BEHAVIORAL AND BRAIN SCIENCES (2017), Page 1 of 72
doi:10.1017/S0140525X16001837, e253

Building machines that learn and think like people

Brenden M. Lake
Department of Psychology and Center for Data Science, New York University, New York, NY 10011
brenden@nyu.edu
http://cims.nyu.edu/~brenden/

Tomer D. Ullman
Department of Brain and Cognitive Sciences and The Center for Brains, Minds and Machines, Massachusetts Institute of Technology, Cambridge, MA 02139
tomeru@mit.edu
http://www.mit.edu/~tomeru/

Joshua B. Tenenbaum
Department of Brain and Cognitive Sciences and The Center for Brains, Minds and Machines, Massachusetts Institute of Technology, Cambridge, MA 02139
jbt@mit.edu
http://web.mit.edu/cocosci/josh.html

Samuel J. Gershman
Department of Psychology and Center for Brain Science, Harvard University, Cambridge, MA 02138, and The Center for Brains, Minds and Machines, Massachusetts Institute of Technology, Cambridge, MA 02139
gershman@fas.harvard.edu
http://gershmanlab.webfactional.com/index.html

Abstract: Recent progress in artificial intelligence has renewed interest in building systems that learn and think like people. Many advances have come from using deep neural networks trained end-to-end in tasks such as object recognition, video games, and board games, achieving performance that equals or even beats that of humans in some respects. Despite their biological inspiration and performance achievements, these systems differ from human intelligence in crucial ways. We review progress in cognitive science suggesting that truly human-like learning and thinking machines will have to reach beyond current engineering trends in both what they learn and how they learn it. Specifically, we argue that these machines should (1) build causal models of the world that support explanation and understanding, rather than merely solving pattern recognition problems; (2) ground learning in intuitive theories of physics and psychology to support and enrich the knowledge that is learned; and (3) harness compositionality and learning-to-learn to rapidly acquire and generalize knowledge to new tasks and situations. We suggest concrete challenges and promising routes toward these goals that can combine the strengths of recent neural network advances with more structured cognitive models.

1. Introduction

Artificial intelligence (AI) has been a story of booms and busts, yet by any traditional measure of success, the last few years have been marked by exceptional progress. Much of this progress has come from recent advances in "deep learning," characterized by learning large neural network-style models with multiple layers of representation (see Glossary in Table 1). These models have achieved remarkable gains in many domains spanning object recognition, speech recognition, and control (LeCun et al. 2015; Schmidhuber 2015). In object recognition, Krizhevsky et al. (2012) trained a deep convolutional neural network (ConvNet [LeCun et al. 1989]) that nearly halved the previous state-of-the-art error rate on the most challenging benchmark to date. In the years since, ConvNets continue to dominate, recently approaching human-level performance on some object recognition benchmarks (He et al. 2016; Russakovsky et al. 2015; Szegedy et al. 2014). In automatic speech recognition, hidden Markov models (HMMs) have been the leading approach since the late 1980s (Juang & Rabiner 1990), yet this framework has been chipped away piece by piece and replaced with deep learning components (Hinton et al. 2012). Now, the leading approaches to speech recognition are fully neural network systems (Graves et al. 2013; Hannun et al. 2014). Ideas from deep learning have also been applied to learning complex control problems. Mnih et al. (2015) combined ideas from deep learning and reinforcement learning to make a "deep reinforcement learning" algorithm that learns to play large classes of simple video games from just frames of pixels and the game

score, achieving human- or superhuman-level performance on many of them (see also Guo et al. 2014; Schaul et al. 2016; Stadie et al. 2016).

These accomplishments have helped neural networks regain their status as a leading paradigm in machine learning, much as they were in the late 1980s and early 1990s. The recent success of neural networks has captured attention beyond academia. In industry, companies such as Google and Facebook have active research divisions exploring these technologies, and object and speech recognition systems based on deep learning have been deployed in core products on smart phones and the web. The media have also covered many of the recent achievements of neural networks, often expressing the view that neural networks have achieved this recent success by virtue of their brain-like computation and, therefore, their ability to emulate human learning and human cognition.

Brenden M. Lake is an Assistant Professor of Psychology and Data Science at New York University. He received his Ph.D. in Cognitive Science from MIT in 2014 and his M.S. and B.S. in Symbolic Systems from Stanford University in 2009. He is a recipient of the Robert J. Glushko Prize for Outstanding Doctoral Dissertation in Cognitive Science. His research focuses on computational problems that are easier for people than they are for machines.

Tomer D. Ullman is a Postdoctoral Researcher at MIT and Harvard University through The Center for Brains, Minds and Machines (CBMM). He received his Ph.D. from the Department of Brain and Cognitive Sciences at MIT in 2015 and his B.S. in Physics and Cognitive Science from the Hebrew University of Jerusalem in 2008. His research interests include intuitive physics, intuitive psychology, and computational models of cognitive development.

Joshua B. Tenenbaum is a Professor of Computational Cognitive Science in the Department of Brain and Cognitive Sciences at MIT and a principal investigator at MIT's Computer Science and Artificial Intelligence Laboratory (CSAIL) and The Center for Brains, Minds and Machines (CBMM). He is a recipient of the Distinguished Scientific Award for Early Career Contribution to Psychology from the American Psychological Association, the Troland Research Award from the National Academy of Sciences, and the Howard Crosby Warren Medal from the Society of Experimental Psychologists. His research centers on perception, learning, and common-sense reasoning in humans and machines, with the twin goals of better understanding human intelligence in computational terms and building more human-like intelligence in machines.

Samuel J. Gershman is an Assistant Professor of Psychology at Harvard University. He received his Ph.D. in Psychology and Neuroscience from Princeton University in 2013 and his B.A. in Neuroscience and Behavior from Columbia University in 2007. He is a recipient of the Robert J. Glushko Prize for Outstanding Doctoral Dissertation in Cognitive Science. His research focuses on reinforcement learning, decision making, and memory.

In this article, we view this excitement as an opportunity to examine what it means for a machine to learn or think like a person. We first review some of the criteria previously offered by cognitive scientists, developmental psychologists, and artificial intelligence (AI) researchers. Second, we articulate what we view as the essential ingredients for building a machine that learns or thinks like a person, synthesizing theoretical ideas and experimental data from research in cognitive science. Third, we consider contemporary AI (and deep learning in particular) in the light of these ingredients, finding that deep learning models have yet to incorporate many of them, and so may be solving some problems in different ways than people do. We end by discussing what we view as the most plausible paths toward building machines that learn and think like people. This includes prospects for integrating deep learning with the core cognitive ingredients we identify, inspired in part by recent work fusing neural networks with lower-level building blocks from classic psychology and computer science (attention, working memory, stacks, queues) that have traditionally been seen as incompatible.

Beyond the specific ingredients in our proposal, we draw a broader distinction between two different computational approaches to intelligence. The statistical *pattern recognition* approach treats prediction as primary, usually in the context of a specific classification, regression, or control task. In this view, learning is about discovering features that have high-value states in common – a shared label in a classification setting or a shared value in a reinforcement learning setting – across a large, diverse set of training data.

The alternative approach treats models of the world as primary, where learning is the process of *model building*. Cognition is about using these models to understand the world, to explain what we see, to imagine what could have happened that didn't, or what could be true that isn't, and then planning actions to make it so. The difference between pattern recognition and model building, between prediction and explanation, is central to our view of human intelligence. Just as scientists seek to *explain* nature, not simply predict it, we see human thought as fundamentally a model building activity. We elaborate this key point with numerous examples below. We also discuss how pattern recognition, even if it is not the core of intelligence, can nonetheless support model building, through "model-free" algorithms that learn through experience how to make essential inferences more computationally efficient.

Before proceeding, we provide a few caveats about the goals of this article, and a brief overview of the key ideas.

1.1. What this article is not

For nearly as long as there have been neural networks, there have been critiques of neural networks (Crick 1989; Fodor & Pylyshyn 1988; Marcus 1998, 2001; Minsky & Papert 1969; Pinker & Prince 1988). Although we are critical of neural networks in this article, our goal is to build on their successes rather than dwell on their shortcomings. We see a role for neural networks in developing more human-like learning machines: They have been applied in compelling ways to many types of machine learning problems, demonstrating the power of gradient-based learning and deep hierarchies of latent variables. Neural networks also have a rich history as computational models of cognition (McClelland et al. 1986; Rumelhart et al. 1986b). It is a

Table 1. *Glossary*

Neural network: A network of simple neuron-like processing units that collectively performs complex computations. Neural networks are often organized into layers, including an input layer that presents the data (e.g., an image), hidden layers that transform the data into intermediate representations, and an output layer that produces a response (e.g., a label or an action). Recurrent connections are also popular when processing sequential data.

Deep learning: A neural network with at least one hidden layer (some networks have dozens). Most state-of-the-art deep networks are trained using the backpropagation algorithm to gradually adjust their connection strengths.

Backpropagation: Gradient descent applied to training a deep neural network. The gradient of the objective function (e.g., classification error or log-likelihood) with respect to the model parameters (e.g., connection weights) is used to make a series of small adjustments to the parameters in a direction that improves the objective function.

Convolutional neural network (ConvNet): A neural network that uses trainable filters instead of (or in addition to) fully connected layers with independent weights. The same filter is applied at many locations across an image or across a time series, leading to neural networks that are effectively larger, but with local connectivity and fewer free parameters.

Model-free and model-based reinforcement learning: Model-free algorithms directly learn a control policy without explicitly building a model of the environment (reward and state transition distributions). Model-based algorithms learn a model of the environment and use it to select actions by planning.

Deep Q-learning: A model-free reinforcement-learning algorithm used to train deep neural networks on control tasks such as playing Atari games. A network is trained to approximate the optimal action-value function $Q(s, a)$, which is the expected long-term cumulative reward of taking action a in state s and then optimally selecting future actions.

Generative model: A model that specifies a probability distribution over the data. For example, in a classification task with examples X and class labels y, a generative model specifies the distribution of data given labels $P(X \mid y)$, as well as a priori on labels $P(y)$, which can be used for sampling new examples or for classification by using Bayes' rule to compute $P(y \mid X)$. A discriminative model specifies $P(y \mid X)$ directly, possibly by using a neural network to predict the label for a given data point, and cannot directly be used to sample new examples or to compute other queries regarding the data. We will generally be concerned with directed generative models (such as Bayesian networks or probabilistic programs), which can be given a causal interpretation, although undirected (non-causal) generative models such as Boltzmann machines are also possible.

Program induction: Constructing a program that computes some desired function, where that function is typically specified by training data consisting of example input-output pairs. In the case of probabilistic programs, which specify candidate generative models for data, an abstract description language is used to define a set of allowable programs, and learning is a search for the programs likely to have generated the data.

history we describe in more detail in the next section. At a more fundamental level, any computational model of learning must ultimately be grounded in the brain's biological neural networks.

We also believe that future generations of neural networks will look very different from the current state-of-the-art neural networks. They may be endowed with intuitive physics, theory of mind, causal reasoning, and other capacities we describe in the sections that follow. More structure and inductive biases could be built into the networks or learned from previous experience with related tasks, leading to more human-like patterns of learning and development. Networks may learn to effectively search for and discover new mental models or intuitive theories, and these improved models will, in turn, enable subsequent learning, allowing systems that learn-to-learn – using previous knowledge to make richer inferences from very small amounts of training data.

It is also important to draw a distinction between AI that purports to emulate or draw inspiration from aspects of human cognition and AI that does not. This article focuses on the former. The latter is a perfectly reasonable and useful approach to developing AI algorithms: avoiding cognitive or neural inspiration as well as claims of cognitive or neural plausibility. Indeed, this is how many researchers have proceeded, and this article has little pertinence to work conducted under this research strategy.[1] On the other hand, we believe that reverse engineering human intelligence can usefully inform AI and machine learning (and has already done so), especially for the types of domains and tasks that people excel at. Despite recent computational achievements, people are better than machines at solving a range of difficult computational problems, including concept learning, scene understanding, language acquisition, language understanding, speech recognition, and so on. Other human cognitive abilities remain difficult to understand computationally, including creativity, common sense, and general-purpose reasoning. As long as natural intelligence remains the best example of intelligence, we believe that the project of reverse engineering the human solutions to difficult computational problems will continue to inform and advance AI.

Finally, whereas we focus on neural network approaches to AI, we do not wish to give the impression that these are the only contributors to recent advances in AI. On the contrary, some of the most exciting recent progress has been in new forms of probabilistic machine learning (Ghahramani 2015). For example, researchers have developed automated statistical reasoning techniques (Lloyd et al. 2014), automated techniques for model building and selection (Grosse et al. 2012), and probabilistic programming languages (e.g., Gelman et al. 2015; Goodman et al. 2008; Mansinghka et al. 2014). We believe that these approaches will play important roles in future AI systems, and they are at least as compatible with the ideas from cognitive science we discuss here. However, a full discussion of those connections is beyond the scope of the current article.

1.2. Overview of the key ideas

The central goal of this article is to propose a set of core ingredients for building more human-like learning and thinking machines. We elaborate on each of these ingredients and topics in Section 4, but here we briefly overview the key ideas.

The first set of ingredients focuses on developmental "start-up software," or cognitive capabilities present early in development. There are several reasons for this focus on development. If an ingredient is present early in development, it is certainly active and available well before a child or adult would attempt to learn the types of tasks discussed in this paper. This is true regardless of whether the early-present ingredient is itself learned from experience or innately present. Also, the earlier an ingredient is present, the more likely it is to be foundational to later development and learning.

We focus on two pieces of developmental start-up software (see Wellman & Gelman [1992] for a review of both). First is *intuitive physics* (sect. 4.1.1): Infants have primitive object concepts that allow them to track objects over time and to discount physically implausible trajectories. For example, infants know that objects will persist over time and that they are solid and coherent. Equipped with these general principles, people can learn more quickly and make more accurate predictions. Although a task may be new, physics still works the same way. A second type of software present in early development is *intuitive psychology* (sect. 4.1.2): Infants understand that other people have mental states like goals and beliefs, and this understanding strongly constrains their learning and predictions. A child watching an expert play a new video game can infer that the avatar has agency and is trying to seek reward while avoiding punishment. This inference immediately constrains other inferences, allowing the child to infer what objects are good and what objects are bad. These types of inferences further accelerate the learning of new tasks.

Our second set of ingredients focus on learning. Although there are many perspectives on learning, we see *model building* as the hallmark of human-level learning, or explaining observed data through the construction of *causal* models of the world (sect. 4.2.2). From this perspective, the early-present capacities for intuitive physics and psychology are also causal models of the world. A primary job of learning is to extend and enrich these models and to build analogous causally structured theories of other domains.

Compared with state-of-the-art algorithms in machine learning, human learning is distinguished by its richness and its efficiency. Children come with the ability and the desire to uncover the underlying causes of sparsely observed events and to use that knowledge to go far beyond the paucity of the data. It might seem paradoxical that people are capable of learning these richly structured models from very limited amounts of experience. We suggest that *compositionality* and *learning-to-learn* are ingredients that make this type of rapid model learning possible (sects. 4.2.1 and 4.2.3, respectively).

A final set of ingredients concerns how the rich models our minds build are put into action, in real time (sect. 4.3). It is remarkable how *fast* we are to perceive and to act. People can comprehend a novel scene in a fraction of a second, or a novel utterance in little more than the time it takes to say it and hear it. An important motivation for using neural networks in machine vision and speech systems is to respond as quickly as the brain does. Although neural networks are usually aiming at pattern recognition rather than model building, we discuss ways in which these "model-free" methods can accelerate slow model-based inferences in perception and cognition (sect. 4.3.1) (see Glossary in Table 1). By learning to recognize patterns in these inferences, the outputs of inference can be predicted without having to go through costly intermediate steps. Integrating neural networks that "learn to do inference" with rich model building learning mechanisms offers a promising way to explain how human minds can understand the world so well and so quickly.

