ARTICLE IN PRESS

CellPress

Opinion

Understanding Human Intelligence through Human Limitations

Thomas L. Griffiths[1,2,*]

Recent progress in artificial intelligence provides the opportunity to ask the question of what is unique about human intelligence, but with a new comparison class. I argue that we can understand human intelligence, and the ways in which it may differ from artificial intelligence, by considering the characteristics of the kind of computational problems that human minds have to solve. I claim that these problems acquire their structure from three fundamental limitations that apply to human beings: limited time, limited computation, and limited communication. From these limitations we can derive many of the properties we associate with human intelligence, such as rapid learning, the ability to break down problems into parts, and the capacity for cumulative cultural evolution.

Different Computational Problems, Different Kinds of Intelligence

As machines begin to outperform humans on an increasing number of tasks, it is natural to ask what is unique about human intelligence. Historically, this has been a question that is asked when comparing humans with other animals. The classical answer (from Aristotle, via the Scholastics) is to view humans as 'rational animals', animals that think [1]. More modern analyses of human uniqueness emphasize the 'cognitive niche' that humans fill, able to use their minds to outsmart the biological defenses of their competitors [2], or contrast this with the 'cultural niche' of being able to accumulate knowledge across individuals and generations in a way that makes it possible to live in an unusually diverse range of environments [3–5]. Asking the same question of what makes humans unique, but changing the contrast class to include intelligent machines, yields a very different kind of answer.

In this article I argue that even as we develop potentially superhuman machines, there is going to be a flavor of intelligence that remains uniquely human. To understand the nature of human intelligence, we need to understand the kinds of computational problems that human minds have to solve. David Marr [6], Roger Shepard [7], and John Anderson [8] all converged on a productive strategy for making sense of specific aspects of human cognition: attempting to understand the abstract computational problem underlying that aspect of cognition and using its ideal solution to gain insight into why people do the things they do. For Marr this was the 'computational level' of analysis, for Shepard a way to try to identify universal laws of cognition, and for Anderson a component of 'rational analysis'. This approach has since been pursued in order to gain insight into a wide range of problems, including reasoning [9], generalization [10], categorization [11,12], and causal learning [13]. However, these applications all focus on particular computational problems, rather than asking what the characteristics are of the kinds of computational problems that human minds need to solve in general.

Following the same approach with this wider lens, we can ask what it is that is common across all of these problems that seems characteristic of all of the computational problems that humans

Highlights

Humans are limited in time, computation, and communication, defining a set of computational problems that human intelligence has to solve.

Considering the structure of these computational problems can help us to understand why human minds have some of the characteristics that they do.

The solutions to these problems involve mathematical formalisms such as Bayesian inference and meta-learning, rational meta-reasoning, and distributed algorithms, which may be particularly relevant to cognitive science.

[1]Department of Psychology, Princeton University, Princeton, NJ, USA
[2]Department of Computer Science, Princeton University, Princeton, NJ, USA

*Correspondence:
tomg@princeton.edu (T.L. Griffiths).

encounter. I suggest that the set of human computational problems all share three important characteristics:

1. Humans have a limited amount of time. Nature may only provide limited opportunities to learn behaviors relevant to survival and the length of human lives imposes an upper bound on the amount of available data.
2. Humans have access to a limited amount of computation. Each human being has a single brain with fixed computational capacity.
3. Humans minds have limited communication. Human beings have no way to directly transfer the contents of their brain to one another.

The constraints imposed by these characteristics cascade: limited time magnifies the effect of limited computation and limited communication makes it harder to draw upon more computation.

It is worth noting that these characteristics are not a necessary constraint on all intelligent systems. While these same constraints apply to cognition in all animals, not just humans, it is worth noting that this is not the case for all intelligent systems. Recent breakthroughs in artificial intelligence (AI) have been driven by an exponential increase in the amount of computation being used to solve problems (http://blog.openai.com/ai-and-compute/) and it has become common for these systems to be provided experience that is equivalent of many human lifetimes. The results of learning in one system can be copied directly to another, making it possible to train a single system through experiences acquired in parallel. Each of these three characteristics of human problems thus reflects a human limitation (Figure 1).

The recent successes of AI research have initiated a discussion of how current AI systems differ from human intelligence. The key differences that have been highlighted include the ability to learn from small amounts of data and the use of structured representations [14]. Rather than reiterating this discussion, my goal is to shift the emphasis from how to why: why do these properties of human intelligence exist? Identifying these properties as a consequence of the nature of the computational problems that human minds have to solve helps to clarify when
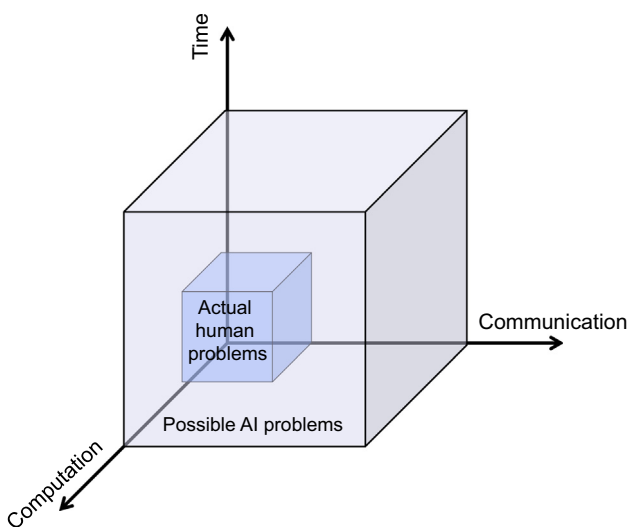

