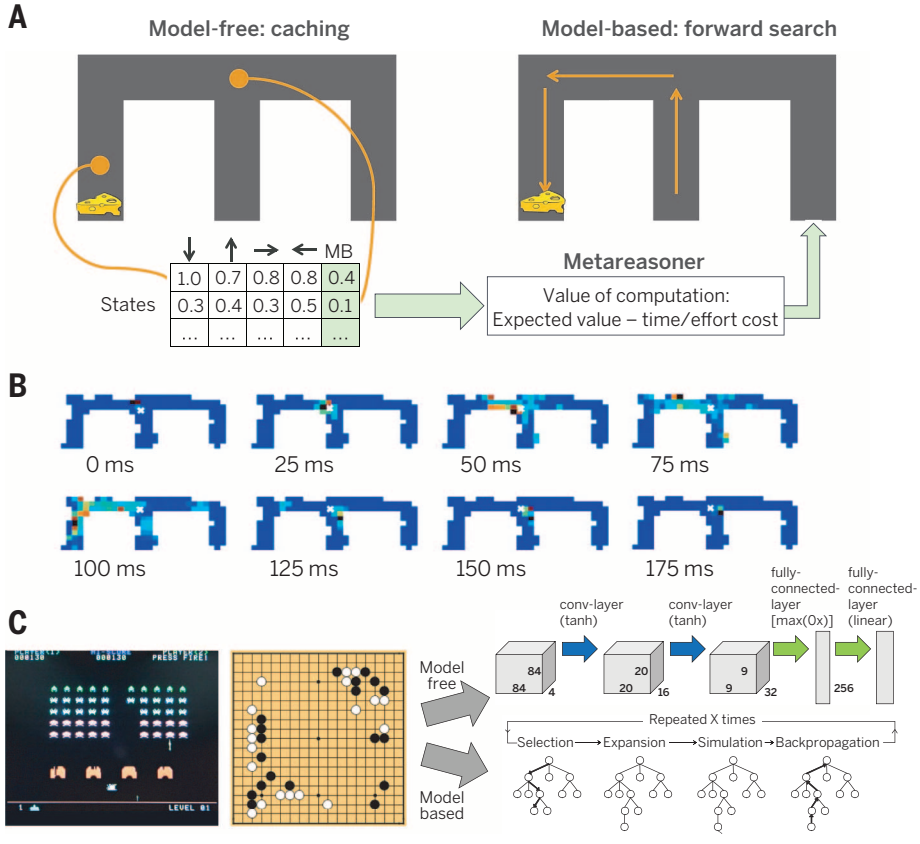

Fig. 4. Computational tradeoffs in use of different decision-making systems. (**A**) A fast but inflexible model-free system stores cached values in a look-up table but can also learn to invoke a slower but more flexible model-based system that uses forward search to construct an optimal plan. The cached value for invoking the model-based system (highlighted in green) is a simple form of metareasoning that weighs the expected value of forward search against time and effort costs. (**B**) Hippocampal place cell–firing patterns show the brain engaged in forward search at a choice point, sweeping ahead of the animal's current location. Each image shows the time-indexed representation intensity of locations in pseudocolor (red, high probability; blue, low probability). The representation intensity is computed by decoding spatial location from ensemble recordings in hippocampal area CA3. [Reprinted with permission from (*56*)] (**C**) Similar principles apply to more complex decision problems, such as Atari and Go (left). AI systems (right) use complex value function approximation architectures, such as deep convolutional nets (top right) [reprinted with permission from (*60*)], for model-free control, and sophisticated forward search strategies, such as Monte Carlo Tree Search (bottom right) [reprinted with permission from (*61*)], for model-based control.

example, reliance on the model-based system decreases when the availability of cognitive resources are transiently disrupted (*53*). Recent data show that the arbitration mechanism may be supported by the lateral prefrontal cortex (*54*), the same region involved in the registration of cognitive demand.