We also discuss the integration of model-based and model-free methods in reinforcement learning (sect. 4.3.2.), an area that has seen rapid recent progress. Once a causal model of a task has been learned, humans can use the model to plan action sequences that maximize future reward. When rewards are used as the metric for successs in model building, this is known as model-based reinforcement learning. However, planning in complex models is cumbersome and slow, making the speed-accuracy trade-off unfavorable for real-time control. By contrast, model-free reinforcement learning algorithms, such as current instantiations of deep reinforcement learning, support fast control, but at the cost of inflexibility and possibly accuracy. We review evidence that humans combine model-based and model-free learning algorithms both competitively and cooperatively and that these interactions are supervised by metacognitive processes. The sophistication of human-like reinforcement learning has yet to be realized in AI systems, but this is an area where crosstalk between cognitive and engineering approaches is especially promising.

2. Cognitive and neural inspiration in artificial intelligence

The questions of whether and how AI should relate to human cognitive psychology are older than the terms *artificial intelligence* and *cognitive psychology*. Alan Turing suspected that it was easier to build and educate a child-machine than try to fully capture adult human cognition (Turing 1950). Turing pictured the child's mind as a notebook with "rather little mechanism and lots of blank sheets," and the mind of a child-machine as filling in the notebook by responding to rewards and punishments, similar to reinforcement learning. This view on representation and learning echoes behaviorism, a dominant psychological tradition in Turing's time. It also echoes the strong empiricism of modern connectionist models – the idea that we can learn almost everything we know from the statistical patterns of sensory inputs.

Cognitive science repudiated the oversimplified behaviorist view and came to play a central role in early AI research (Boden 2006). Newell and Simon (1961) developed their "General Problem Solver" as both an AI algorithm and a model of human problem solving, which they subsequently tested experimentally (Newell & Simon 1972). AI pioneers in other areas of research explicitly referenced human cognition and even published papers in cognitive psychology journals (e.g., Bobrow &

https://doi.org/10.1017/S0140525X17000127

Winograd 1977; Hayes-Roth & Hayes-Roth 1979; Winograd 1972). For example, Schank (1972), writing in the journal *Cognitive Psychology*, declared that "We hope to be able to build a program that can learn, as a child does, how to do what we have described in this paper instead of being spoon-fed the tremendous information necessary" (p. 629).

A similar sentiment was expressed by Minsky (1974): "I draw no boundary between a theory of human thinking and a scheme for making an intelligent machine; no purpose would be served by separating these today since neither domain has theories good enough to explain—or to produce—enough mental capacity" (p. 6).

Much of this research assumed that human knowledge representation is symbolic and that reasoning, language, planning and vision could be understood in terms of symbolic operations. Parallel to these developments, a radically different approach was being explored based on neuron-like "sub-symbolic" computations (e.g., Fukushima 1980; Grossberg 1976; Rosenblatt 1958). The representations and algorithms used by this approach were more directly inspired by neuroscience than by cognitive psychology, although ultimately it would flower into an influential school of thought about the nature of cognition: *parallel distributed processing* (PDP) (McClelland et al. 1986; Rumelhart et al. 1986b). As its name suggests, PDP emphasizes parallel computation by combining simple units to collectively implement sophisticated computations. The knowledge learned by these neural networks is thus distributed across the collection of units rather than localized as in most symbolic data structures. The resurgence of recent interest in neural networks, more commonly referred to as "deep learning," shares the same representational commitments and often even the same learning algorithms as the earlier PDP models. "Deep" refers to the fact that more powerful models can be built by composing many layers of representation (see LeCun et al. [2015] and Schmidhuber [2015] for recent reviews), still very much in the PDP style while utilizing recent advances in hardware and computing capabilities, as well as massive data sets, to learn deeper models.

It is also important to clarify that the PDP perspective is compatible with "model building" in addition to "pattern recognition." Some of the original work done under the banner of PDP (Rumelhart et al. 1986b) is closer to model building than pattern recognition, whereas the recent large-scale discriminative deep learning systems more purely exemplify pattern recognition (see Bottou [2014] for a related discussion). But, as discussed, there is also a question of the nature of the learned representations within the model – their form, compositionality, and transferability – and the developmental start-up software that was used to get there. We focus on these issues in this article.

Neural network models and the PDP approach offer a view of the mind (and intelligence more broadly) that is sub-symbolic and often populated with minimal constraints and inductive biases to guide learning. Proponents of this approach maintain that many classic types of structured knowledge, such as graphs, grammars, rules, objects, structural descriptions, and programs, can be useful yet misleading metaphors for characterizing thought. These structures are more epiphenomenal than real, emergent properties of more fundamental sub-symbolic cognitive processes (McClelland et al. 2010). Compared with other paradigms for studying cognition, this position on the nature of representation is often accompanied by a relatively "blank slate" vision of initial knowledge and representation, much like Turing's blank notebook.

When attempting to understand a particular cognitive ability or phenomenon within this paradigm, a common scientific strategy is to train a relatively generic neural network to perform the task, adding additional ingredients only when necessary. This approach has shown that neural networks can behave as if they learned explicitly structured knowledge, such as a rule for producing the past tense of words (Rumelhart & McClelland 1986), rules for solving simple balance beam physics problems (McClelland 1988), or a tree to represent types of living things (plants and animals) and their distribution of properties (Rogers & McClelland 2004). Training large-scale relatively generic networks is also the best current approach for object recognition (He et al. 2016; Krizhevsky et al. 2012; Russakovsky et al. 2015; Szegedy et al. 2014), where the high-level feature representations of these convolutional nets have also been used to predict patterns of neural response in human and macaque IT cortex (Khaligh-Razavi & Kriegeskorte 2014; Kriegeskorte 2015; Yamins et al. 2014), as well as human typicality ratings (Lake et al. 2015b) and similarity ratings (Peterson et al. 2016) for images of common objects. Moreover, researchers have trained generic networks to perform structured and even strategic tasks, such as the recent work on using a Deep Q-learning Network (DQN) to play simple video games (Mnih et al. 2015) (see Glossary in Table 1). If neural networks have such broad application in machine vision, language, and control, and if they can be trained to emulate the rule-like and structured behaviors that characterize cognition, do we need more to develop truly human-like learning and thinking machines? How far can relatively generic neural networks bring us toward this goal?

3. Challenges for building more human-like machines

Although cognitive science has not yet converged on a single account of the mind or intelligence, the claim that a mind is a collection of general-purpose neural networks with few initial constraints is rather extreme in contemporary cognitive science. A different picture has emerged that highlights the importance of early inductive biases, including core concepts such as number, space, agency, and objects, as well as powerful learning algorithms that rely on prior knowledge to extract knowledge from small amounts of training data. This knowledge is often richly organized and theory-like in structure, capable of the graded inferences and productive capacities characteristic of human thought.

Here we present two challenge problems for machine learning and AI: learning simple visual concepts (Lake et al. 2015a) and learning to play the Atari game Frostbite (Mnih et al. 2015). We also use the problems as running examples to illustrate the importance of core cognitive ingredients in the sections that follow.

3.1. The Characters Challenge

The first challenge concerns handwritten character recognition, a classic problem for comparing different types of

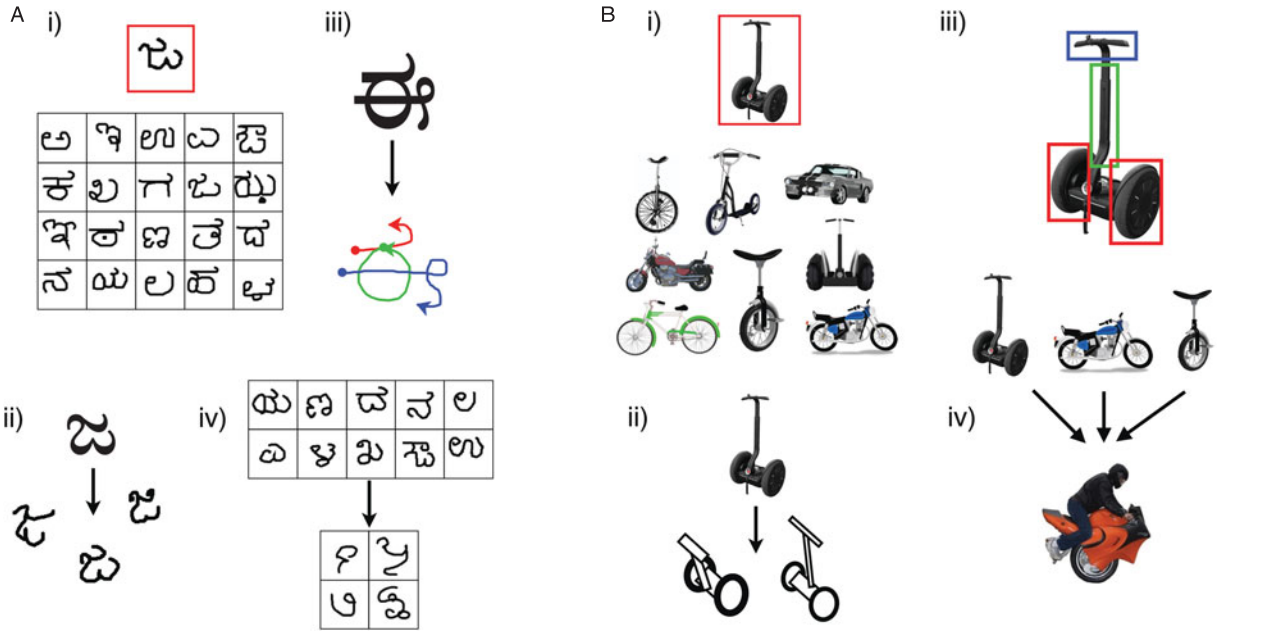

Figure 1. The Characters Challenge: Human-level learning of novel handwritten characters (A), with the same abilities also illustrated for a novel two-wheeled vehicle (B). A single example of a new visual concept (*red box*) can be enough information to support the (i) classification of new examples, (ii) generation of new examples, (iii) parsing an object into parts and relations, and (iv) generation of new concepts from related concepts. Adapted from Lake et al. (2015a).

machine learning algorithms. Hofstadter (1985) argued that the problem of recognizing characters in all of the ways people do – both handwritten and printed – contains most, if not all, of the fundamental challenges of AI. Whether or not this statement is correct, it highlights the surprising complexity that underlies even "simple" human-level concepts like letters. More practically, handwritten character recognition is a real problem that children and adults must learn to solve, with practical applications ranging from reading envelope addresses or checks in an automated teller machine (ATM). Handwritten character recognition is also simpler than more general forms of object recognition; the object of interest is two-dimensional, separated from the background, and usually unoccluded. Compared with how people learn and see other types of objects, it seems possible, in the near term, to build algorithms that can see most of the structure in characters that people can see.

The standard benchmark is the Mixed National Institute of Standards and Technology (MNIST) data set for digit recognition, which involves classifying images of digits into the categories '0' to '9' (LeCun et al. 1998). The training set provides 6,000 images per class for a total of 60,000 training images. With a large amount of training data available, many algorithms achieve respectable performance, including *K*-nearest neighbors (5% test error), support vector machines (about 1% test error), and convolutional neural networks (below 1% test error [LeCun et al. 1998]). The best results achieved using deep convolutional nets are very close to human-level performance at an error rate of 0.2% (Ciresan et al. 2012). Similarly, recent results applying convolutional nets to the far more challenging ImageNet object recognition benchmark have shown that human-level performance is within reach on that data set as well (Russakovsky et al. 2015).

Although humans and neural networks may perform equally well on the MNIST digit recognition task and other large-scale image classification tasks, it does not mean that they learn and think in the same way. There are at least two important differences: people learn from fewer examples and they learn richer representations, a comparison true for both learning handwritten characters and for learning more general classes of objects (Fig. 1). People can learn to recognize a new handwritten character from a single example (Fig. 1A-i), allowing them to discriminate between novel instances drawn by other people and similar looking non-instances (Lake et al. 2015a; Miller et al. 2000). Moreover, people learn more than how to do pattern recognition: they learn a concept, that is, a model of the class that allows their acquired knowledge to be flexibly applied in new ways. In addition to recognizing new examples, people can also generate new examples (Fig. 1A-ii), parse a character into its most important parts and relations (Fig. 1A-iii) (Lake et al. 2012), and generate new characters given a small set of related characters (Fig. 1A-iv). These additional abilities come for free along with the acquisition of the underlying concept.

Even for these simple visual concepts, people are still better and more sophisticated learners than the best algorithms for character recognition. People learn a lot more from a lot less, and capturing these human-level learning abilities in machines is the Characters Challenge. We recently reported progress on this challenge using probabilistic program induction (Lake et al. 2015a) (see Glossary in Table 1), yet aspects of the full human cognitive ability remain out of reach. Although both people and models represent characters as a sequence of pen strokes and relations, people have a far richer repertoire of structural relations between strokes. Furthermore, people can efficiently integrate across multiple examples of a character to infer which have optional elements, such as the horizontal cross-bar in '7's, combining different variants of the same character into a single coherent representation.

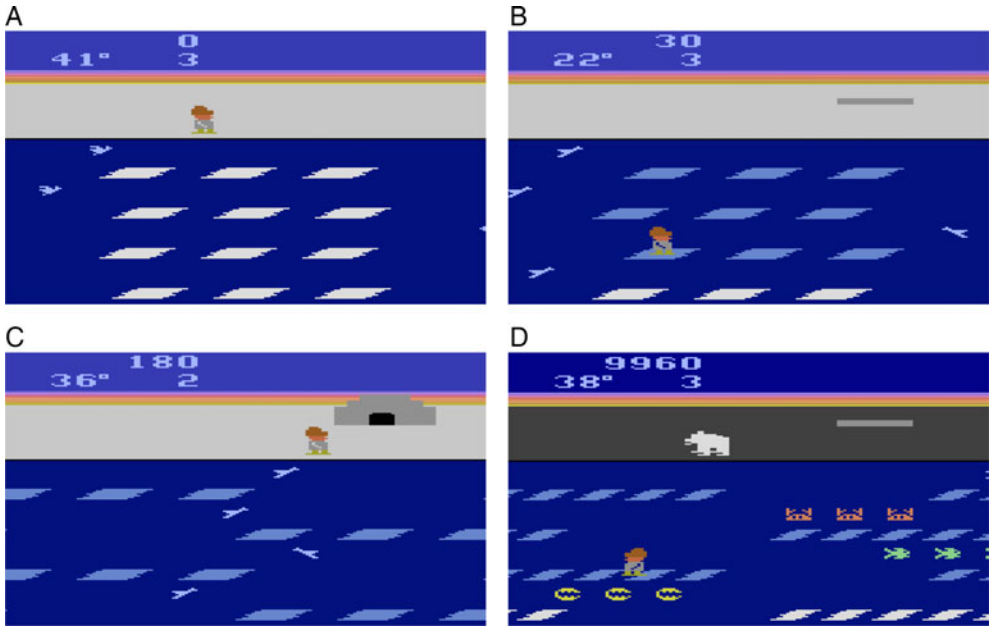

Figure 2. Screenshots of Frostbite, a 1983 video game designed for the Atari game console. (A) The start of a level in Frostbite. The agent must construct an igloo by hopping between ice floes and avoiding obstacles such as birds. The floes are in constant motion (either left or right), making multi-step planning essential to success. (B) The agent receives pieces of the igloo (top right) by jumping on the active ice floes (*white*), which then deactivates them (*blue*). (C) At the end of a level, the agent must safely reach the completed igloo. (D) Later levels include additional rewards (fish) and deadly obstacles (crabs, clams, and bears).