Figure 1. Schematic Depiction of Computational Problems Potentially Faced by Humans and Machines. Human minds solve a set of problems that are a specific subset of those potentially faced by artificial intelligence (AI) systems, resulting from limited time, computation, and communication.

Trends in Cognitive Sciences

we might want AI systems to have the same properties: when they face the same constraints as humans.

Understanding human intelligence via human limitations is also helpful for identifying the formal tools that will be most relevant to cognitive science. In the spirit of Marr, Shepard, and Anderson, we can ask what the ideal solutions to human computational problems look like. Each of these limitations imposes its own structure on problems, with a corresponding set of mathematical tools required to solve them. Limited time means being able to make inferences from limited data, so formalisms such as Bayesian inference that allow us to characterize the inductive biases of learners are valuable. Limited computation requires that computation be used efficiently, making a connection to the literature on rational meta-reasoning in AI. Limited communication implies that solving problems that go beyond the computational or temporal limits of those individuals requires developing mechanisms for cumulative cultural evolution, for which the relevant formalisms come from parallel and distributed algorithms.

If the goal of cognitive science is to understand human intelligence, and the intelligence of entities that operate under similar constraints, it will require making use of each of these tools. While dealing with these constraints is a necessary component of understanding human cognition, it is also a desirable characteristic for AI systems. Understanding how human minds navigate these constraints is thus potentially also relevant to developing systems capable of learning quickly, thinking efficiently, and communicating about the results (see Box 1 for further discussion).

In the remainder of the article, I discuss each human limitation and their cascading consequences in further detail.

Limitation 1: Time

Limited time means having to learn from limited amounts of data. There are at least three timescales at which human learning has to operate, reflecting different kinds of constraints: timescales

---

Box 1. When Do Machines Need to Be Like Humans?

As should be clear from Figure 1 (in main text), I am not arguing that the computational problems faced by humans and those potentially faced by machines are completely disjoint; rather, that human problems are a specific subset of potential machine problems that result from the intersection of three limitations. A natural question is then what kind of machines encounter that same subset. The main text highlights some contexts where learning from limited data may be necessary: certain scientific applications, or settings that require interacting with people. Likewise, systems that need to make timely decisions might potentially draw inspiration from people. Finally, systems that need to coordinate with limited communication may need something like language or teaching. But when do machines need to confront all three human limitations simultaneously?

Viewed from the perspective of an engineer, humans are a very specific kind of system: a self-contained autonomous system capable of real-time performance of a wide range of tasks. We carry our own power supply, operate without external control, do so in time-critical settings, and are able to navigate, find energy sources, make inferences about our environment, and execute precision movements. Each of these aspects of humans, being autonomous, real-time, and general-purpose, is an engineering challenge and it is usually possible to design a machine that performs a task without having to overcome these challenges. As a consequence, most machines do not face the particular configuration of computational problems that humans do.

This does not mean that there are no use cases for machines that align well with those of humans. One example is space exploration, where we might seek to create autonomous probes that are able to gather data to make intelligent inferences and actions while preserving limited power, with low bandwidth and high latency for communications. Such a system faces many of the same limitations as humans and might be expected to have similar characteristics. Indeed, each human birth launches a probe into a universe full of potential new discoveries.

imposed by survival, timescales imposed by the explore/exploit trade-off, and timescales imposed by the absolute limits of the human lifespan.

The first timescale appears in cases where the limits on the time available to learn are imposed by the need to survive. Humans who required thousands of examples in order to be able to reliably identify a tiger would not have survived in environments where hungry black and orange beasts abound. Being able to quickly learn to identify different animals, plants, and other natural categories has been, at least in an evolutionary sense, an important part of the computational problems that humans face. In this context, the capacity to learn from limited data is comparable with the ability of baby gazelles to run shortly after birth, a consequence of the limited time available to build up the requisite skills for survival.

The second timescale that affects human learning is an indirect consequence of the length of human lives. Many of the problems that humans have to solve require navigating the explore/exploit trade-off [15]. Faced with situations we are likely to encounter again in the future, we have to decide whether to take actions that may provide useful information (explore) or instead take the action we know is most likely to result in a good outcome (exploit). A key variable in the decision about whether to explore or exploit is how many more times we will make similar decisions: the value of information increases the more opportunities we will have to use it [16]. We should thus expect people to switch from exploration to exploitation with age, an idea that has been proposed as a way of understanding human childhood [17].