Finer-grained metareasoning may play a role within the richer model-based systems themselves. One way to approximate values is to adapt the sampling hypothesis to the sequential decision setting, stochastically exploring trajectories through the state space and using these sample paths to construct a Monte Carlo estimator. Recently, a class of sampling algorithms known as Monte Carlo Tree Search (MCTS) has gained considerable traction on complex problems by balancing exploration and exploitation to determine which trajectories to sample. MCTS has achieved state-of-the-art performance in computer Go as well as a number of other difficult sequential decision problems (*55*). A recent study analyzed MCTS within a computational rationality framework and showed how simulation decisions can be chosen to optimize the value of computation (*20*).

There is evidence that the brain might use an algorithm resembling MCTS to solve spatial navigation problems. In the hippocampus, "place cells" respond selectively when an animal is in a particular spatial location and are activated sequentially when an animal considers two different trajectories (Fig. 4B) (*56*). Pfeiffer and Foster (*57*) have shown that these sequences predict an animal's immediate behavior, even for new start and goal locations. It is unknown whether forward sampling observed in place cells balances exploration and exploitation as in MCTS, exploring spatial environments the way MCTS explores game trees, or whether they are sensitive to the value of computation. These are important standing questions in the computational neuroscience of decision-making.

At the same time, AI researchers are beginning to explore powerful interactions between model-based and model-free decision-making systems parallel to the hybrid approaches that computational cognitive neuroscientists have investigated (Fig. 4C). Model-free methods for game-playing based on deep neural networks can, with extensive training, match or exceed model-based MCTS approaches in the regimes that they have been trained on (*58*). Yet, combinations of MCTS and deep-network approaches beat either approach on its own (*59*) and may be a promising route to explain how human decision-making in complex sequential tasks can be so accurate and so fast yet still flexible to replan when circumstances change—the essence of acting intelligently in an uncertain world.

Looking forward

Computational rationality offers a potential unifying framework for the study of intelligence in minds, brains, and machines, based on three core ideas: that intelligent agents fundamentally seek to form beliefs and plan actions in support of maximizing expected utility; that ideal MEU calculations may be intractable for real-world problems, but can be effectively approximated by rational algorithms that maximize a more general expected utility incorporating the costs of computation; and that these algorithms can be rationally adapted to the organism's specific needs, either offline through engineering or evolutionary design, or online through metareasoning mechanisms for selecting the best approximation strategy in a given situation. We discussed case studies in which these ideas are being fruitfully applied across the disciplines of intelligence, but we admit that a genuine unifying theory remains mostly a promise for the future. We see great value in pursuing new studies that seek additional confirmation (or disconfirmation) of the roles of machinery for cost-sensitive computation in human cognition, and for enabling advances in AI. Although we cannot foresee precisely where this road leads, our best guess is that the pursuit itself is a good bet—and as far as we can see, the best bet that we have.