Additional progress may come by combining deep learning and probabilistic program induction to tackle even richer versions of the Characters Challenge.

3.2. The Frostbite Challenge

The second challenge concerns the Atari game Frostbite (Fig. 2), which was one of the control problems tackled by the DQN of Mnih et al. (2015). The DQN was a significant advance in reinforcement learning, showing that a single algorithm can learn to play a wide variety of complex tasks. The network was trained to play 49 classic Atari games, proposed as a test domain for reinforcement learning (Bellemare et al. 2013), impressively achieving human-level performance or above on 29 of the games. It did, however, have particular trouble with Frostbite and other games that required temporally extended planning strategies.

In Frostbite, players control an agent (Frostbite Bailey) tasked with constructing an igloo within a time limit. The igloo is built piece by piece as the agent jumps on ice floes in water (Fig. 2A–C). The challenge is that the ice floes are in constant motion (moving either left or right), and ice floes only contribute to the construction of the igloo if they are visited in an active state (white, rather than blue). The agent may also earn extra points by gathering fish while avoiding a number of fatal hazards (falling in the water, snow geese, polar bears, etc.). Success in this game requires a temporally extended plan to ensure the agent can accomplish a sub-goal (such as reaching an ice floe) and then safely proceed to the next sub-goal. Ultimately, once all of the pieces of the igloo are in place, the agent must proceed to the igloo and complete the level before time expires (Fig. 2C).

The DQN learns to play Frostbite and other Atari games by combining a powerful pattern recognizer (a deep convolutional neural network) and a simple model-free reinforcement learning algorithm (Q-learning [Watkins & Dayan 1992]). These components allow the network to map sensory inputs (frames of pixels) onto a policy over a small set of actions, and both the mapping and the policy are trained to optimize long-term cumulative reward (the game score). The network embodies the strongly empiricist approach characteristic of most connectionist models: very little is built into the network apart from the assumptions about image structure inherent in convolutional networks, so the network has to essentially learn a visual and conceptual system from scratch for each new game. In Mnih et al. (2015), the network architecture and hyper-parameters were fixed, but the network was trained anew for each game, meaning the visual system and the policy are highly specialized for the games it was trained on. More recent work has shown how these game-specific networks can share visual features (Rusu et al. 2016) or be used to train a multitask network (Parisotto et al. 2016), achieving modest benefits of transfer when learning to play new games.

Although it is interesting that the DQN learns to play games at human-level performance while assuming very little prior knowledge, the DQN may be learning to play Frostbite and other games in a very different way than people do. One way to examine the differences is by considering the amount of experience required for learning. In Mnih et al. (2015), the DQN was compared with a professional gamer who received approximately 2 hours of practice on each of the 49 Atari games (although he or she likely had prior experience with some of the games). The DQN was trained on 200 million frames from each of the games, which equates to approximately 924 hours of

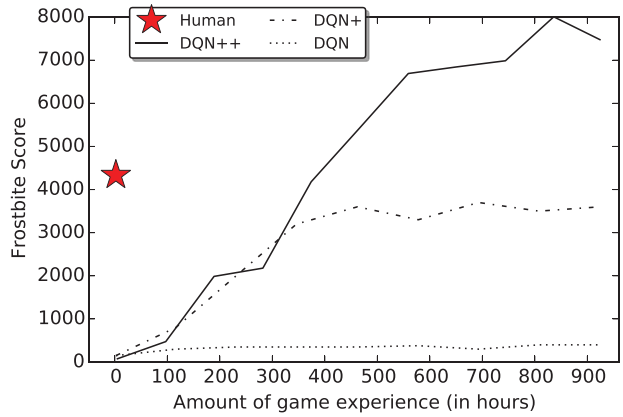

Figure 3. Comparing learning speed for people versus Deep Q-Networks (DQNs). Performance on the Atari 2600 game Frostbite is plotted as a function of game experience (in hours at a frame rate of 60 fps), which does not include additional experience replay. Learning curves and scores are shown from different networks: DQN (Mnih et al. 2015), DQN+ (Schaul et al. 2016), and DQN++ (Wang et al. 2016). Random play achieves a score of 65.2.

game time (about 38 days), or almost 500 times as much experience as the human received.[2] Additionally, the DQN incorporates experience replay, where each of these frames is replayed approximately eight more times on average over the course of learning.

With the full 924 hours of unique experience and additional replay, the DQN achieved less than 10% of human-level performance during a controlled test session (see DQN in Fig. 3). More recent variants of the DQN perform better, and can even outperform the human tester (Schaul et al. 2016; Stadie et al. 2016; van Hasselt et al. 2016; Wang et al. 2016), reaching 83% of the professional gamer's score by incorporating smarter experience replay (Schaul et al. 2016), and 172% by using smarter replay and more efficient parameter sharing (Wang et al. 2016) (see DQN+ and DQN++ in Fig. 3).[3] But they require a lot of experience to reach this level. The learning curve for the model of Wang et al. (2016) shows performance is approximately 44% after 200 hours, 8% after 100 hours, and less than 2% after 5 hours (which is close to random play, approximately 1.5%). The differences between the human and machine learning curves suggest that they may be learning different kinds of knowledge, using different learning mechanisms, or both.

The contrast becomes even more dramatic if we look at the very earliest stages of learning. Although both the original DQN and these more recent variants require multiple hours of experience to perform reliably better than random play, even non-professional humans can grasp the basics of the game after just a few minutes of play. We speculate that people do this by inferring a general schema to describe the goals of the game and the object types and their interactions, using the kinds of intuitive theories, model-building abilities and model-based planning mechanisms we describe below. Although novice players may make some mistakes, such as inferring that fish are harmful rather than helpful, they can learn even faster. In informal experiments with two of the authors playing Frostbite on a Javascript emulator (http://www.virtualatari.org/soft.php?soft=Frostbite), after watching videos of expert play on YouTube for just 2 minutes, we found that we were able to reach scores comparable to or better than the human expert reported in Mnih et al. (2015) after at most 15 to 20 minutes of total practice.[4]

There are other behavioral signatures that suggest fundamental differences in representation and learning between people and the DQN. For example, the game of Frostbite provides incremental rewards for reaching each active ice floe, providing the DQN with the relevant sub-goals for completing the larger task of building an igloo. Without these sub-goals, the DQN would have to take random actions until it accidentally builds an igloo and is rewarded for completing the entire level. In contrast, people likely do not rely on incremental scoring in the same way when figuring out how to play a new game. In Frostbite, it is possible to figure out the higher-level goal of building an igloo without incremental feedback; similarly, sparse feedback is a source of difficulty in other Atari 2600 games such as Montezuma's Revenge, in which people substantially outperform current DQN approaches.

The learned DQN network is also rather inflexible to changes in its inputs and goals. Changing the color or appearance of objects or changing the goals of the network would have devastating consequences on performance if the network is not retrained. Although any specific model is necessarily simplified and should not be held to the standard of general human intelligence, the contrast between DQN and human flexibility is striking nonetheless. For example, imagine you are tasked with playing Frostbite with any one of these new goals:

1. Get the lowest possible score.
2. Get closest to 100, or 300, or 1,000, or 3,000, or any level, without going over.
3. Beat your friend, who's playing next to you, but just barely, not by too much, so as not to embarrass them.
4. Go as long as you can without dying.
5. Die as quickly as you can.
6. Pass each level at the last possible minute, right before the temperature timer hits zero and you die (i.e., come as close as you can to dying from frostbite without actually dying).
7. Get to the furthest unexplored level without regard for your score.
8. See if you can discover secret Easter eggs.
9. Get as many fish as you can.
10. Touch all of the individual ice floes on screen once and only once.
11. Teach your friend how to play as efficiently as possible.

This range of goals highlights an essential component of human intelligence: people can learn models and use them for arbitrary new tasks and goals. Although neural networks can learn multiple mappings or tasks with the same set of stimuli – adapting their outputs depending on a specified goal – these models require substantial training or reconfiguration to add new tasks (e.g., Collins & Frank 2013; Eliasmith et al. 2012; Rougier et al. 2005). In contrast, people require little or no retraining or reconfiguration, adding new tasks and goals to their repertoire with relative ease.

The Frostbite example is a particularly telling contrast when compared with human play. Even the best deep networks learn gradually over many thousands of game episodes, take a long time to reach good performance, and are locked into particular input and goal patterns. Humans, after playing just a small number of games over a span of minutes, can understand the game and its goals well enough to perform better than deep networks do after almost a thousand hours of experience. Even more impressively, people understand enough to invent or accept new goals, generalize over changes to the input, and explain the game to others. Why are people different? What core ingredients of human intelligence might the DQN and other modern machine learning methods be missing?

One might object that both the Frostbite and Characters challenges draw an unfair comparison between the speed of human learning and neural network learning. We discuss this objection in detail in Section 5, but we feel it is important to anticipate it here as well. To paraphrase one reviewer of an earlier draft of this article, "It is not that DQN and people are solving the same task differently. They may be better seen as solving different tasks. Human learners – unlike DQN and many other deep learning systems – approach new problems armed with extensive prior experience. The human is encountering one in a years-long string of problems, with rich overlapping structure. Humans as a result often have important domain-specific knowledge for these tasks, even before they 'begin.' The DQN is starting completely from scratch."

We agree, and indeed this is another way of putting our point here. Human learners fundamentally take on different learning tasks than today's neural networks, and if we want to build machines that learn and think like people, our machines need to confront the kinds of tasks that human learners do, not shy away from them. People never start completely from scratch, or even close to "from scratch," and that is the secret to their success. The challenge of building models of human learning and thinking then becomes: How do we bring to bear rich prior knowledge to learn new tasks and solve new problems so quickly? What form does that prior knowledge take, and how is it constructed, from some combination of inbuilt capacities and previous experience? The core ingredients we propose in the next section offer one route to meeting this challenge.

4. Core ingredients of human intelligence

In the Introduction, we laid out what we see as core ingredients of intelligence. Here we consider the ingredients in detail and contrast them with the current state of neural network modeling. Although these are hardly the only ingredients needed for human-like learning and thought (see our discussion of language in sect. 5), they are key building blocks, which are not present in most current learning-based AI systems – certainly not all present together – and for which additional attention may prove especially fruitful. We believe that integrating them will produce significantly more powerful and more human-like learning and thinking abilities than we currently see in AI systems.

Before considering each ingredient in detail, it is important to clarify that by "core ingredient" we do not necessarily mean an ingredient that is innately specified by genetics or must be "built in" to any learning algorithm. We intend our discussion to be agnostic with regards to the origins of the key ingredients. By the time a child or an adult is picking up a new character or learning how to play Frostbite, he or she is armed with extensive real-world experience that deep learning systems do not benefit from – experience that would be hard to emulate in any general sense. Certainly, the core ingredients are enriched by this experience, and some may even be a product of the experience itself. Whether learned, built in, or enriched, the key claim is that these ingredients play an active and important role in producing human-like learning and thought, in ways contemporary machine learning has yet to capture.

4.1. Developmental start-up software

Early in development, humans have a foundational understanding of several core domains (Spelke 2003; Spelke & Kinzler 2007). These domains include number (numerical and set operations), space (geometry and navigation), physics (inanimate objects and mechanics), and psychology (agents and groups). These core domains cleave cognition at its conceptual joints, and each domain is organized by a set of entities and abstract principles relating the entities to each other. The underlying cognitive representations can be understood as "intuitive theories," with a causal structure resembling a scientific theory (Carey 2004; 2009; Gopnik et al. 2004; Gopnik & Meltzo 1999; Gweon et al. 2010; Schulz 2012b; Wellman & Gelman 1992; 1998). The "child as scientist" proposal further views the process of learning itself as also scientist-like, with recent experiments showing that children seek out new data to distinguish between hypotheses, isolate variables, test causal hypotheses, make use of the data-generating process in drawing conclusions, and learn selectively from others (Cook et al. 2011; Gweon et al. 2010; Schulz et al. 2007; Stahl & Feigenson 2015; Tsividis et al. 2013). We address the nature of learning mechanisms in Section 4.2.

Each core domain has been the target of a great deal of study and analysis, and together the domains are thought to be shared cross-culturally and partly with non-human animals. All of these domains may be important augmentations to current machine learning, though below, we focus in particular on the early understanding of objects and agents.

4.1.1. Intuitive physics. Young children have a rich knowledge of intuitive physics. Whether learned or innate, important physical concepts are present at ages far earlier than when a child or adult learns to play Frostbite, suggesting these resources may be used for solving this and many everyday physics-related tasks.

At the age of 2 months, and possibly earlier, human infants expect inanimate objects to follow principles of persistence, continuity, cohesion, and solidity. Young infants believe objects should move along smooth paths, not wink in and out of existence, not inter-penetrate and not act at a distance (Spelke 1990; Spelke et al. 1995). These expectations guide object segmentation in early infancy, emerging before appearance-based cues such as color, texture, and perceptual goodness (Spelke 1990).

These expectations also go on to guide later learning. At around 6 months, infants have already developed different expectations for rigid bodies, soft bodies, and liquids (Rips & Hespos 2015). Liquids, for example, are expected to go

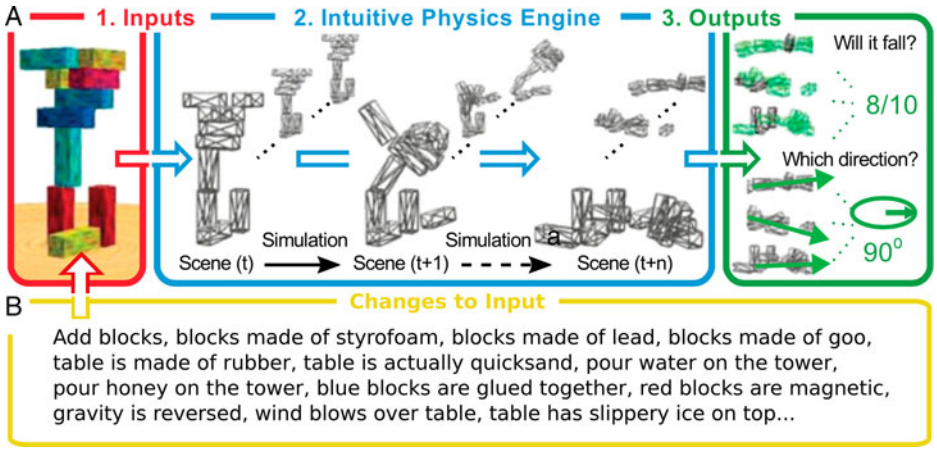

Figure 4. The intuitive physics-engine approach to scene understanding, illustrated through tower stability. (A) The engine takes in inputs through perception, language, memory, and other faculties. It then constructs a physical scene with objects, physical properties, and forces; simulates the scene's development over time; and hands the output to other reasoning systems. (B) Many possible "tweaks" to the input can result in very different scenes, requiring the potential discovery, training, and evaluation of new features for each tweak. Adapted from Battaglia et al. (2013).

through barriers, while solid objects cannot (Hespos et al. 2009). By their first birthday, infants have gone through several transitions of comprehending basic physical concepts such as inertia, support, containment, and collisions (Baillargeon 2004; Baillargeon et al. 2009; Hespos & Baillargeon 2008).