Crucially, the point at which we switch from exploration (learning about our world) to exploitation (using that knowledge) will depend on the total length of our lifespan. The longer we live, the greater the value of information provided by exploration, and the longer it makes sense to explore. So, the length of human lifespans imposes a natural scale on the length of many of the learning problems we face: even when our environment affords us the luxury of time to learn, there is a natural limit on the amount of data we will have access to in order to learn the skills that are essential to our lives as adults.

Finally, the third timescale is an even more direct consequence of our limited lifespan. Even when we have the luxury of spending our entire lives learning, the maximum amount of data we can learn from is that which can be accumulated in a single lifetime. Whatever we learn must be acquired in (significantly) less than a million hours of real-time experience.

It is by no means necessary that machines operate under the same constraints. Learning from less data may save money, but the amount of data available to machine learning systems is fundamentally a function of how much data has being collected in that domain and the goals of the person building the system. The AlphaGo system that beat one of the best human players in history had the benefit of multiple human lifetimes of simulated play [18]. The most successful neural network language model, GPT-3, was recently trained on over 400 billion tokens of text [19]. Assuming speech at a rate of 150 words per minute, that is how many words would be produced speaking continuously since the reign of the earliest Egyptian pharaohs over 5000 years ago. However, there are settings where being able to learn from small amounts of data is important. One example is in sciences where data is very limited, such as particle physics or cosmology [20]. Another example is building software systems that have to adapt to individual users, where new users may decide whether or not to adopt a system based on how quickly it learns from their behavior [21]. More generally, in interactions with humans, where expectations about the speed of learning are set by other humans, there is pressure to quickly learn new tasks or concepts from examples [22]. In these settings, we might expect machine learning systems to be more similar to human learning (Box 1).

The mathematical formalisms tools most useful for making the best use of limited time are those that help us understand the inductive biases of learners. In machine learning, inductive bias is defined as anything other than the data that influences the conclusion the system reaches [23]. A reinforcement learning system that establishes how to move the joints of a robot through trial and error is relying on data to get there, while a newborn gazelle is making strong use of inductive bias. To be able to learn from less data, a system needs inductive biases that point towards the right solutions.

Abstractly, humans and machines both face the challenge of acquiring enough knowledge about the world around them in order to produce intelligent action. What differs between human learning and machine learning is the variables that can be manipulated in order to achieve that goal. For machines, data is typically flexible. How much data is available to learn from depends mostly on the budget for solving the problem, and recent research in deep learning has succeeded in part by massively increasing the amount of data available to learners (http://blog.openai.com/ai-and-compute/). Getting more data is often easier, and more successful, than trying to engineer good inductive biases, something that has been termed the 'bitter lesson' [24] (http://www.incompleteideas.net/IncIdeas/BitterLesson.html). For humans, data is typically fixed. (That said, human minds work hard to make the best use of the limited time available to them, actively selecting the data that they learn from whenever they can [25].) We have to learn from the data available to us and the only way to do so is to establish good inductive biases. The project of the cognitive scientist, then, becomes one of trying to reverse-engineer these inductive biases.

Bayesian inference provides one tool for exploring inductive biases, using different prior distributions as an explicit means of characterizing the predispositions of learners [26]. Bayesian models of cognition can thus be used to evaluate hypotheses about human inductive biases. Machine learning systems that learn from massive amounts of data can also be informative about human inductive biases. One way of understanding how people are able to learn from limited data is by viewing evolution as providing the equivalent of 'pretraining' on a larger data set. Machine learning methods are thus solving in a single learning step a problem that for humans has been broken into separate parts, one solved by evolution and the other by learning. Recent work in machine learning on meta-learning, which is a framework for learning an appropriate inductive bias through experience, provides a way to see how these parts could be separated again (Box 2).

Limitation 2: Computation

Limited time is a concern not just for the amount of data the learner gets exposed to, but for the amount of computation that can be expended in solving a problem. When faced with limited time, computer scientists will often increase the amount of computation devoted to a problem. However, humans do not have access to that solution. Equipped with a brain with fixed processing power, we have to use these limited resources to solve all of the problems that we encounter.

This limitation sets up a new kind of problem: how do we make the best use of our limited computational resources? A key attribute of human intelligence is being able to break problems into parts that can individually be solved more easily, or that make it possible to reuse partial solutions discovered through previous experience. These methods for making computational problems more tractable are such a ubiquitous part of human intelligence that they seem to be an obligatory component of intelligence more generally. One example of this is forming subgoals. The early AI literature, inspired by human problem-solving, put a significant emphasis on reducing tasks to a series of subgoals [27]. However, forming subgoals is not a necessary part of intelligence, it is a consequence of having limited computation. With a sufficiently large amount of computation, there is no

Box 2. Meta-learning, Bayes, and Evolution

An active topic of research in machine learning is 'meta-learning', or learning to learn [42]. Rather than optimizing an agent to perform a single task, meta-learning focuses on optimizing agents to quickly learn to perform a range of tasks that have some shared structure (Figure I). Effectively, this approach aims to learn an inductive bias that is appropriate for the distribution of tasks that the agent encounters. This can be done in a variety of ways, including using a recurrent neural network to learn regularities in strategies across tasks [43,44] or optimizing performance across tasks using the same gradient descent algorithms used for learning within tasks [45–47]. Learning an inductive bias in this way is similar to learning a prior distribution for Bayesian inference, a relationship that can be made exact for a least one popular approach to meta-learning [48].