REFERENCES AND NOTES

1. D. Koller, N. Friedman, *Probabilistic Graphical Models: Principles and Techniques* (MIT Press, Cambridge, MA, 2009).
2. J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference* (Morgan Kaufmann Publishers, Los Altos, CA, 1988).
3. S. Russell, P. Norvig, *Artificial Intelligence: A Modern Approach* (Pearson, Upper Saddle River, NJ, 2009).
4. J. von Neumann, O. Morgenstern, *Theory of Games and Economic Behavior* (Princeton Univ. Press, Princeton, NJ, 1947).
5. J. B. Tenenbaum, C. Kemp, T. L. Griffiths, N. D. Goodman, *Science* **331**, 1279–1285 (2011).
6. A. M. Turing, *Proc. Lond. Math. Soc.* **2**, 230–265 (1936).
7. A. M. Turing, *Mind* **59**, 433–460 (1950).
8. J. von Neumann, *The Computer and the Brain* (Yale Univ. Press, New Haven, CT, 1958).
9. H. A. Simon, *Models of Man* (Wiley, New York, 1957).
10. I. J. Good, *J. R. Stat. Soc. B* **14**, 107–114 (1952).
11. E. Horvitz, in *Proceedings of the 3rd International Conference on Uncertainty in Artificial Intelligence* (Mountain View, CA, July 1987), pp. 429–444 (1987).
12. S. Russell, E. Wefald, *Artif. Intell.* **49**, 361–395 (1991).
13. E. Horvitz, G. Cooper, D. Heckerman, in *Proceedings of IJCAI*, January 1989, pp. 1121–1127 (1989).
14. E. Horvitz, G. Rutledge, in *Proceedings of the 7th International Conference on Uncertainty in Artificial Intelligence* (Morgan Kaufmann Publishers, San Francisco, 1991), pp. 151–158.
15. E. Horvitz, Y. Ruan, G. Gomes, H. Kautz, B. Selman, D. M. Chickering, in *Proceedings of 17th Conference on Uncertainty in Artificial Intelligence* (Morgan Kaufmann Publishers, San Francisco, 2001), pp. 235–244.
16. E. Horvitz, *Artif. Intell.* **126**, 159–196 (2001).
17. E. Burns, W. Ruml, M. B. Do, *J. Artif. Intell. Res.* **47**, 697–740 (2013).
18. C. H. Lin, A. Kolobov, A. Kamar, E. Horvitz, Metareasoning for planning under uncertainty. In *Proceedings of IJCAI* (2015).
19. T. Dean, L. P. Kaelbling, J. Kirman, A. Nicholson, *Artif. Intell.* **76**, 35–74 (1995).
20. N. Hay, S. Russell, D. Tolpin, S. Shimony, in *Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence* (2012), pp. 346–355.
21. D. Heckerman, J. S. Breese, E. Horvitz, in *Proceedings of the 5th Conference on Uncertainty in Artifical Intelligence*, July 1989 (1989), pp. 162–173.
22. G. Cooper, *Artif. Intell.* **42**, 393–405 (1990).
23. C. R. Peterson, L. R. Beach, *Psychol. Bull.* **68**, 29–46 (1967).
24. A. Tversky, D. Kahneman, *Science* **185**, 1124–1131 (1974).
25. G. Gigerenzer, *Rationality for Mortals: How People Cope with Uncertainty* (Oxford Univ. Press, Oxford, 2008).
26. J. R. Anderson, *The Adaptive Character of Thought* (Lawrence Erlbaum, Hillsdale, NJ, 1990).
27. M. Oaksford, N. Chater, *Bayesian Rationality* (Oxford Univ. Press, Oxford, 2007).
28. T. L. Griffiths, J. B. Tenenbaum, *Cognit. Psychol.* **51**, 334–384 (2005).
29. K. Doya, S. Ishii, A. Pouget, R. P. N. Rao, Eds. *The Bayesian Brain: Probabilistic Approaches to Neural Coding* (MIT Press, Cambridge, MA, 2007).
30. T. L. Griffiths, F. Lieder, N. D. Goodman, *Top. Cogn. Sci.* **7**, 217–229 (2015).
31. S. Denison, E. Bonawitz, A. Gopnik, T. L. Griffiths, *Cognition* **126**, 285–300 (2013).