There is no single agreed-upon computational account of these early physical principles and concepts, and previous suggestions have ranged from decision trees (Baillargeon et al. 2009), to cues, to lists of rules (Siegler & Chen 1998). A promising recent approach sees intuitive physical reasoning as similar to inference over a physics software engine, the kind of simulators that power modern-day animations and games (Bates et al. 2015; Battaglia et al. 2013; Gerstenberg et al. 2015; Sanborn et al. 2013). According to this hypothesis, people reconstruct a perceptual scene using internal representations of the objects and their physically relevant properties (such as mass, elasticity, and surface friction) and forces acting on objects (such as gravity, friction, or collision impulses). Relative to physical ground truth, the intuitive physical state representation is approximate and probabilistic, and oversimplified and incomplete in many ways. Still, it is rich enough to support mental simulations that can predict how objects will move in the immediate future, either on their own or in responses to forces we might apply.

This "intuitive physics engine" approach enables flexible adaptation to a wide range of everyday scenarios and judgments in a way that goes beyond perceptual cues. For example, (Fig. 4), a physics-engine reconstruction of a tower of wooden blocks from the game Jenga can be used to predict whether (and how) a tower will fall, finding close quantitative fits to how adults make these predictions (Battaglia et al. 2013), as well as simpler kinds of physical predictions that have been studied in infants (Téglás et al. 2011). Simulation-based models can also capture how people make hypothetical or counterfactual predictions: What would happen if certain blocks were taken away, more blocks were added, or the table supporting the tower was jostled? What if certain blocks were glued together, or attached to the table surface? What if the blocks were made of different materials (Styrofoam, lead, ice)? What if the blocks of one color were much heavier than those of other colors? Each of these physical judgments may require new features or new training for a pattern recognition account to work at the same level as the model-based simulator.

What are the prospects for embedding or acquiring this kind of intuitive physics in deep learning systems? Connectionist models in psychology have previously been applied to physical reasoning tasks such as balance-beam rules (McClelland 1988; Shultz 2003) or rules relating to distance, velocity, and time in motion (Buckingham & Shultz 2000). However, these networks do not attempt to work with complex scenes as input, or a wide range of scenarios and judgments as in Figure 4. A recent paper from Facebook AI researchers (Lerer et al. 2016) represents an exciting step in this direction. Lerer et al. (2016) trained a deep convolutional network-based system (PhysNet) to predict the stability of block towers from simulated images similar to those in Figure 4A, but with much simpler configurations of two, three, or four cubical blocks stacked vertically. Impressively, PhysNet generalized to simple real images of block towers, matching human performance on these images, meanwhile exceeding human performance on synthetic images. Human and PhysNet confidence were also correlated across towers, although not as strongly as for the approximate probabilistic simulation models and experiments of Battaglia et al. (2013). One limitation is that PhysNet currently requires extensive training – between 100,000 and 200,000 scenes – to learn judgments for just a single task (will the tower fall?) on a narrow range of scenes (towers with two to four cubes). It has been shown to generalize, but also only in limited ways (e.g., from towers of two and three cubes to towers of four cubes). In contrast, people require far less experience to perform any particular task, and can generalize to many novel judgments and complex scenes with no new training required (although they receive large amounts of physics experience through interacting with the world more generally). Could deep learning systems such as PhysNet capture this flexibility, without explicitly

simulating the causal interactions between objects in three dimensions? We are not sure, but we hope this is a challenge they will take on.

Alternatively, instead of trying to make predictions without simulating physics, could neural networks be trained to emulate a general-purpose physics simulator, given the right type and quantity of training data, such as the raw input experienced by a child? This is an active and intriguing area of research, but it too faces significant challenges. For networks trained on object classification, deeper layers often become sensitive to successively higher-level features, from edges to textures to shape-parts to full objects (Yosinski et al. 2014; Zeiler & Fergus 2014). For deep networks trained on physics-related data, it remains to be seen whether higher layers will encode objects, general physical properties, forces, and approximately Newtonian dynamics. A generic network trained on dynamic pixel data might learn an implicit representation of these concepts, but would it generalize broadly beyond training contexts as people's more explicit physical concepts do? Consider, for example, a network that learns to predict the trajectories of several balls bouncing in a box (Kodratoff & Michalski 2014). If this network has actually learned something like Newtonian mechanics, then it should be able to generalize to interestingly different scenarios – at a minimum different numbers of differently shaped objects, bouncing in boxes of different shapes and sizes and orientations with respect to gravity, not to mention more severe generalization tests such as all of the tower tasks discussed above, which also fall under the Newtonian domain. Neural network researchers have yet to take on this challenge, but we hope they will. Whether such models can be learned with the kind (and quantity) of data available to human infants is not clear, as we discuss further in Section 5.

It may be difficult to integrate object and physics-based primitives into deep neural networks, but the payoff in terms of learning speed and performance could be great for many tasks. Consider the case of learning to play Frostbite. Although it can be difficult to discern exactly how a network learns to solve a particular task, the DQN probably does not parse a Frostbite screenshot in terms of stable objects or sprites moving according to the rules of intuitive physics (Fig. 2). But incorporating a physics-engine–based representation could help DQNs learn to play games such as Frostbite in a faster and more general way, whether the physics knowledge is captured implicitly in a neural network or more explicitly in a simulator. Beyond reducing the amount of training data, and potentially improving the level of performance reached by the DQN, it could eliminate the need to retrain a Frostbite network if the objects (e.g., birds, ice floes, and fish) are slightly altered in their behavior, reward structure, or appearance. When a new object type such as a bear is introduced, as in the later levels of Frostbite (Fig. 2D), a network endowed with intuitive physics would also have an easier time adding this object type to its knowledge (the challenge of adding new objects was also discussed in Marcus [1998; 2001]). In this way, the integration of intuitive physics and deep learning could be an important step toward more human-like learning algorithms.

4.1.2. Intuitive psychology. Intuitive psychology is another early-emerging ability with an important influence on human learning and thought. Pre-verbal infants distinguish animate agents from inanimate objects. This distinction is partially based on innate or early-present detectors for low-level cues, such as the presence of eyes, motion initiated from rest, and biological motion (Johnson et al. 1998; Premack & Premack 1997; Schlottmann et al. 2006; Tremoulet & Feldman 2000). Such cues are often sufficient but not necessary for the detection of agency.

Beyond these low-level cues, infants also expect agents to act contingently and reciprocally, to have goals, and to take efficient actions toward those goals subject to constraints (Csibra 2008; Csibra et al. 2003; Spelke & Kinzler 2007). These goals can be socially directed; at around 3 months of age, infants begin to discriminate antisocial agents that hurt or hinder others from neutral agents (Hamlin 2013; Hamlin et al. 2010), and they later distinguish between anti-social, neutral, and pro-social agents (Hamlin et al. 2007; 2013).

It is generally agreed that infants expect agents to act in a goal-directed, efficient, and socially sensitive fashion (Spelke & Kinzler 2007). What is less agreed on is the computational architecture that supports this reasoning and whether it includes any reference to mental states and explicit goals.

One possibility is that intuitive psychology is simply cues "all the way down" (Schlottmann et al. 2013; Scholl & Gao 2013), though this would require more and more cues as the scenarios become more complex. Consider, for example, a scenario in which an agent A is moving toward a box, and an agent B moves in a way that blocks A from reaching the box. Infants and adults are likely to interpret B's behavior as "hindering" (Hamlin 2013). This inference could be captured by a cue that states, "If an agent's expected trajectory is prevented from completion, the blocking agent is given some negative association."

Although the cue is easily calculated, the scenario is also easily changed to necessitate a different type of cue. Suppose A was already negatively associated (a "bad guy"); acting negatively toward A could then be seen as good (Hamlin 2013). Or suppose something harmful was in the box, which A did not know about. Now B would be seen as helping, protecting, or defending A. Suppose A knew there was something bad in the box and wanted it anyway. B could be seen as acting paternalistically. A cue-based account would be twisted into gnarled combinations such as, "If an expected trajectory is prevented from completion, the blocking agent is given some negative association, unless that trajectory leads to a negative outcome or the blocking agent is previously associated as positive, or the blocked agent is previously associated as negative, or…."

One alternative to a cue-based account is to use generative models of action choice, as in the Bayesian inverse planning, or Bayesian theory of mind (ToM), models of Baker et al. (2009) or the naive utility calculus models of Jara-Ettinger et al. (2015) (see also Jern and Kemp [2015] and Tauber and Steyvers [2011] and a related alternative based on predictive coding from Kilner et al. [2007]). These models formalize explicitly mentalistic concepts such as "goal," "agent," "planning," "cost," "efficiency," and "belief," used to describe core psychological reasoning in infancy. They assume adults and children treat agents as approximately rational planners who choose the most efficient means to their goals. Planning computations may be

formalized as solutions to Markov decision processes (MDPs) or partially observable Markov decision processes (POMDPs), taking as input utility and belief functions defined over an agent's state-space and the agent's state-action transition functions, and returning a series of actions the agent should perform to most efficiently fulfill their goals (or maximize their utility). By simulating these planning processes, people can predict what agents might do next, or use inverse reasoning from observing a series of actions to infer the utilities and beliefs of agents in a scene. This is directly analogous to how simulation engines can be used for intuitive physics, to predict what will happen next in a scene or to infer objects' dynamical properties from how they move. It yields similarly flexible reasoning abilities: Utilities and beliefs can be adjusted to take into account how agents might act for a wide range of novel goals and situations. Importantly, unlike in intuitive physics, simulation-based reasoning in intuitive psychology can be nested recursively to understand social interactions. We can think about agents thinking about other agents.

As in the case of intuitive physics, the success that generic deep networks will have in capturing intuitive psychological reasoning will depend in part on the representations humans use. Although deep networks have not yet been applied to scenarios involving theory of mind and intuitive psychology, they could probably learn visual cues, heuristics and summary statistics of a scene that happens to involve agents.[5] If that is all that underlies human psychological reasoning, a data-driven deep learning approach can likely find success in this domain.

However, it seems to us that any full formal account of intuitive psychological reasoning needs to include representations of agency, goals, efficiency, and reciprocal relations. As with objects and forces, it is unclear whether a complete representation of these concepts (agents, goals, etc.) could emerge from deep neural networks trained in a purely predictive capacity. Similar to the intuitive physics domain, it is possible that with a tremendous number of training trajectories in a variety of scenarios, deep learning techniques could approximate the reasoning found in infancy even without learning anything about goal-directed or socially directed behavior more generally. But this is also unlikely to resemble how humans learn, understand, and apply intuitive psychology unless the concepts are genuine. In the same way that altering the setting of a scene or the target of inference in a physics-related task may be difficult to generalize without an understanding of objects, altering the setting of an agent or their goals and beliefs is difficult to reason about without understanding intuitive psychology.

In introducing the Frostbite challenge, we discussed how people can learn to play the game extremely quickly by watching an experienced player for just a few minutes and then playing a few rounds themselves. Intuitive psychology provides a basis for efficient learning from others, especially in teaching settings with the goal of communicating knowledge efficiently (Shafto et al. 2014). In the case of watching an expert play Frostbite, whether or not there is an explicit goal to teach, intuitive psychology lets us infer the beliefs, desires, and intentions of the experienced player. For example, we can learn that the birds are to be avoided from seeing how the experienced player appears to avoid them. We do not need to experience a single example of encountering a bird, and watching Frostbite Bailey die because of the bird, to infer that birds are probably dangerous. It is enough to see that the experienced player's avoidance behavior is best explained as acting under that belief.

Similarly, consider how a sidekick agent (increasingly popular in video games) is expected to help a player achieve his or her goals. This agent can be useful in different ways in different circumstances, such as getting items, clearing paths, fighting, defending, healing, and providing information, all under the general notion of being helpful (Macindoe 2013). An explicit agent representation can predict how such an agent will be helpful in new circumstances, whereas a bottom-up pixel-based representation is likely to struggle.

There are several ways that intuitive psychology could be incorporated into contemporary deep learning systems. Although it could be built in, intuitive psychology may arise in other ways. Connectionists have argued that innate constraints in the form of hard-wired cortical circuits are unlikely (Elman 2005; Elman et al. 1996), but a simple inductive bias, for example, the tendency to notice things that move other things, can bootstrap reasoning about more abstract concepts of agency (Ullman et al. 2012a).[6] Similarly, a great deal of goal-directed and socially directed actions can also be boiled down to a simple utility calculus (e.g., Jara-Ettinger et al. 2015), in a way that could be shared with other cognitive abilities. Although the origins of intuitive psychology are still a matter of debate, it is clear that these abilities are early emerging and play an important role in human learning and thought, as exemplified in the Frostbite challenge and when learning to play novel video games more broadly.

4.2. Learning as rapid model building

Since their inception, neural networks models have stressed the importance of learning. There are many learning algorithms for neural networks, including the perceptron algorithm (Rosenblatt 1958), Hebbian learning (Hebb 1949), the BCM rule (Bienenstock et al. 1982), backpropagation (Rumelhart et al. 1986a), the wake-sleep algorithm (Hinton et al. 1995), and contrastive divergence (Hinton 2002). Whether the goal is supervised or unsupervised learning, these algorithms implement learning as a process of gradual adjustment of connection strengths. For supervised learning, the updates are usually aimed at improving the algorithm's pattern recognition capabilities. For unsupervised learning, the updates work toward gradually matching the statistics of the model's internal patterns with the statistics of the input data.

In recent years, machine learning has found particular success using backpropagation and large data sets to solve difficult pattern recognition problems (see Glossary in Table 1). Although these algorithms have reached human-level performance on several challenging benchmarks, they are still far from matching human-level learning in other ways. Deep neural networks often need more data than people do to solve the same types of problems, whether it is learning to recognize a new type of object or learning to play a new game. When learning the meanings of words in their native language, children make meaningful generalizations from very sparse data (Carey & Bartlett 1978; Landau et al. 1988; Markman 1989; Smith et al. 2002; Xu & Tenenbaum 2007; although see Horst & Samuelson 2008

https://doi.org/10.1017/S0140525X17000127

regarding memory limitations). Children may only need to see a few examples of the concepts *hairbrush*, *pineapple*, and *lightsaber*, before they largely "get it," grasping the boundary of the infinite set that defines each concept from the infinite set of all possible objects. Children are far more practiced than adults at learning new concepts, learning roughly 9 or 10 new words each day, after beginning to speak through the end of high school (Bloom 2000; Carey 1978). Yet the ability for rapid "one-shot" learning does not disappear in adulthood. An adult may need to see a single image or movie of a novel two-wheeled vehicle to infer the boundary between this concept and others, allowing him or her to discriminate new examples of that concept from similar-looking objects of a different type (Fig. 1B-i).

Contrasting with the efficiency of human learning, neural networks, by virtue of their generality as highly flexible function approximators, are notoriously data hungry (the bias/variance dilemma [Geman et al. 1992]). Benchmark tasks such as the ImageNet data set for object recognition provide hundreds or thousands of examples per class (Krizhevsky et al. 2012; Russakovsky et al. 2015): 1,000 hairbrushes, 1,000 pineapples, and so on. In the context of learning new, handwritten characters or learning to play Frostbite, the MNIST benchmark includes 6,000 examples of each handwritten digit (LeCun et al. 1998), and the DQN of Mnih et al. (2015) played each Atari video game for approximately 924 hours of unique training experience (Fig. 3). In both cases, the algorithms are clearly using information less efficiently than a person learning to perform the same tasks.

It is also important to mention that there are many classes of concepts that people learn more slowly. Concepts that are learned in school are usually far more challenging and more difficult to acquire, including mathematical functions, logarithms, derivatives, integrals, atoms, electrons, gravity, DNA, and evolution. There are also domains for which machine learners outperform human learners, such as combing through financial or weather data. But for the vast majority of cognitively natural concepts – the types of things that children learn as the meanings of words – people are still far better learners than machines. This is the type of learning we focus on in this section, which is more suitable for the enterprise of reverse engineering and articulating additional principles that make human learning successful. It also opens the possibility of building these ingredients into the next generation of machine learning and AI algorithms, with potential for making progress on learning concepts that are both easy and difficult for humans to acquire.