Meta-learning algorithms typically consist of an 'inner loop', in which a set of agents learns to perform different tasks, and an 'outer loop', in which the inductive biases of those agents are tuned to improve their inner-loop performance. For example, one popular algorithm, Model-Agnostic Meta-Learning [46], minimizes the following objective function

$$\mathcal{L}(\theta) = \sum_t \mathcal{L}_t(\varphi_t) \qquad \varphi_t = \theta - \alpha\nabla_\theta\mathcal{L}_t(\theta) \tag{I}$$

where θ is the initial weights of an artificial neural network, $\mathcal{L}_t(\varphi_t)$ is the loss (error) of a neural network with weights φ_t on task t, and $\varphi_t = \theta - \alpha\nabla_\theta\mathcal{L}_t(\theta)$ are the weights obtained after one step of gradient descent initialized with θ with learning rate α. Calculating φ_t is the 'inner loop' of learning and finding θ is the 'outer loop' that optimizes the inductive bias, here in the form of the initialization θ.

This decomposition of agent performance into two separate optimization problems resembles the decomposition of human behavior into ontogenetic and phylogenetic processes, learning and evolution. The outer-loop optimization is the problem that evolutionary processes address, tuning the characteristics of organisms based on the environments that those organisms will encounter. This tuning only makes sense if the amount of data available in the inner loop is fixed, if learners have limited time, as depicted in Figure I. Indeed, the primary application for meta-learning algorithms is trying to reduce the amount of data learners require to perform a new task.

Viewed in this way, meta-learning may also offer a unique tool for exploring human inductive biases. By providing a way to impart learners with inductive biases that are sufficient to solve a particular set of learning problems, these algorithms potentially allow us to explore what kinds of inductive biases are necessary to reach human-level solutions from human-limited data (for a preliminary example of this approach, see [49]).
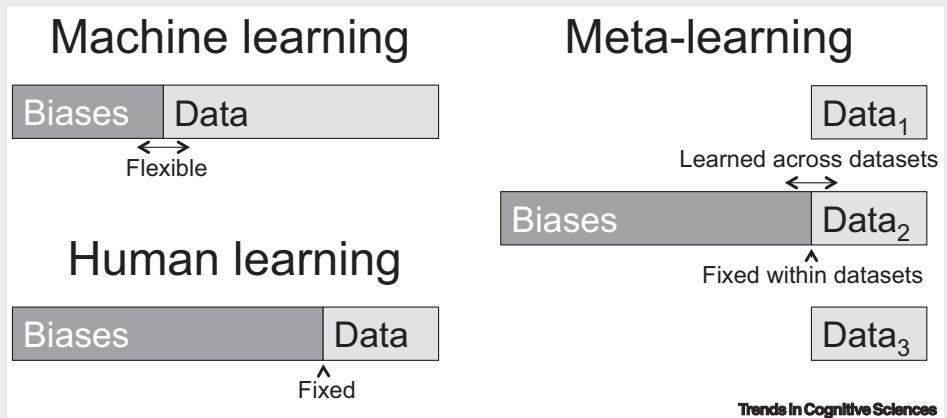

Figure I. Different Regimes for Learning. In machine learning, there is typically some flexibility about the amount of data available to the learner. This contrasts with human learning, where the total amount of data available is limited. Meta-learning provides a way to bridge these paradigms, optimizing inductive bias across a large number of distinct datasets, each limited in size.

need to have subgoals: the problem can be solved by simply planning all the way to the final goal. Go experts have commented that new AI systems sometimes produce play that seems alien, precisely because it was hard to identify goals that motivated particular actions [20] (http://www.theatlantic.com/technology/archive/2017/10/alphago-zero-the-ai-that-taught-itself-go-543450/). This makes

perfect sense, since the actions taken by these systems are justified by the fact that they are most likely to yield a small expected advantage many steps in the future rather than because they satisfy some specific subgoal.

Another example where human intelligence looks very different from machine intelligence is in solving the Rubik's cube. Thanks to some careful analysis and a significant amount of computation, the Rubik's cube is a solved problem: the shortest path from any configuration to an unscrambled cube has been identified, taking no more than 20 moves [28]. However, the solution does not have a huge amount of underlying structure, those shortest paths are stored in a gigantic lookup table. Contrast this with the solutions used by human solvers. A variety of methods for solving the cube exist, but those used by the fastest human solvers require around 50 moves. These solutions require memorizing a few dozen to a few hundred 'algorithms' that specify transformations to be used at particular points in the process. Methods also have intermediate subgoals, such as first solving an entire side. While human solutions may take longer to execute, they are far more efficient to describe and implement, being able to be summarized in a short booklet.