32. E. Vul, M. Frank, G. Alvarez, J. B. Tenenbaum, *Adv. Neural Inf. Process. Syst.* **29**, 1955–1963 (2009).
33. S. J. Gershman, E. Vul, J. B. Tenenbaum, *Neural Comput.* **24**, 1–24 (2012).
34. A. N. Sanborn, T. L. Griffiths, D. J. Navarro, *Psychol. Rev.* **117**, 1144–1167 (2010).
35. L. Buesing, J. Bill, B. Nessler, W. Maass, *PLOS Comput. Biol.* **7**, e1002211 (2011).
36. M. Isard, A. Blake, *Int. J. Comput. Vis.* **29**, 5–28 (1998).
37. R. Levy, F. Reali, T. L. Griffiths, *Adv. Neural Inf. Process. Syst.* **21**, 937–944 (2009).
38. E. Vul, N. Goodman, T. L. Griffiths, J. B. Tenenbaum, *Cogn. Sci.* **38**, 599–637 (2014).
39. W. Kool, J. T. McGuire, Z. B. Rosen, M. M. Botvinick, *J. Exp. Psychol. Gen.* **139**, 665–682 (2010).
40. W. Kool, M. Botvinick, *J. Exp. Psychol. Gen.* **143**, 131–141 (2014).
41. J. T. McGuire, M. M. Botvinick, *Proc. Natl. Acad. Sci. U.S.A.* **107**, 7922–7926 (2010).
42. F. Lieder *et al*., *Adv. Neural Inf. Process. Syst.* **27**, 2870–2878 (2014).
43. R. L. Lewis, A. Howes, S. Singh, *Top. Cogn. Sci.* **6**, 279–311 (2014).
44. J. W. Payne, J. R. Bettman, E. J. Johnson, *J. Exp. Psychol. Learn. Mem. Cogn.* **14**, 534–552 (1988).
45. J. Rieskamp, P. E. Otto, *J. Exp. Psychol. Gen.* **135**, 207–236 (2006).
46. F. Lieder, M. Hsu, T. L. Griffiths, in *Proc. 36th Ann. Conf. Cognitive Science Society* (Austin, TX, 2014).
47. N. D. Daw, Y. Niv, P. Dayan, *Nat. Neurosci.* **8**, 1704–1711 (2005).
48. S. Killcross, E. Coutureau, *Cereb. Cortex* **13**, 400–408 (2003).
49. H. H. Yin, B. J. Knowlton, B. W. Balleine, *Eur. J. Neurosci.* **19**, 181–189 (2004).
50. N. D. Daw, S. J. Gershman, B. Seymour, P. Dayan, R. J. Dolan, *Neuron* **69**, 1204–1215 (2011).
51. M. Keramati, A. Dezfouli, P. Piray, *PLOS Comput. Biol.* **7**, e1002055 (2011).
52. A. Dickinson, *Philos. Trans. R. Soc. London B Biol. Sci.* **308**, 67–78 (1985).
53. A. R. Otto, S. J. Gershman, A. B. Markman, N. D. Daw, *Psychol. Sci.* **24**, 751–761 (2013).
54. S. W. Lee, S. Shimojo, J. P. O'Doherty, *Neuron* **81**, 687–699 (2014).
55. S. Gelly *et al*., *Commun. ACM* **55**, 106–113 (2012).
56. A. Johnson, A. D. Redish, *J. Neurosci.* **27**, 12176–12189 (2007).
57. B. E. Pfeiffer, D. J. Foster, *Nature* **497**, 74–79 (2013).
58. V. Mnih *et al*., *Nature* **518**, 529–533 (2015).
59. C. J. Maddison, A. Huang, I. Sutskever, D. Silver, http://arxiv.org/abs/1412.6564 (2014).
60. X. Guo, S. Singh, H. Lee, R. Lewis, X. Wang, *Adv. Neural Inf. Process. Syst.* **27**, 3338–3346 (2014).
61. G. M. J.-B. Chaslot, S. Bakkes, I. Szita, P. Spronck, in *Proc. Artif. Intell. Interact. Digit. Entertain. Conf.* (Stanford, CA, 2008), pp. 216–217.

ACKNOWLEDGMENTS

We are grateful to A. Gershman and the three referees for helpful comments. This research was partly supported by the Center for Brains, Minds and Machines (CBMM), funded by National Science Foundation Science and Technology Center award CCF-1231216.

10.1126/science.aac6076

Computational rationality: A converging paradigm for intelligence in brains, minds, and machines

Samuel J. Gershman *et al.*

Science **349**, 273 (2015);
DOI: 10.1126/science.aac6076