Even with just a few examples, people can learn remarkably rich conceptual models. One indicator of richness is the variety of functions that these models support (Markman & Ross 2003; Solomon et al. 1999). Beyond classification, concepts support prediction (Murphy & Ross 1994; Rips 1975), action (Barsalou 1983), communication (Markman & Makin 1998), imagination (Jern & Kemp 2013; Ward 1994), explanation (Lombrozo 2009; Williams & Lombrozo 2010), and composition (Murphy 1988; Osherson & Smith 1981). These abilities are not independent; rather they hang together and interact (Solomon et al. 1999), coming for free with the acquisition of the underlying concept. Returning to the previous example of a novel two-wheeled vehicle, a person can sketch a range of new instances (Fig. 1B-ii), parse the concept into its most important components (Fig. 1B-iii), or even create a new complex concept through the combination of familiar concepts (Fig. 1B-iv). Likewise, as discussed in the context of Frostbite, a learner who has acquired the basics of the game could flexibly apply his or her knowledge to an infinite set of Frostbite variants (sect. 3.2). The acquired knowledge supports reconfiguration to new tasks and new demands, such as modifying the goals of the game to survive, while acquiring as few points as possible, or to efficiently teach the rules to a friend.

This richness and flexibility suggest that learning as model building is a better metaphor than learning as pattern recognition. Furthermore, the human capacity for one-shot learning suggests that these models are built upon rich domain knowledge rather than starting from a blank slate (Mikolov et al. 2016; Mitchell et al. 1986). In contrast, much of the recent progress in deep learning has been on pattern recognition problems, including object recognition, speech recognition, and (model-free) video game learning, that use large data sets and little domain knowledge.

There has been recent work on other types of tasks, including learning generative models of images (Denton et al. 2015; Gregor et al. 2015), caption generation (Karpathy & Fei-Fei 2017; Vinyals et al. 2014; Xu et al. 2015), question answering (Sukhbaatar et al. 2015; Weston et al. 2015b), and learning simple algorithms (Graves et al. 2014; Grefenstette et al. 2015). We discuss question answering and learning simple algorithms in Section 6.1. Yet, at least for image and caption generation, these tasks have been mostly studied in the big data setting that is at odds with the impressive human ability to generalize from small data sets (although see Rezende et al. [2016] for a deep learning approach to the Character Challenge). And it has been difficult to learn neural network–style representations that effortlessly generalize new tasks that they were not trained on (see Davis & Marcus 2015; Marcus 1998; 2001). What additional ingredients may be needed to rapidly learn more powerful and more general-purpose representations?

A relevant case study is from our own work on the Characters Challenge (sect. 3.1; Lake 2014; Lake et al. 2015a). People and various machine learning approaches were compared on their ability to learn new handwritten characters from the world's alphabets. In addition to evaluating several types of deep learning models, we developed an algorithm using Bayesian program learning (BPL) that represents concepts as simple stochastic programs: structured procedures that generate new examples of a concept when executed (Fig. 5A). These programs allow the model to express causal knowledge about how the raw data are formed, and the probabilistic semantics allow the model to handle noise and perform creative tasks. Structure sharing across concepts is accomplished by the compositional re-use of stochastic primitives that can combine in new ways to create new concepts.

Note that we are overloading the word *model* to refer to the BPL framework as a whole (which is a generative model), as well as the individual probabilistic models (or concepts) that it infers from images to represent novel handwritten characters. There is a hierarchy of models: a higher-level program that generates different types of concepts, which are themselves programs that can be run to generate tokens of a concept. Here, describing learning as "rapid model building" refers to the fact that BPL

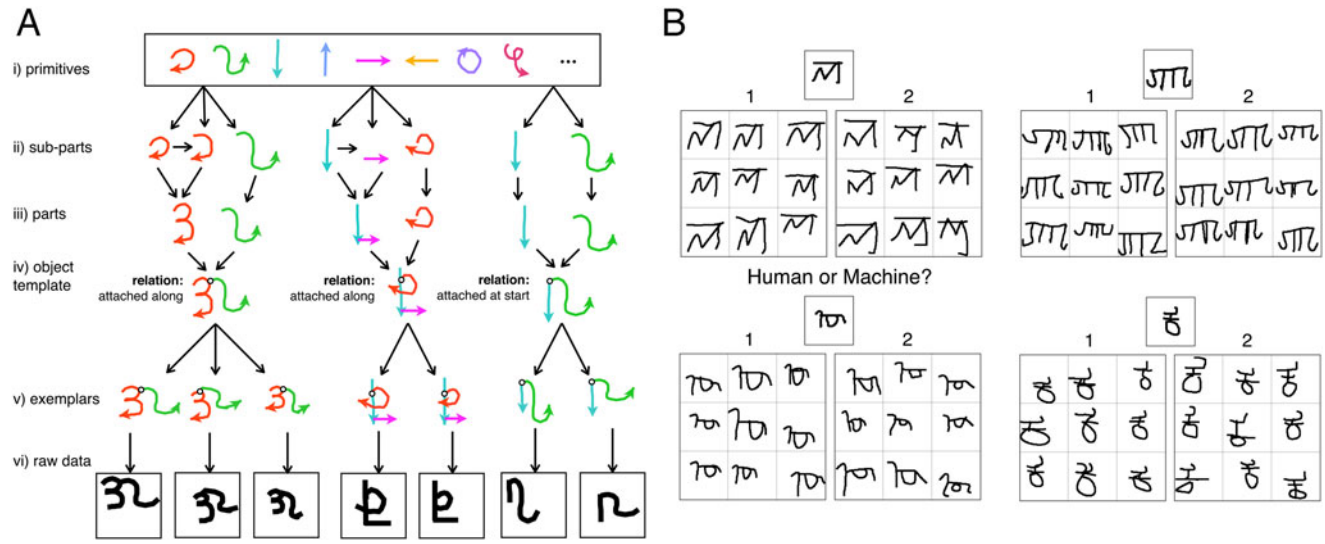

Figure 5. A causal, compositional model of handwritten characters. (A) New types are generated compositionally by choosing primitive actions (color coded) from a library (i), combining these sub-parts (ii) to make parts (iii), and combining parts with relations to define simple programs (iv). These programs can create different tokens of a concept (v) that are rendered as binary images (vi). (B) Probabilistic inference allows the model to generate new examples from just one example of a new concept; shown here in a visual Turing test. An example image of a new concept is shown above each pair of grids. One grid was generated by nine people and the other is nine samples from the BPL model. Which grid in each pair (A or B) was generated by the machine? Answers by row: 1,2;1,1. Adapted from Lake et al. (2015a).

constructs generative models (lower-level programs) that produce tokens of a concept (Fig. 5B).

Learning models of this form allows BPL to perform a challenging one-shot classification task at human-level performance (Fig. 1A-i) and to outperform current deep learning models such as convolutional networks (Koch et al. 2015).[7] The representations that BPL learns also enable it to generalize in other, more creative, human-like ways, as evaluated using "visual Turing tests" (e.g., Fig. 5B). These tasks include generating new examples (Figs. 1A-ii and 5B), parsing objects into their essential components (Fig. 1A-iii), and generating new concepts in the style of a particular alphabet (Fig. 1A-iv). The following sections discuss the three main ingredients – compositionality, causality, and learning-to-learn – that were important to the success of this framework and, we believe, are important to understanding human learning as rapid model building more broadly. Although these ingredients fit naturally within a BPL or a probabilistic program induction framework, they could also be integrated into deep learning models and other types of machine learning algorithms, prospects we discuss in more detail below.

4.2.1. Compositionality. Compositionality is the classic idea that new representations can be constructed through the combination of primitive elements. In computer programming, primitive functions can be combined to create new functions, and these new functions can be further combined to create even more complex functions. This function hierarchy provides an efficient description of higher-level functions, such as a hierarchy of parts for describing complex objects or scenes (Bienenstock et al. 1997). Compositionality is also at the core of productivity: an infinite number of representations can be constructed from a finite set of primitives, just as the mind can think an infinite number of thoughts, utter or understand an infinite number of sentences, or learn new concepts from a seemingly infinite space of possibilities (Fodor 1975; Fodor & Pylyshyn 1988; Marcus 2001; Piantadosi 2011).

Compositionality has been broadly influential in both AI and cognitive science, especially as it pertains to theories of object recognition, conceptual representation, and language. Here, we focus on compositional representations of object concepts for illustration. Structural description models represent visual concepts as compositions of parts and relations, which provides a strong inductive bias for constructing models of new concepts (Biederman 1987; Hummel & Biederman 1992; Marr & Nishihara 1978; van den Hengel et al. 2015; Winston 1975). For instance, the novel two-wheeled vehicle in Figure 1B might be represented as two wheels connected by a platform, which provides the base for a post, which holds the handlebars, and so on. Parts can themselves be composed of subparts, forming a "partonomy" of part-whole relationships (Miller & Johnson-Laird 1976; Tversky & Hemenway 1984). In the novel vehicle example, the parts and relations can be shared and re-used from existing related concepts, such as cars, scooters, motorcycles, and unicycles. Because the parts and relations are themselves a product of previous learning, their facilitation of the construction of new models is also an example of learning-to-learn, another ingredient that is covered below. Although compositionality and learning-to-learn fit naturally together, there are also forms of compositionality that rely less on previous learning, such as the bottom-up, parts-based representation of Hoffman and Richards (1984).

Learning models of novel handwritten characters can be operationalized in a similar way. Handwritten characters are inherently compositional, where the parts are pen strokes, and relations describe how these strokes connect to each other. Lake et al. (2015a) modeled these parts using an additional layer of compositionality, where parts are complex movements created from simpler sub-part movements. New characters can be constructed by

combining parts, sub-parts, and relations in novel ways (Fig. 5). Compositionality is also central to the construction of other types of symbolic concepts beyond characters, where new spoken words can be created through a novel combination of phonemes (Lake et al. 2014), or a new gesture or dance move can be created through a combination of more primitive body movements.

An efficient representation for Frostbite should be similarly compositional and productive. A scene from the game is a composition of various object types, including birds, fish, ice floes, igloos, and so on (Fig. 2). Representing this compositional structure explicitly is both more economical and better for generalization, as noted in previous work on object-oriented reinforcement learning (Diuk et al. 2008). Many repetitions of the same objects are present at different locations in the scene, and therefore, representing each as an identical instance of the same object with the same properties is important for efficient representation and quick learning of the game. Further, new levels may contain different numbers and combinations of objects, where a compositional representation of objects – using intuitive physics and intuitive psychology as glue – would aid in making these crucial generalizations (Fig. 2D).

Deep neural networks have at least a limited notion of compositionality. Networks trained for object recognition encode part-like features in their deeper layers (Zeiler & Fergus 2014), whereby the presentation of new types of objects can activate novel combinations of feature detectors. Similarly, a DQN trained to play Frostbite may learn to represent multiple replications of the same object with the same features, facilitated by the invariance properties of a convolutional neural network architecture. Recent work has shown how this type of compositionality can be made more explicit, where neural networks can be used for efficient inference in more structured generative models (both neural networks and three-dimensional scene models) that explicitly represent the number of objects in a scene (Eslami et al. 2016). Beyond the compositionality inherent in parts, objects, and scenes, compositionality can also be important at the level of goals and sub-goals. Recent work on hierarchical DQNs shows that by providing explicit object representations to a DQN, and then defining sub-goals based on reaching those objects, DQNs can learn to play games with sparse rewards (such as Montezuma's Revenge) by combining these sub-goals together to achieve larger goals (Kulkarni et al. 2016).

We look forward to seeing these new ideas continue to develop, potentially providing even richer notions of compositionality in deep neural networks that lead to faster and more flexible learning. To capture the full extent of the mind's compositionality, a model must include explicit representations of objects, identity, and relations, all while maintaining a notion of "coherence" when understanding novel configurations. Coherence is related to our next principle, causality, which is discussed in the section that follows.

4.2.2. Causality. In concept learning and scene understanding, causal models represent hypothetical real-world processes that produce the perceptual observations. In control and reinforcement learning, causal models represent the structure of the environment, such as modeling state-to-state transitions or action/state-to-state transitions.

Concept learning and vision models that use causality are usually generative (as opposed to discriminative; see Glossary in Table 1), but not every generative model is also causal. Although a generative model describes a process for generating data, or at least assigns a probability distribution over possible data points, this generative process may not resemble how the data are produced in the real world. Causality refers to the subclass of generative models that resemble, at an abstract level, how the data are actually generated. Although generative neural networks such as Deep Belief Networks (Hinton et al. 2006) or variational auto-encoders (Gregor et al. 2016; Kingma et al. 2014) may generate compelling handwritten digits, they mark one end of the "causality spectrum," because the steps of the generative process bear little resemblance to steps in the actual process of writing. In contrast, the generative model for characters using BPL does resemble the steps of writing, although even more causally faithful models are possible.

Causality has been influential in theories of perception. "Analysis-by-synthesis" theories of perception maintain that sensory data can be more richly represented by modeling the process that generated it (Bever & Poeppel 2010; Eden 1962; Halle & Stevens 1962; Neisser 1966). Relating data to their causal source provides strong priors for perception and learning, as well as a richer basis for generalizing in new ways and to new tasks. The canonical examples of this approach are speech and visual perception. For example, Liberman et al. (1967) argued that the richness of speech perception is best explained by inverting the production plan, at the level of vocal tract movements, to explain the large amounts of acoustic variability and the blending of cues across adjacent phonemes. As discussed, causality does not have to be a literal inversion of the actual generative mechanisms, as proposed in the motor theory of speech. For the BPL of learning handwritten characters, causality is operationalized by treating concepts as motor programs, or abstract causal descriptions of how to produce examples of the concept, rather than concrete configurations of specific muscles (Fig. 5A). Causality is an important factor in the model's success in classifying and generating new examples after seeing just a single example of a new concept (Lake et al. 2015a) (Fig. 5B).

Causal knowledge has also been shown to influence how people learn new concepts; providing a learner with different types of causal knowledge changes how he or she learns and generalizes. For example, the structure of the causal network underlying the features of a category influences how people categorize new examples (Rehder 2003; Rehder & Hastie 2001). Similarly, as related to the Characters Challenge, the way people learn to write a novel handwritten character influences later perception and categorization (Freyd 1983; 1987).

To explain the role of causality in learning, conceptual representations have been likened to intuitive theories or explanations, providing the glue that lets core features stick, whereas other equally applicable features wash away (Murphy & Medin 1985). Borrowing examples from Murphy and Medin (1985), the feature "flammable" is more closely attached to wood than money because of the underlying causal roles of the concepts, even though

the feature is equally applicable to both. These causal roles derive from the *functions* of objects. Causality can also glue some features together by relating them to a deeper underlying cause, explaining why some features such as "can fly," "has wings," and "has feathers" co-occur across objects, whereas others do not.

Beyond concept learning, people also understand scenes by building causal models. Human-level scene understanding involves composing a story that explains the perceptual observations, drawing upon and integrating the ingredients of intuitive physics, intuitive psychology, and compositionality. Perception without these ingredients, and absent the causal glue that binds them, can lead to revealing errors. Consider image captions generated by a deep neural network (Fig. 6) (Karpathy & Fei-Fei 2017). In many cases, the network gets the key objects in a scene correct, but fails to understand the physical forces at work, the mental states of the people, or the causal relationships between the objects. In other words, it does not build the right causal model of the data.