As these examples make clear, human intelligence reflects an expertise in finding solutions to problems that can be implemented with the limited amount of computation contained inside a single human brain. The search for solutions to problems that trade-off speed or accuracy with the cost of computation can be formalized using an idea originally introduced in the AI literature: rational meta-reasoning [29,30]. If reasoning is solving a problem, meta-reasoning is solving the problem of how to solve the problem, in this case, working out how to solve problems using limited amounts of computation. Rational meta-reasoning explores what optimal solutions to these meta-reasoning problems look like (Box 3). In AI, this provides a way to characterize

---

Box 3. Rational Meta-reasoning and the Human Paradox

Anybody who talks to both psychologists and computer scientists can come away with a paradoxical view of human nature: humans disappoint psychologists as error-prone decision-makers who often act irrationally, but inspire computer scientists with their ability to efficiently solve a remarkably wide range of problems. One way to resolve this paradox is to recognize that both of these attributes stem from having limited computational resources. Those limits mean that humans have to follow heuristic strategies that result in errors [50,51]. However, they also mean that humans have had to develop a finely-honed ability to deploy the limited resources that we have as effectively as possible.

Rational meta-reasoning provides a framework for answering the question of how a rational agent should deploy limited resources [29,30]. The classic characterization of rationality in terms of maximizing expected utility tells us how we should take actions in the external world. Rational meta-reasoning focuses on our inner world, telling us what computations we should execute in order to obtain the information that we need to decide how to act, trading off the costs of the errors we might make with the costs of executing those computations. As such, it provides an excellent framework in which to ask the questions of cognitive psychology, which are typically more about our internal processes than our external actions (for a review see [31]).

More formally, the classical approach to rationality emphasizes maximizing expected utility. The goal of the agent is thus to maximize $E_{p(x|a)}[u(x)]$ across actions a, where $p(x \mid a)$ is the probability of outcome x under action a and $u(x)$ is its utility. This is fundamentally a 'behaviorist' notion of rationality, insofar as it focuses on the action the agent takes and not how the agent decides that action is appropriate. The rational meta-reasoner instead maximizes $\max_a E_{p_c(x|a)}[u(x)] - cost(c)$ over computations c, where $p_c(x \mid a)$ is the estimate of the probability of x yielded by those computations and $cost(c)$ is their cost. Maximizing over computations focuses on the cognitive processes behind a decision.

Rational meta-reasoning also provides a framework for thinking about how to create AI systems that are capable of the same flexibility as humans. Building artificial general intelligence requires training AI systems on a range of tasks that have the same generality as those executed by humans. Part of the way that humans achieve that generality is by being able to reuse knowledge or subroutines that they have acquired when performing one task in service of performing another. This can be reduced to a decision about which of a learned set computations to execute when solving a problem: rational meta-reasoning [52].

effective strategies for systems that need to reduce responses under time constraints, relevant to applications such as driving or healthcare. However, rational meta-reasoning may be even more valuable for understanding the solutions to computational problems that characterize human intelligence [31,32].

Limitation 3: Communication

Limited time and computation need not be major constraints if it is possible to solve problems in serial or parallel. For example, being eaten by a tiger may not be a bad outcome for an intelligent system if it is able to transfer the data it has acquired to another instance of that system, effectively increasing the amount of experience each system gets over time. Likewise, limited computation is not necessarily an issue if a problem can be broken up into components that are distributed across many processes and then aggregated into a solution. The problem is that both serial and parallel computation require information to be shared between processors and human brains have no direct interfaces for copying information in this way: the experiences of each individual are encapsulated within their brain.

Limited communication results in a third key component of human intelligence: mechanisms that support cumulative cultural evolution. When humans need to solve a problem that transcends the limits of individual lifespans or computational power, we develop cultural institutions that allow us to accumulate the results of individual learning and action [3]. Commercial companies and scientific journals are examples of institutions that allow us to transcend individual limits. Joining together in groups that support division of labor allows some individuals to specialize and thus learn from more data. Using pictures or writing allows us to communicate information beyond a single lifespan. These same principles are behind the formation of modern corporations and scientific journals.

At a more basic level, key components of human intelligence can be seen as a consequence of limited communication. Teaching is one example: if individual learners have a limited amount of time in which to learn, then making sure that they are provided with the very best data is a way to expand their capacity. Language is another, being a system that allows us to come closer to copying information across brains. A human Go player potentially has access to more than a single lifespan of play through the documentation and discussion of what has been learned about the game by previous generations. Even under a richer view of language as a mechanism for coordination and negotiation [33], an implicit premise is that the states of the agents coordinating and negotiating and are not transparent to one another. The compositional structure of language has itself been suggested to be a consequence of having limited time in which to transfer knowledge from one brain to another [34,35].

While there has been significant interest in topics such as teaching and the emergence of language in AI research (e.g., [36,37]), these need only be a property of AI systems if encapsulated agents are necessary. The capacity to directly transfer data or states of computation from one machine to another, or to have a single intelligence control multiple bodies [38], means that it is perfectly reasonable to have AI systems that are not subject to this kind of constraint. As a consequence, such systems might be expected to omit these aspects of human intelligence.