There have been steps toward deep neural networks and related approaches that learn causal models. Lopez-Paz et al. (2015) introduced a discriminative, data-driven framework for distinguishing the direction of causality from examples. Although it outperforms existing methods on various causal prediction tasks, it is unclear how to apply the approach to inferring rich hierarchies of latent causal variables, as needed for the Frostbite Challenge and especially the Characters Challenge. Graves (2014) learned a generative model of cursive handwriting using a recurrent neural network trained on handwriting data. Although it synthesizes impressive examples of handwriting in various styles, it requires a large training corpus and has not been applied to other tasks. The DRAW network performs both recognition and generation of handwritten digits using recurrent neural networks with a window of attention, producing a limited circular area of the image at each time step (Gregor et al. 2015). A more recent variant of DRAW was applied to generating examples of a novel character from just a single training example (Rezende et al. 2016). The model demonstrates an impressive ability to make plausible generalizations that go beyond the training examples, yet it generalizes too broadly in other cases, in ways that are not especially human-like. It is not clear that it could yet pass any of the "visual Turing tests" in Lake et al. (2015a) (Fig. 5B), although we hope DRAW-style networks will continue to be extended and enriched, and could be made to pass these tests.

Incorporating causality may greatly improve these deep learning models; they were trained without access to causal data about how characters are actually produced, and without any incentive to learn the true causal process. An attentional window is only a crude approximation of the true causal process of drawing with a pen, and in Rezende et al. (2016) the attentional window is not pen-like at all, although a more accurate pen model could be incorporated. We anticipate that these sequential generative neural networks could make sharper one-shot inferences, with the goal of tackling the full Characters Challenge by incorporating additional causal, compositional, and hierarchical structure (and by continuing to use learning-to-learn, described next), potentially leading to a more computationally efficient and neurally grounded variant of the BPL model of handwritten characters (Fig. 5).

A causal model of Frostbite would have to be more complex, gluing together object representations and explaining their interactions with intuitive physics and intuitive psychology, much like the game engine that generates the game dynamics and, ultimately, the frames of pixel images. Inference is the process of inverting this causal generative model, explaining the raw pixels as objects and their interactions, such as the agent stepping on an ice floe to deactivate it or a crab pushing the agent into the water (Fig. 2). Deep neural networks could play a role in two ways: by serving as a bottom-up proposer to make probabilistic inference more tractable in a structured generative model (sect. 4.3.1) or by serving as the causal generative model if imbued with the right set of ingredients.

4.2.3. Learning-to-learn. When humans or machines make inferences that go far beyond the data, strong prior knowledge (or inductive biases or constraints) must be making up the difference (Geman et al. 1992; Griffiths et al. 2010; Tenenbaum et al. 2011). One way people acquire this prior knowledge is through "learning-to-learn," a term introduced by Harlow (1949) and closely related to the machine learning notions of "transfer learning," "multitask learning," and "representation learning." These terms refer to ways that learning a new task or a new concept can be accelerated through previous or parallel learning of other related tasks or other related concepts. The strong priors, constraints, or inductive bias needed to learn a particular

Figure 6. Perceiving scenes without intuitive physics, intuitive psychology, compositionality, and causality. Image captions are generated by a deep neural network (Karpathy & Fei-Fei 2017) using code from github.com/karpathy/neuraltalk2. Image credits: Gabriel Villena Fernández (left), TVBS Taiwan/Agence France-Presse (middle), and AP Photo/Dave Martin (right). Similar examples using images from Reuters news can be found at twitter.com/interesting_jpg.

task quickly are often shared to some extent with other related tasks. A range of mechanisms have been developed to adapt the learner's inductive bias as they learn specific tasks and then apply these inductive biases to new tasks.

In hierarchical Bayesian modeling (Gelman et al. 2004), a general prior on concepts is shared by multiple specific concepts, and the prior itself is learned over the course of learning the specific concepts (Salakhutdinov et al. 2012; 2013). These models have been used to explain the dynamics of human learning-to-learn in many areas of cognition, including word learning, causal learning, and learning intuitive theories of physical and social domains (Tenenbaum et al. 2011). In machine vision, for deep convolutional networks or other discriminative methods that form the core of recent recognition systems, learning-to-learn can occur through the sharing of features between the models learned for old objects or old tasks and the models learned for new objects or new tasks (Anselmi et al. 2016; Baxter 2000; Bottou 2014; Lopez-Paz et al. 2016; Rusu et al. 2016; Salakhutdinov et al. 2011; Srivastava & Salakhutdinov, 2013; Torralba et al. 2007; Zeiler & Fergus 2014). Neural networks can also learn-to-learn by optimizing hyper-parameters, including the form of their weight update rule (Andrychowicz et al. 2016), over a set of related tasks.

Although transfer learning and multitask learning are already important themes across AI, and in deep learning in particular, they have not yet led to systems that learn new tasks as rapidly and flexibly as humans do. Capturing more human-like learning-to-learn dynamics in deep networks and other machine learning approaches could facilitate much stronger transfer to new tasks and new problems. To gain the full benefit that humans get from learning-to-learn, however, AI systems might first need to adopt the more compositional (or more language-like, see sect. 5) and causal forms of representations that we have argued for above.

We can see this potential in both of our challenge problems. In the Characters Challenge as presented in Lake et al. (2015a), all viable models use "pre-training" on many character concepts in a background set of alphabets to tune the representations they use to learn new character concepts in a test set of alphabets. But to perform well, current neural network approaches require much more pre-training than do people or our Bayesian program learning approach. Humans typically learn only one or a few alphabets, and even with related drawing experience, this likely amounts to the equivalent of a few hundred character-like visual concepts at most. For BPL, pre-training with characters in only five alphabets (for around 150 character types in total) is sufficient to perform human-level one-shot classification and generation of new examples. With this level of pre-training, current neural networks perform much worse on classification and have not even attempted generation; they are still far from solving the Characters Challenge.[8]

We cannot be sure how people get to the knowledge they have in this domain, but we do understand how this works in BPL, and we think people might be similar. BPL transfers readily to new concepts because it learns about object parts, sub-parts, and relations, capturing learning about what each concept is like and what concepts are like in general. It is crucial that learning-to-learn occurs at multiple levels of the hierarchical generative process. Previously learned primitive actions and larger generative pieces can be re-used and re-combined to define new generative models for new characters (Fig. 5A). Further transfer occurs by learning about the typical levels of variability within a typical generative model. This provides knowledge about how far and in what ways to generalize when we have seen only one example of a new character, which on its own could not possibly carry any information about variance. BPL could also benefit from deeper forms of learning-to-learn than it currently does. Some of the important structure it exploits to generalize well is built in to the prior and not learned from the background pre-training, whereas people might learn this knowledge, and ultimately, a human-like machine learning system should as well.

Analogous learning-to-learn occurs for humans in learning many new object models, in vision and cognition: Consider the novel two-wheeled vehicle in Figure 1B, where learning-to-learn can operate through the transfer of previously learned parts and relations (sub-concepts such as wheels, motors, handle bars, attached, powered by) that reconfigure compositionally to create a model of the new concept. If deep neural networks could adopt similarly compositional, hierarchical, and causal representations, we expect they could benefit more from learning-to-learn.

In the Frostbite Challenge, and in video games more generally, there is a similar interdependence between the form of the representation and the effectiveness of learning-to-learn. People seem to transfer knowledge at multiple levels, from low-level perception to high-level strategy, exploiting compositionality at all levels. Most basically, they immediately parse the game environment into objects, types of objects, and causal relations between them. People also understand that video games like these have goals, which often involve approaching or avoiding objects based on their type. Whether the person is a child or a seasoned gamer, it seems obvious that interacting with the birds and fish will change the game state in some way, either good or bad, because video games typically yield costs or rewards for these types of interactions (e.g., dying or points). These types of hypotheses can be quite specific and rely on prior knowledge: When the polar bear first appears and tracks the agent's location during advanced levels (Fig. 2D), an attentive learner is sure to avoid it. Depending on the level, ice floes can be spaced far apart (Fig. 2A–C) or close together (Fig. 2D), suggesting the agent may be able to cross some gaps, but not others. In this way, general world knowledge and previous video games may help inform exploration and generalization in new scenarios, helping people learn maximally from a single mistake or avoid mistakes altogether.

Deep reinforcement learning systems for playing Atari games have had some impressive successes in transfer learning, but they still have not come close to learning to play new games as quickly as humans can. For example, Parisotto et al. (2016) present the "actor-mimic" algorithm that first learns 13 Atari games by watching an expert network play and trying to mimic the expert network action selection and/or internal states (for about 4 million frames of experience each, or 18.5 hours per game). This algorithm can then learn new games faster than a randomly initialized DQN: Scores that might have taken 4 or 5 million frames of learning to reach might now be reached

after 1 or 2 million frames of practice. But anecdotally, we find that humans can still reach these scores with a few minutes of practice, requiring far less experience than the DQNs.

In sum, the interaction between representation and previous experience may be key to building machines that learn as fast as people. A deep learning system trained on many video games may not, by itself, be enough to learn new games as quickly as people. Yet, if such a system aims to learn compositionally structured causal models of each game – built on a foundation of intuitive physics and psychology – it could transfer knowledge more efficiently and thereby learn new games much more quickly.

4.3. Thinking Fast

The previous section focused on learning rich models from sparse data and proposed ingredients for achieving these human-like learning abilities. These cognitive abilities are even more striking when considering the speed of perception and thought: the amount of time required to understand a scene, think a thought, or choose an action. In general, richer and more structured models require more complex and slower inference algorithms, similar to how complex models require more data, making the speed of perception and thought all the more remarkable.

The combination of rich models with efficient inference suggests another way psychology and neuroscience may usefully inform AI. It also suggests an additional way to build on the successes of deep learning, where efficient inference and scalable learning are important strengths of the approach. This section discusses possible paths toward resolving the conflict between fast inference and structured representations, including Helmholtz machine–style approximate inference in generative models (Dayan et al. 1995; Hinton et al. 1995) and cooperation between model-free and model-based reinforcement learning systems.

4.3.1. Approximate inference in structured models. Hierarchical Bayesian models operating over probabilistic programs (Goodman et al. 2008; Lake et al. 2015a; Tenenbaum et al. 2011) are equipped to deal with theory-like structures and rich causal representations of the world, yet there are formidable algorithmic challenges for efficient inference. Computing a probability distribution over an entire space of programs is usually intractable, and often even finding a single high-probability program poses an intractable search problem. In contrast, whereas representing intuitive theories and structured causal models is less natural in deep neural networks, recent progress has demonstrated the remarkable effectiveness of gradient-based learning in high-dimensional parameter spaces. A complete account of learning and inference must explain how the brain does so much with limited computational resources (Gershman et al. 2015; Vul et al. 2014).

Popular algorithms for approximate inference in probabilistic machine learning have been proposed as psychological models (see Griffiths et al. [2012] for a review). Most prominently, it has been proposed that humans can approximate Bayesian inference using Monte Carlo methods, which stochastically sample the space of possible hypotheses and evaluate these samples according to their consistency with the data and prior knowledge (Bonawitz et al. 2014; Gershman et al. 2012; Ullman et al. 2012b; Vul et al. 2014). Monte Carlo sampling has been invoked to explain behavioral phenomena ranging from children's response variability (Bonawitz et al. 2014), to garden-path effects in sentence processing (Levy et al. 2009) and perceptual multistability (Gershman et al. 2012; Moreno-Bote et al. 2011). Moreover, we are beginning to understand how such methods could be implemented in neural circuits (Buesing et al. 2011; Huang & Rao 2014; Pecevski et al. 2011).[9]

Although Monte Carlo methods are powerful and come with asymptotic guarantees, it is challenging to make them work on complex problems like program induction and theory learning. When the hypothesis space is vast, and only a few hypotheses are consistent with the data, how can good models be discovered without exhaustive search? In at least some domains, people may not have an especially clever solution to this problem, instead grappling with the full combinatorial complexity of theory learning (Ullman et al. 2012b). Discovering new theories can be slow and arduous, as testified by the long time scale of cognitive development, and learning in a saltatory fashion (rather than through gradual adaptation) is characteristic of aspects of human intelligence, including discovery and insight during development (Schulz 2012b), problem-solving (Sternberg & Davidson 1995), and epoch-making discoveries in scientific research (Langley et al. 1987). Discovering new theories can also occur much more quickly. A person learning the rules of Frostbite will probably undergo a loosely ordered sequence of "Aha!" moments: He or she will learn that jumping on ice floes causes them to change color, that changing the color of ice floes causes an igloo to be constructed piece-by-piece, that birds make him or her lose points, that fish make him or her gain points, that he or she can change the direction of ice floes at the cost of one igloo piece, and so on. These little fragments of a "Frostbite theory" are assembled to form a causal understanding of the game relatively quickly, in what seems more like a guided process than arbitrary proposals in a Monte Carlo inference scheme. Similarly, as described in the Characters Challenge, people can quickly infer motor programs to draw a new character in a similarly guided processes.

For domains where program or theory learning occurs quickly, it is possible that people employ inductive biases not only to evaluate hypotheses, but also to guide hypothesis selection. Schulz (2012b) has suggested that abstract structural properties of problems contain information about the abstract forms of their solutions. Even without knowing the answer to the question, "Where is the deepest point in the Pacific Ocean?" one still knows that the answer must be a location on a map. The answer "20 inches" to the question, "What year was Lincoln born?" can be invalidated *a priori*, even without knowing the correct answer. In recent experiments, Tsividis et al. (2015) found that children can use high-level abstract features of a domain to guide hypothesis selection, by reasoning about distributional properties like the ratio of seeds to flowers, and dynamical properties like periodic or monotonic relationships between causes and effects (see also Magid et al. 2015).

How might efficient mappings from questions to a plausible subset of answers be learned? Recent work in AI,

spanning both deep learning and graphical models, has attempted to tackle this challenge by "amortizing" probabilistic inference computations into an efficient feed-forward mapping (Eslami et al. 2014; Heess et al. 2013; Mnih & Gregor, 2014; Stuhlmüller et al. 2013). We can also think of this as "learning to do inference," which is independent from the ideas of learning as model building discussed in the previous section. These feed-forward mappings can be learned in various ways, for example, using paired generative/recognition networks (Dayan et al. 1995; Hinton et al. 1995) and variational optimization (Gregor et al. 2015; Mnih & Gregor 2014; Rezende et al. 2014), or nearest-neighbor density estimation (Kulkarni et al. 2015a; Stuhlmüller et al. 2013). One implication of amortization is that solutions to different problems will become correlated because of the sharing of amortized computations. Some evidence for inferential correlations in humans was reported by Gershman and Goodman (2014). This trend is an avenue of potential integration of deep learning models with probabilistic models and probabilistic programming: Training neural networks to help perform probabilistic inference in a generative model or a probabilistic program (Eslami et al. 2016; Kulkarni et al. 2015b; Yildirim et al. 2015). Another avenue for potential integration is through differentiable programming (Dalrymple 2016), by ensuring that the program-like hypotheses are differentiable and thus learnable via gradient descent – a possibility discussed in the concluding section (Section 6.1).

4.3.2. Model-based and model-free reinforcement learning. The DQN introduced by Mnih et al. (2015) used a simple form of model-free reinforcement learning in a deep neural network that allows for fast selection of actions. There is indeed substantial evidence that the brain uses similar model-free learning algorithms in simple associative learning or discrimination learning tasks (see Niv 2009, for a review). In particular, the phasic firing of midbrain dopaminergic neurons is qualitatively (Schultz et al. 1997) and quantitatively (Bayer & Glimcher 2005) consistent with the reward prediction error that drives updating of model-free value estimates.

Model-free learning is not, however, the whole story. Considerable evidence suggests that the brain also has a model-based learning system, responsible for building a "cognitive map" of the environment and using it to plan action sequences for more complex tasks (Daw et al. 2005; Dolan & Dayan 2013). Model-based planning is an essential ingredient of human intelligence, enabling flexible adaptation to new tasks and goals; it is where all of the rich model-building abilities discussed in the previous sections earn their value as guides to action. As we argued in our discussion of Frostbite, one can design numerous variants of this simple video game that are identical except for the reward function; that is, governed by an identical environment model of state-action–dependent transitions. We conjecture that a competent Frostbite player can easily shift behavior appropriately, with little or no additional learning, and it is hard to imagine a way of doing that other than having a model-based planning approach in which the environment model can be modularly combined with arbitrary new reward functions and then deployed immediately for planning. One boundary condition on this flexibility is the fact that the skills become "habitized" with routine application, possibly reflecting a shift from model-based to model-free control. This shift may arise from a rational arbitration between learning systems to balance the trade-off between flexibility and speed (Daw et al. 2005; Keramati et al. 2011).

Similarly to how probabilistic computations can be amortized for efficiency (see previous section), plans can be amortized into cached values by allowing the model-based system to simulate training data for the model-free system (Sutton 1990). This process might occur offline (e.g., in dreaming or quiet wakefulness), suggesting a form of consolidation in reinforcement learning (Gershman et al. 2014). Consistent with the idea of cooperation between learning systems, a recent experiment demonstrated that model-based behavior becomes automatic over the course of training (Economides et al. 2015). Thus, a marriage of flexibility and efficiency might be achievable if we use the human reinforcement learning systems as guidance.

Intrinsic motivation also plays an important role in human learning and behavior (Berlyne 1966; Harlow 1950; Ryan & Deci 2007). Although much of the previous discussion assumes the standard view of behavior as seeking to maximize reward and minimize punishment, all externally provided rewards are reinterpreted according to the "internal value" of the agent, which may depend on the current goal and mental state. There may also be an intrinsic drive to reduce uncertainty and construct models of the environment (Edelman 2015; Schmidhuber 2015), closely related to learning-to-learn and multitask learning. Deep reinforcement learning is only just starting to address intrinsically motivated learning (Kulkarni et al. 2016; Mohamed & Rezende 2015).

5. Responses to common questions

In disussing the arguments in this article with colleagues, three lines of questioning or critiques have frequently arisen. We think it is helpful to address these points directly, to maximize the potential for moving forward together.

5.1. Comparing the learning speeds of humans and neural networks on specific tasks is not meaningful, because humans have extensive prior experience

It may seem unfair to compare neural networks and humans on the amount of training experience required to perform a task, such as learning to play new Atari games or learning new handwritten characters, when humans have had extensive prior experience that these networks have not benefited from. People have had many hours playing other games, and experience reading or writing many other handwritten characters, not to mention experience in a variety of more loosely related tasks. If neural networks were "pre-trained" on the same experience, the argument goes, then they might generalize similarly to humans when exposed to novel tasks.

This has been the rationale behind multitask learning or transfer learning, a strategy with a long history that has shown some promising results recently with deep networks (e.g., Donahue et al. 2014; Luong et al. 2015; Parisotto et al. 2016). Furthermore, some deep learning advocates

argue the human brain effectively benefits from even more experience through evolution. If deep learning researchers see themselves as trying to capture the equivalent of humans' collective evolutionary experience, this would be equivalent to a truly immense "pre-training" phase.

We agree that humans have a much richer starting point than neural networks when learning most new tasks, including learning a new concept or learning to play a new video game. That is the point of the "developmental start-up software" and other building blocks that we argued are key to creating this richer starting point. We are less committed to a particular story regarding the origins of the ingredients, including the relative roles of genetically programmed and experience-driven developmental mechanisms in building these components in early infancy. Either way, we see them as fundamental building blocks for facilitating rapid learning from sparse data.

Learning-to-learn across multiple tasks is conceivably one route to acquiring these ingredients, but simply training conventional neural networks on many related tasks may not be sufficient to generalize in human-like ways for novel tasks. As we argued in Section 4.2.3, successful learning-to-learn – or, at least, human-level transfer learning – is enabled by having models with the right representational structure, including the other building blocks discussed in this article. Learning-to-learn is a powerful ingredient, but it can be more powerful when operating over compositional representations that capture the underlying causal structure of the environment, while also building on intuitive physics and psychology.

Finally, we recognize that some researchers still hold out hope that if only they can just get big enough training data sets, sufficiently rich tasks, and enough computing power – far beyond what has been tried out so far – then deep learning methods might be sufficient to learn representations equivalent to what evolution and learning provide humans. We can sympathize with that hope, and believe it deserves further exploration, although we are not sure it is a realistic one. We understand in principle how evolution could build a brain with the cognitive ingredients we discuss here. Stochastic hill climbing is slow. It may require massively parallel exploration, over millions of years with innumerable dead ends, but it can build complex structures with complex functions if we are willing to wait long enough. In contrast, trying to build these representations from scratch using backpropagation, Deep Q-learning, or any stochastic gradient-descent weight update rule in a fixed network architecture, may be unfeasible regardless of how much training data are available. To build these representations from scratch might require exploring fundamental structural variations in the network's architecture, which gradient-based learning in weight space is not prepared to do. Although deep learning researchers do explore many such architectural variations, and have been devising increasingly clever and powerful ones recently, it is the researchers who are driving and directing this process. Exploration and creative innovation in the space of network architectures have not yet been made algorithmic. Perhaps they could, using genetic programming methods (Koza 1992) or other structure-search algorithms (Yamins et al. 2014). We think this would be a fascinating and promising direction to explore, but we may have to acquire more patience than machine-learning researchers typically express with their algorithms: the dynamics of structure search may look much more like the slow random hill climbing of evolution than the smooth, methodical progress of stochastic gradient descent. An alternative strategy is to build in appropriate infant-like knowledge representations and core ingredients as the starting point for our learning-based AI systems, or to build learning systems with strong inductive biases that guide them in this direction.

Regardless of which way an AI developer chooses to go, our main points are orthogonal to this objection. There are a set of core cognitive ingredients for human-like learning and thought. Deep learning models could incorporate these ingredients through some combination of additional structure and perhaps additional learning mechanisms, but for the most part have yet to do so. Any approach to human-like AI, whether based on deep learning or not, is likely to gain from incorporating these ingredients.

5.2. Biological plausibility suggests theories of intelligence should start with neural networks

We have focused on how cognitive science can motivate and guide efforts to engineer human-like AI, in contrast to some advocates of deep neural networks who cite neuroscience for inspiration. Our approach is guided by a pragmatic view that the clearest path to a computational formalization of human intelligence comes from understanding the "software" before the "hardware." In the case of this article, we proposed key ingredients of this software in previous sections.

Nonetheless, a cognitive approach to intelligence should not ignore what we know about the brain. Neuroscience can provide valuable inspirations for both cognitive models and AI researchers: The centrality of neural networks and model-free reinforcement learning in our proposals for "thinking fast" (sect. 4.3) are prime exemplars. Neuroscience can also, in principle, impose constraints on cognitive accounts, at both the cellular and systems levels. If deep learning embodies brain-like computational mechanisms and those mechanisms are incompatible with some cognitive theory, then this is an argument against that cognitive theory and in favor of deep learning. Unfortunately, what we "know" about the brain is not all that clear-cut. Many seemingly well-accepted ideas regarding neural computation are in fact biologically dubious, or uncertain at best, and therefore should not disqualify cognitive ingredients that pose challenges for implementation within that approach.

For example, most neural networks use some form of gradient-based (e.g., backpropagation) or Hebbian learning. It has long been argued, however, that backpropagation is not biologically plausible. As Crick (1989) famously pointed out, backpropagation seems to require that information be transmitted backward along the axon, which does not fit with realistic models of neuronal function (although recent models circumvent this problem in various ways [Liao et al. 2015; Lillicrap et al. 2014; Scellier & Bengio 2016]). This has not prevented backpropagation from being put to good use in connectionist models of cognition or in building deep neural networks for AI. Neural network researchers must regard it as a very good thing, in this case, that concerns of biological plausibility did not hold back research on this particular algorithmic approach to learning.[10] We strongly agree: Although neuroscientists

https://doi.org/10.1017/S0140525X17000127

have not found any mechanisms for implementing backpropagation in the brain, neither have they produced definitive evidence against it. The existing data simply offer little constraint either way, and backpropagation has been of obviously great value in engineering today's best pattern recognition systems.

Hebbian learning is another case in point. In the form of long-term potentiation (LTP) and spike-timing dependent plasticity (STDP), Hebbian learning mechanisms are often cited as biologically supported (Bi & Poo 2001). However, the cognitive significance of any biologically grounded form of Hebbian learning is unclear. Gallistel and Matzel (2013) have persuasively argued that the critical interstimulus interval for LTP is orders of magnitude smaller than the intervals that are behaviorally relevant in most forms of learning. In fact, experiments that simultaneously manipulate the interstimulus and intertrial intervals demonstrate that no critical interval exists. Behavior can persist for weeks or months, whereas LTP decays to baseline over the course of days (Power et al. 1997). Learned behavior is rapidly re-acquired after extinction (Bouton 2004), whereas no such facilitation is observed for LTP (Jonge & Racine 1985). Most relevantly for our focus, it would be especially challenging to try to implement the ingredients described in this article using purely Hebbian mechanisms.

Claims of biological plausibility or implausibility usually rest on rather stylized assumptions about the brain that are wrong in many of their details. Moreover, these claims usually pertain to the cellular and synaptic levels, with few connections made to systems-level neuroscience and subcortical brain organization (Edelman 2015). Understanding which details matter and which do not requires a computational theory (Marr 1982). Moreover, in the absence of strong constraints from neuroscience, we can turn the biological argument around: Perhaps a hypothetical biological mechanism should be viewed with skepticism if it is cognitively implausible. In the long run, we are optimistic that neuroscience will eventually place more constraints on theories of intelligence. For now, we believe cognitive plausibility offers a surer foundation.

5.3. Language is essential for human intelligence. Why is it not more prominent here?

We have said little in this article about people's ability to communicate and think in natural language, a distinctively human cognitive capacity where machine capabilities strikingly lag. Certainly one could argue that language should be included on any short list of key ingredients in human intelligence: For example, Mikolov et al. (2016) featured language prominently in their recent paper sketching challenge problems and a road map for AI. Moreover, whereas natural language processing is an active area of research in deep learning (e.g., Bahdanau et al. 2015; Mikolov et al. 2013; Xu et al. 2015), it is widely recognized that neural networks are far from implementing human language abilities. The question is, how do we develop machines with a richer capacity for language?

We believe that understanding language and its role in intelligence goes hand-in-hand with understanding the building blocks discussed in this article. It is also true that language builds on the core abilities for intuitive physics, intuitive psychology, and rapid learning with compositional, causal models that we focus on. These capacities are in place before children master language, and they provide the building blocks for linguistic meaning and language acquisition (Carey 2009; Jackendoff 2003; Kemp 2007; O'Donnell 2015; Pinker 2007; Xu & Tenenbaum 2007). We hope that by better understanding these earlier ingredients and how to implement and integrate them computationally, we will be better positioned to understand linguistic meaning and acquisition in computational terms and to explore other ingredients that make human language possible.

What else might we need to add to these core ingredients to get language? Many researchers have speculated about key features of human cognition that give rise to language and other uniquely human modes of thought: Is it recursion, or some new kind of recursive structure building ability (Berwick & Chomsky 2016; Hauser et al. 2002)? Is it the ability to re-use symbols by name (Deacon 1998)? Is it the ability to understand others intentionally and build shared intentionality (Bloom 2000; Frank et al. 2009; Tomasello 2010)? Is it some new version of these things, or is it just *more* of the aspects of these capacities that are already present in infants? These are important questions for future work with the potential to expand the list of key ingredients; we did not intend our list to be complete.

Finally, we should keep in mind all of the ways that acquiring language extends and enriches the ingredients of cognition that we focus on in this article. The intuitive physics and psychology of infants are likely limited to reasoning about objects and agents in their immediate spatial and temporal vicinity and to their simplest properties and states. But with language, older children become able to reason about a much wider range of physical and psychological situations (Carey 2009). Language also facilitates more powerful learning-to-learn and compositionality (Mikolov et al. 2016), allowing people to learn more quickly and flexibly by representing new concepts and thoughts in relation to existing concepts (Lupyan & Bergen 2016; Lupyan & Clark 2015). Ultimately, the full project of building machines that learn and think like humans must have language at its core.

6. Looking forward

In the last few decades, AI and machine learning have made remarkable progress: Computer programs beat chess masters; AI systems beat *Jeopardy* champions; apps recognize photos of your friends; machines rival humans on large-scale object recognition; smart phones recognize (and, to a limited extent, understand) speech. The coming years promise still more exciting AI applications, in areas as varied as self-driving cars, medicine, genetics, drug design, and robotics. As a field, AI should be proud of these accomplishments, which have helped move research from academic journals into systems that improve our daily lives.

We should also be mindful of what AI has and has not achieved. Although the pace of progress has been impressive, natural intelligence is still by far the best example of intelligence. Machine performance may rival or exceed human performance on particular tasks, and algorithms may take inspiration from neuroscience or aspects of psychology, but it does not follow that the algorithm learns

or thinks like a person. This is a higher bar worth reaching for, potentially leading to more powerful algorithms, while also helping unlock the mysteries of the human mind.

When comparing people with the current best algorithms in AI and machine learning, people learn from fewer data and generalize in richer and more flexible ways. Even for relatively simple concepts such as handwritten characters, people need to see just one or a few examples of a new concept before being able to recognize new examples, generate new examples, and generate new concepts based on related ones (Fig. 1A). So far, these abilities elude even the best deep neural networks for character recognition (Ciresan et al. 2012), which are trained on many examples of each concept and do not flexibly generalize to new tasks. We suggest that the comparative power and flexibility of people's inferences come from the causal and compositional nature of their representations.

We believe that deep learning and other learning paradigms can move closer to human-like learning and thought if they incorporate psychological ingredients, including those outlined in this article. Before closing, we discuss some recent trends that we see as some of the most promising developments in deep learning – trends we hope will continue and lead to more important advances.

6.1. Promising directions in deep learning

There has been recent interest in integrating psychological ingredients with deep neural networks, especially selective attention (Bahdanau et al. 2015; Mnih et al. 2014; Xu et al. 2015), augmented working memory (Graves et al. 2014; 2016; Grefenstette et al. 2015; Sukhbaatar et al. 2015; Weston et al. 2015b), and experience replay (McClelland et al. 1995; Mnih et al. 2015). These ingredients are lower-level than the key cognitive ingredients discussed in this article. yet they suggest a promising trend of using insights from cognitive psychology to improve deep learning, one that may be even furthered by incorporating higher-level cognitive ingredients.

Paralleling the human perceptual apparatus, selective attention forces deep learning models to process raw, perceptual data as a series of high-resolution "foveal glimpses" rather than all at once. Somewhat surprisingly, the incorporation of attention has led to substantial performance gains in a variety of domains, including in machine translation (Bahdanau et al. 2015), object recognition (Mnih et al. 2014), and image caption generation (Xu et al. 2015). Attention may help these models in several ways. It helps to coordinate complex, often sequential, outputs by attending to only specific aspects of the input, allowing the model to focus on smaller sub-tasks rather than solving an entire problem in one shot. For example, during caption generation, the attentional window has been shown to track the objects as they are mentioned in the caption, where the network may focus on a boy and then a Frisbee when producing a caption like, "A boy throws a Frisbee" (Xu et al. 2015). Attention also allows larger models to be trained without requiring every model parameter to affect every output or action. In generative neural network models, attention has been used to concentrate on generating particular regions of the image rather than the whole image at once (Gregor et al. 2015). This could be a stepping stone toward building more causal generative models in neural networks, such as a neural version of the Bayesian program learning model that could be applied to tackling the Characters Challenge (sect. 3.1).