The formal tools required to understand cumulative cultural evolution come from yet another part of computer science: the study of distributed computation. Cognitive scientists are used to analyzing computational problems expressed at the level of individuals, but we can apply the same lens at the level of groups or societies. Thinking about a group of people jointly solving a computational problem allows us to ask what algorithms might be effective for solving that problem,

Box 4. Algorithms for Cumulative Cultural Evolution

Broadly speaking, humans face two kinds of problems that require resources that go beyond the limitations of individual brains: inference and optimization. Inference, trying to discover the truth about an aspect of the world, requires aggregating data that go beyond the experience of one individual. An example is science: no single human has personal experience of the entire history of experimental work in a given field, so we need to develop mechanisms that allow us to aggregate that knowledge. Optimization, trying to develop better solutions to a problem, involves computational resources that are greater than a single human mind. An example is technological development: each technology accretes innovations made by many different people.

Inference and optimization are also at the heart of many of the problems we want our computers to solve and, as a consequence, there is an extensive literature on algorithms for solving these problems. Some of these algorithms are designed for exactly the situation that humans face: many processors, each of which has limited capacity, requiring parallelization of a challenging computation. These parallel algorithms may provide a rich source of inspiration for understanding human intelligence and in particular how we are able to accumulate knowledge across individuals and generations.

One concrete example is the particle filter, an algorithm that parallelizes elements of Bayesian inference. One of the properties of Bayesian inference is that 'yesterday's posterior is todays prior', meaning that the process of iteratively updating beliefs in light of evidence can be expressed as repeated applications of Bayes' rule. More formally,

$$p(h \mid d_1, ..., d_n) \propto p(d_n \mid h)p(h \mid d_1, ..., d_{n-1}) \qquad [\text{I}]$$

where $p(h \mid d_1, ..., d_n)$ is the posterior probability of h after observing data $d_1, ..., d_n$, $p(d_n \mid h)$ is the probability of d_n under h, and $p(h \mid d_1, ..., d_{n-1})$ is the posterior probability of h after observing data $d_1, ..., d_{n-1}$ (and we assume $d_1, ..., d_n$ are conditionally independent of one another given h).

The particle filter takes advantage of this structure. In this algorithm, we begin with a set of hypotheses sampled from $p(h \mid d_1, ..., d_{n-1})$. As each piece of data is observed, those hypotheses are assigned a weight based on how well they explain the observed data, reflected in $p(d_n \mid h)$. Then, a new set of hypotheses is sampled from this set, giving higher probability to those of other theories that have higher weight. The result is a set of samples that approximate the posterior distribution $p(h \mid d_1, ..., d_n)$.

To translate this to an algorithm for cultural evolution, we associate each hypothesis with a person assume each person entertains a single hypothesis and equate the sampling process with the choice by the next generation of people of who to learn from, as shown in Figure I. The success of the previous generation in explaining observed data is determined by $p(d_n \mid h)$ and we assume that more successful individuals are more likely to be chosen to learn from. The hypotheses maintained by each generation will represent samples from the Bayesian posterior distribution. For a direct application of this approach in cultural evolution, see [41]. For a more detailed articulation of the parallels between evolutionary dynamics and the particle filter, see [24].
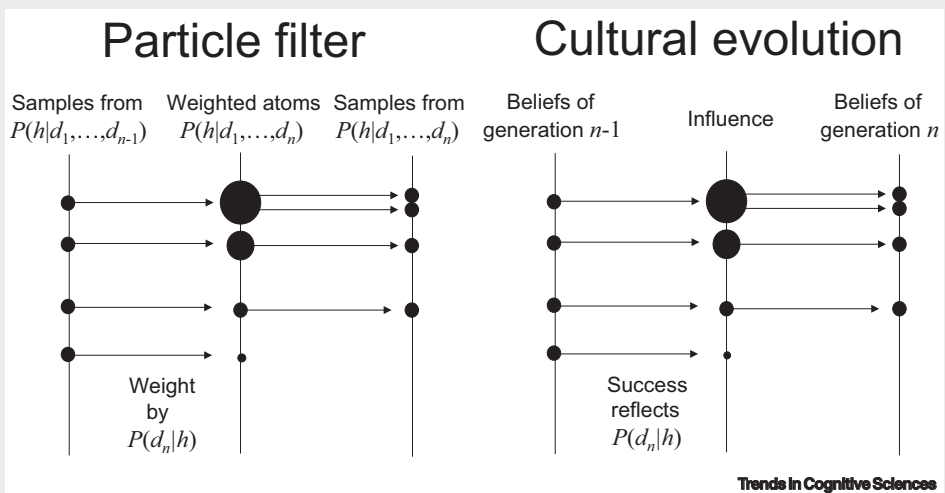

Figure I. Analogies between Parallel Algorithms and Cumulative Cultural Evolution. The particle filter is an algorithm for approximating the iterative update of a probability distribution by Bayesian inference. As described in the text, its components map directly to a population of individuals influencing the beliefs of the next generation. Sampled hypotheses become the beliefs of individuals and the process of assigning weights to those samples become the influence that individuals have when determining the beliefs of the next generation.

under the constraint that those algorithms have to be executed by a distributed set of processors (individual humans) [39]. For example, probabilistic inference, trying to compute a Bayesian posterior distribution over the true state of the world, can be executed as a serial computation across the sequence of agents, each of whom gathers some data about the environment [40], or as a parallel computation across a group of agents [41]. The literature on distributed algorithms for solving the kinds of computational problems that groups of humans have to solve is likely to yield new insights into how human minds transcend some of the limits of time and computation under which they operate (Box 4).

Concluding Remarks

The intersection of the limitations of time, computation, and communication define a set of computational problems at a very specific human scale. To the extent that human minds are adapted to solve such problems, these limitations potentially provide insight into the nature of human intelligence (but see Outstanding Questions). If AI systems are not operating under all of these constraints, then we might expect those systems to not exhibit traits that we associate with human intelligence. This does not mean that they are not intelligent systems, just that they instantiate a different flavor of intelligence that is shaped by a different set of computational problems. Understanding the human flavor of intelligence is going to require a particular set of mathematical tools, with Bayesian inference and meta-learning, rational meta-reasoning, and parallel algorithms being of particular relevance to the project of cognitive science.

The limitations that shape human intelligence also provide a way to answer the question of what makes humans unique in a world where machines are becoming increasingly intelligent. These limitations are largely a consequence of being biological organisms and are shared by all vertebrates. Asking the question of what makes humans unique, but changing the contrast class from animals to machines, suggests that perhaps we can recycle the classical answer with a corresponding change in emphasis: rather than being animals that *think*, we are *animals* that think.

Outstanding Questions

What other limitations shape human intelligence? Are time, computation, and communication the most significant?

Why do humans have the limitations that we do? Are they the intrinsic consequence of being biological organisms, or themselves the result of optimization?

Is there a point where artificial intelligence and cognitive science diverge in the mathematical tools that they require as a consequence of differences in the underlying computational problems?

Despite our biological natures being significant in characterizing the difference between human and machine intelligence, it is also clear that humans differ from other animals. Is that difference something that progress in artificial intelligence can help us to understand?

Roger Shepard was interested in how ideal solutions to computational problems could result in universal laws of cognition, principles that hold for intelligent entities throughout the universe. To the extent that human constraints are a consequence of biology, should we expect extraterrestrial biological entities to have a similar kind of intelligence?

References

1. Goodey, C.F. (1996) On Aristotle's 'animal capable of reason'. *Anc. Philos.* 16, 389–403
2. Pinker, S. (2010) The cognitive niche: coevolution of intelligence, sociality, and language. *Proc. Natl. Acad. Sci.* 107, 8993–8999
3. Boyd, R. *et al.* (2011) The cultural niche: Why social learning is essential for human adaptation. *Proc. Natl. Acad. Sci.* 108, 10918–10925
4. Henrich, J. (2017) *The Secret of Our Success*, Princeton University Press
5. Heyes, C. (2018) *Cognitive gadgets: the cultural evolution of thinking*, Harvard University Press
6. Marr, D. (1982) *Vision*, W. H. Freeman
7. Shepard, R. (1987) Towards a universal law of generalization for psychological science. *Science* 237, 1317–1323
8. Anderson, J.R. (1990) *The Adaptive Character of Thought*, Erlbaum
9. Oaksford, M. and Chater, N. (1994) A rational analysis of the selection task as optimal data selection. *Psychol. Rev.* 101, 608–631
10. Tenenbaum, J.B. and Griffiths, T.L. (2001) Generalization, similarity, and Bayesian inference. *Behav. Brain Sci.* 24, 629–641
11. Anderson, J.R. (1991) The Adaptive Nature of Human Categorization. *Psychol. Rev.* 98, 409–429
12. Ashby, F.G. and Alfonso-Reese, L.A. (1995) Categorization as probability density estimation. *J. Math. Psychol.* 39, 216–233
13. Griffiths, T.L. and Tenenbaum, J.B. (2005) Structure and strength in causal induction. *Cogn. Psychol.* 51, 354–384
14. Lake, B.M. *et al.* (2017) Building machines that learn and think like people. *Behav. Brain Sci.* 40, E253
15. Kaelbling, L.P. *et al.* (1998) Planning and acting in partially observable stochastic domains. *Artif. Intell.* 101, 99–134
16. Bellman, R.E. (1957) *Dynamic Programming*, Princeton University Press
17. Gopnik, A. *et al.* (2017) Changes in cognitive flexibility and hypothesis search across human life history from childhood to adolescence to adulthood. *Proc. Natl. Acad. Sci.* 114, 7892–7899
18. Silver, D. *et al.* (2016) Mastering the game of Go with deep neural networks and tree search. *Nature* 529, 484–489

19. Brown, T.B. *et al.* (2020) Language models are few-shot learners. *arXiv*. Published online May 28, 2020. https://arxiv.org/abs/2005.14165

20. Carleo, G. *et al.* (2019) Machine learning and the physical sciences. *Rev. Mod. Phys.* 91, 045002

21. Bobadilla, J. *et al.* (2012) A collaborative filtering approach to mitigate the new user cold start problem. *Knowl.-Based Syst.* 26, 225–238

22. Hadfield-Menell, D. and *et al.*Lee, D.D. *et al.* (2016) Cooperative inverse reinforcement learning. In *Advances in Neural Information Processing Systems (NIPS)* (29), pp. 3909–3917, NIPS

23. Mitchell, T.M. (1997) *Machine Learning*, McGraw Hill

24. Suchow, J.W. *et al.* (2017) Evolution in mind: evolutionary dynamics, cognitive processes, and Bayesian inference. *Trends Cogn. Sci.* 21, 522–530

25. Gureckis, T.M. and Markant, D.B. (2012) Self-directed learning: a cognitive and computational perspective. *Perspect. Psychol. Sci.* 7, 464–481

26. Griffiths, T.L. *et al.* (2010) Probabilistic models of cognition: exploring representations and inductive biases. *Trends Cogn. Sci.* 14, 357–364

27. Newell, A. *et al.* (1958) Elements of a theory of human problem solving. *Psychol. Rev.* 65, 151–166

28. Rokicki, T. *et al.* (2014) The diameter of the Rubik's cube group is twenty. *SIAM Rev.* 56, 645–670

29. Horvitz, E.J. (1990) *Rational Metareasoning and Compilation for Optimizing Decisions under Bounded Resources (tech. rep.)*, Stanford University

30. Russell, S.J. and Wefald, E.H. (1991) *Do the Right Thing: Studies in Limited Rationality*, MIT Press

31. Lieder, F. and Griffiths, T.L. (2019) Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behav. Brain Sci.* 43, e1

32. Lieder, F. *et al.* (2018) Rational metareasoning and the plasticity of cognitive control. *PLoS Comput. Biol.* 14, e1006043

33. Dingemanse, M. (2017) Brain-to-brain interfaces and the role of language in distributing agency. In *Distributed Agency* (Enfield, N.J. and Kockelman, P., eds), pp. 59–66, Oxford University Press

34. Kirby, S. (2001) Spontaneous evolution of linguistic structure: an iterated learning model of the emergence of regularity and irregularity. *IEEE J. Evol. Comput.* 5, 102–110

35. Kirby, S. *et al.* (2004) From UG to universals: linguistic adaptation through iterated learning. *Stud. Lang.* 28, 587–607

36. Mordatch, I. and Abbeel, P. (2018) *Emergence of grounded compositional language in multi-agent populations*, Thirty-Second AAAI Conference on Artificial Intelligence

37. Zhu, X. (2015) *Machine teaching: an inverse problem to machine learning and an approach toward optimal education*, Twenty-Ninth AAAI Conference on Artificial Intelligence

38. Boutilier, C. (1996) Planning, learning and coordination in multiagent decision processes. In *Proceedings of the 6th Conference on Theoretical Aspects of Rationality and Knowledge*, pp. 195–210

39. Krafft, P.M. and Griffiths, T.L. (2018) Levels of analysis in computational social science. In *Proceedings of the 42nd Annual Conference of the Cognitive Science Society*

40. Beppu, A. and Griffiths, T.L. (2009) Iterated learning and the cultural ratchet. In *Proceedings of the 31st Annual Meeting of the Cognitive Science Society*

41. Krafft, P.M. *et al.* (2020) Bayesian collective learning emerges from heuristic social learning. *Cognition* (in press)

42. Thrun, S. and Pratt, L. (2012) *Learning to Learn*, Springer Science & Business Media

43. Duan, Y. *et al.* (2016) RL2: fast reinforcement learning via slow reinforcement learning. *arXiv*. Published online November 9, 2016. https://arxiv.org/abs/1611.02779

44. Wang, J.X. *et al.* (2016) Learning to reinforcement learn. *arXiv* 1611, 05763

45. Andrychowicz, M. *et al.* (2016) Learning to learn by gradient descent by gradient descent. In *Advances in Neural Information Processing Systems (NIPS)* (29) (Lee, D.D. *et al.*, eds), pp. 3981–3989, NIPS

46. Finn, C. *et al.* (2017) model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*

47. Ravi, S. and Larochelle, H. (2017) Optimization as a model for few-shot learning. In *Proceedings of the 4th International Conference on Learning Representations (ICLR)*

48. Grant, E. *et al.* (2018) Recasting gradient-based meta-learning as hierarchical Bayes. In *Proceedings of the 6th International Conference on Learning Representations (ICLR)*

49. McCoy, R.T. *et al.* (2020) Universal linguistic inductive biases via meta-learning. In *Proceedings of the 44th Annual Conference of the Cognitive Science Society*

50. Kahneman, D. and Tversky, A. (1972) Subjective probability: a judgment of representativeness. *Cogn. Psychol.* 3, 430–454

51. Simon, H.A. (1955) A behavioral model of rational choice. *Q. J. Econ.* 69, 99–118

52. Chang, M.B. *et al.* (2019) Automatically composing representation transformations as a means for generalization. In *Proceedings of the International Conference on Learning Representations (ICLR)*