Researchers are also developing neural networks with "working memories" that augment the shorter-term memory provided by unit activation and the longer-term memory provided by the connection weights (Graves et al. 2014; 2016; Grefenstette et al. 2015; Reed & Freitas 2016; Sukhbaatar et al. 2015; Weston et al. 2015b). These developments are also part of a broader trend toward "differentiable programming," the incorporation of classic data structures, such as random access memory, stacks, and queues, into gradient-based learning systems (Dalrymple 2016). For example, the neural Turing machine (NTM) (Graves et al. 2014) and its successor the differentiable neural computer (DNC) (Graves et al. 2016) are neural networks augmented with a random access external memory with read and write operations that maintain end-to-end differentiability. The NTM has been trained to perform sequence-to-sequence prediction tasks such as sequence copying and sorting, and the DNC has been applied to solving block puzzles and finding paths between nodes in a graph after memorizing the graph. Additionally, neural programmer-interpreters learn to represent and execute algorithms such as addition and sorting from fewer examples, by observing input-output pairs (like the NTM and DNC), as well as execution traces (Reed & Freitas 2016). Each model seems to learn genuine programs from examples, albeit in a representation more like assembly language than a high-level programming language.

Although this new generation of neural networks has yet to tackle the types of challenge problems introduced in this article, differentiable programming suggests the intriguing possibility of combining the best of program induction and deep learning. The types of structured representations and model building ingredients discussed in this article – objects, forces, agents, causality, and compositionality – help explain important facets of human learning and thinking, yet they also bring challenges for performing efficient inference (sect. 4.3.1). Deep learning systems have not yet shown they can work with these representations, but they have demonstrated the surprising effectiveness of gradient descent in large models with high-dimensional parameter spaces. A synthesis of these approaches, able to perform efficient inference over programs that richly model the causal structure an infant sees in the world, would be a major step forward in building human-like AI.

Another example of combining pattern recognition and model-based search comes from recent AI research into the game Go. Go is considerably more difficult for AI than chess, and it was only recently that a computer program – *AlphaGo* – first beat a world-class player (Chouard 2016) by using a combination of deep convolutional neural networks (ConvNets) and Monte-Carlo Tree Search (Silver et al. 2016). Each of these components has made gains against artificial and real Go players (Gelly & Silver 2008; 2011; Silver et al. 2016; Tian & Zhu 2016), and the notion of combining pattern recognition and model-based search goes back decades in Go and other games. Showing that these approaches can be integrated to beat a human Go champion is an important AI accomplishment (see Fig. 7). Just as important, however, are the new questions and directions they open up for the long-term project of building genuinely human-like AI.

https://doi.org/10.1017/S0140525X17000127

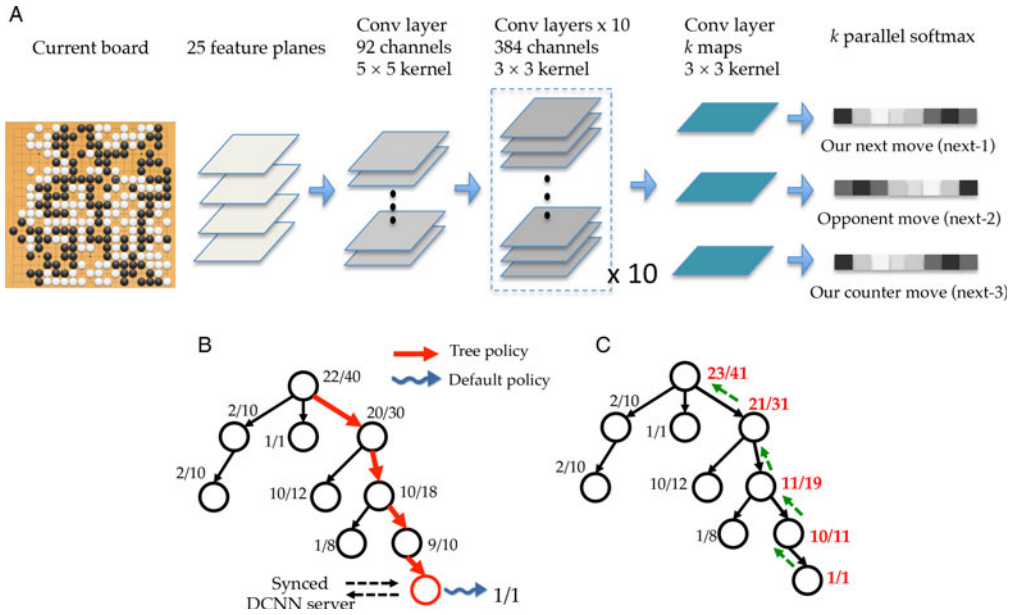

Figure 7. An AI system for playing Go, combining a deep convolutional network (ConvNet) and model-based search through Monte-Carlo Tree Search (MCTS). (A) The ConvNet on its own can be used to predict the next k moves given the current board. (B) A search tree with the current board state as its root and the current "win/total" statistics at each node. A new MCTS rollout selects moves along the tree according to the MCTS policy (*red arrows*) until it reaches a new leaf (*red circle*), where the next move is chosen by the ConvNet. From there, play proceeds until the game's end according to a pre-defined default policy based on the Pachi program (Baudiš & Gailly 2012), itself based on MCTS. (C) The end-game result of the new leaf is used to update the search tree. Adapted from Tian and Zhu (2016) with permission.

One worthy goal would be to build an AI system that beats a world-class player with the amount and kind of training human champions receive, rather than overpowering them with Google-scale computational resources. AlphaGo is initially trained on 28.4 million positions and moves from 160,000 unique games played by human experts; it then improves through reinforcement learning, playing 30 million more games against itself. Between the publication of Silver et al. (2016) and facing world champion Lee Sedol, AlphaGo was iteratively retrained several times in this way. The basic system always learned from 30 million games, but it played against successively stronger versions of itself, effectively learning from 100 million or more games altogether (D. Silver, personal communication, 2017). In contrast, Lee has probably played around 50,000 games in his entire life. Looking at numbers like these, it is impressive that Lee can even compete with AlphaGo. What would it take to build a professional-level Go AI that learns from only 50,000 games? Perhaps a system that combines the advances of AlphaGo with some of the complementary ingredients for intelligence we argue for here would be a route to that end.

Artificial intelligence could also gain much by trying to match the learning speed and flexibility of normal human Go players. People take a long time to master the game of Go, but as with the Frostbite and Characters challenges (sects. 3.1 and 3.2), humans can quickly learn the basics of the game through a combination of explicit instruction, watching others, and experience. Playing just a few games teaches a human enough to beat someone who has just learned the rules but never played before. Could AlphaGo model these earliest stages of real human learning curves? Human Go players can also adapt what they have learned to innumerable game variants. The Wikipedia page "Go variants" describes versions such as playing on bigger or smaller board sizes (ranging from 9×9 to 38×38, not just the usual 19×19 board), or playing on boards of different shapes and connectivity structures (rectangles, triangles, hexagons, even a map of the English city Milton Keynes). The board can be a torus, a mobius strip, a cube, or a diamond lattice in three dimensions. Holes can be cut in the board, in regular or irregular ways. The rules can be adapted to what is known as First Capture Go (the first player to capture a stone wins), NoGo (the player who avoids capturing any enemy stones longer wins), or Time Is Money Go (players begin with a fixed amount of time and at the end of the game, the number of seconds remaining on each player's clock is added to his or her score). Players may receive bonuses for creating certain stone patterns or capturing territory near certain landmarks. There could be four or more players, competing individually or in teams. In each of these variants, effective play needs to change from the basic game, but a skilled player can adapt, and does not simply have to relearn the game from scratch. Could AlphaGo quickly adapt to new variants of Go? Although techniques for handling variable-sized inputs in ConvNets may help in playing on different board sizes (Sermanet et al. 2014), the value functions and policies that AlphaGo learns seem unlikely to generalize as flexibly and automatically as people. Many of the variants described above would require significant reprogramming and retraining, directed by the smart humans who programmed AlphaGo, not the system itself. As impressive as AlphaGo is in beating the world's best players at the standard game – and it is extremely impressive – the fact that it cannot even conceive of these variants, let alone adapt to them autonomously, is a sign that it does not understand the game as humans do. Human players can

understand these variants and adapt to them because they explicitly represent Go *as* a game, with a goal to beat an adversary who is playing to achieve the same goal he or she is, governed by rules about how stones can be placed on a board and how board positions are scored. Humans represent their strategies as a response to these constraints, such that if the game changes, they can begin to adjust their strategies accordingly.

In sum, Go presents compelling challenges for AI beyond matching world-class human performance, in trying to match human levels of understanding and generalization, based on the same kinds and amounts of data, explicit instructions, and opportunities for social learning afforded to people. In learning to play Go as quickly and as flexibly as they do, people are drawing on most of the cognitive ingredients this article has laid out. They are learning-to-learn with compositional knowledge. They are using their core intuitive psychology and aspects of their intuitive physics (spatial and object representations). And like AlphaGo, they are also integrating model-free pattern recognition with model-based search. We believe that Go AI systems could be built to do all of these things, potentially better capturing how humans learn and understand the game. We believe it would be richly rewarding for AI and cognitive science to pursue this challenge together and that such systems could be a compelling testbed for the principles this article suggests, as well as building on all of the progress to date that AlphaGo represents.

6.2. Future applications to practical AI problems

In this article, we suggested some ingredients for building computational models with more human-like learning and thought. These principles were explained in the context of the Characters and Frostbite Challenges, with special emphasis on reducing the amount of training data required and facilitating transfer to novel yet related tasks. We also see ways these ingredients can spur progress on core AI problems with practical applications. Here we offer some speculative thoughts on these applications.

1. *Scene understanding*. Deep learning is moving beyond object recognition and toward scene understanding, as evidenced by a flurry of recent work focused on generating natural language captions for images (Karpathy & Fei-Fei 2017; Vinyals et al. 2014; Xu et al. 2015). Yet current algorithms are still better at recognizing objects than understanding scenes, often getting the key objects right but their causal relationships wrong (Fig. 6). We see compositionality, causality, intuitive physics, and intuitive psychology as playing an increasingly important role in reaching true scene understanding. For example, picture a cluttered garage workshop with screw drivers and hammers hanging from the wall, wood pieces and tools stacked precariously on a work desk, and shelving and boxes framing the scene. For an autonomous agent to effectively navigate and perform tasks in this environment, the agent would need intuitive physics to properly reason about stability and support. A holistic model of the scene would require the composition of individual object models, glued together by relations. Finally, causality helps infuse the recognition of existing tools or the learning of new ones with an understanding of their use, helping to connect different object models in the proper way (e.g., hammering a nail into a wall, or using a saw horse to support a beam being cut by a saw). If the scene includes people acting or interacting, it will be nearly impossible to understand their actions without thinking about their thoughts and especially their goals and intentions toward the other objects and agents they believe are present.

2. *Autonomous agents and intelligent devices*. Robots and personal assistants such as cell phones cannot be pre-trained on all possible concepts they may encounter. Like a child learning the meaning of new words, an intelligent and adaptive system should be able to learn new concepts from a small number of examples, as they are encountered naturally in the environment. Common concept types include new spoken words (names like "Ban Ki-Moon" and "Kofi Annan"), new gestures (a secret handshake and a "fist bump"), and new activities, and a human-like system would be able to learn both to recognize and to produce new instances from a small number of examples. As with handwritten characters, a system may be able to quickly learn new concepts by constructing them from pre-existing primitive actions, informed by knowledge of the underlying causal process and learning-to-learn.

3. *Autonomous driving*. Perfect autonomous driving requires intuitive psychology. Beyond detecting and avoiding pedestrians, autonomous cars could more accurately predict pedestrian behavior by inferring mental states, including their beliefs (e.g., Do they think it is safe to cross the street? Are they paying attention?) and desires (e.g., Where do they want to go? Do they want to cross? Are they retrieving a ball lost in the street?). Similarly, other drivers on the road have similarly complex mental states underlying their behavior (e.g., Does he or she want to change lanes? Pass another car? Is he or she swerving to avoid a hidden hazard? Is he or she distracted?). This type of psychological reasoning, along with other types of model-based causal and physical reasoning, are likely to be especially valuable in challenging and novel driving circumstances for which there are few relevant training data (e.g., navigating unusual construction zones, natural disasters).

4. *Creative design*. Creativity is often thought to be a pinnacle of human intelligence. Chefs design new dishes, musicians write new songs, architects design new buildings, and entrepreneurs start new businesses. Although we are still far from developing AI systems that can tackle these types of tasks, we see compositionality and causality as central to this goal. Many commonplace acts of creativity are combinatorial, meaning they are unexpected combinations of familiar concepts or ideas (Boden 1998; Ward 1994). As illustrated in Figure 1-iv, novel vehicles can be created as a combination of parts from existing vehicles, and similarly, novel characters can be constructed from the parts of stylistically similar characters, or familiar characters can be re-conceptualized in novel styles (Rehling 2001). In each case, the free combination of parts is not enough on its own: Although compositionality and learning-to-learn can provide the parts for new ideas, causality provides the glue that gives them coherence and purpose.

6.3. Toward more human-like learning and thinking machines

Since the birth of AI in the 1950s, people have wanted to build machines that learn and think like people. We hope

https://doi.org/10.1017/S0140525X17000127

researchers in AI, machine learning, and cognitive science will accept our challenge problems as a testbed for progress. Rather than just building systems that recognize handwritten characters and play Frostbite or Go as the end result of an asymptotic process, we suggest that deep learning and other computational paradigms should aim to tackle these tasks using as few training data as people need, and also to evaluate models on a range of human-like generalizations beyond the one task on which the model was trained. We hope that the ingredients outlined in this article will prove useful for working toward this goal: seeing objects and agents rather than features, building causal models and not just recognizing patterns, recombining representations without needing to retrain, and learning-to-learn rather than starting from scratch.

ACKNOWLEDGMENTS

We are grateful to Peter Battaglia, Matt Botvinick, Y-Lan Boureau, Shimon Edelman, Nando de Freitas, Anatole Gershman, George Kachergis, Leslie Kaelbling, Andrej Karpathy, George Konidaris, Tejas Kulkarni, Tammy Kwan, Michael Littman, Gary Marcus, Kevin Murphy, Steven Pinker, Pat Shafto, David Sontag, Pedro Tsividis, and four anonymous reviewers for helpful comments on early versions of this article. Tom Schaul and Matteo Hessel were very helpful in answering questions regarding the DQN learning curves and Frostbite scoring. This work was supported by The Center for Minds, Brains and Machines (CBMM), under National Science Foundation (NSF) Science and Technology Centers (NTS) Award CCF-1231216, and the Moore–Sloan Data Science Environment at New York University.

NOTES

1. In their influential textbook, Russell and Norvig (2003) state that "The quest for 'artificial flight' succeeded when the Wright brothers and others stopped imitating birds and started using wind tunnels and learning about aerodynamics" (p. 3).

2. The time required to train the DQN (compute time) is not the same as the game (experience) time.

3. The Atari games are deterministic, raising the possibility that a learner can succeed by memorizing long sequences of actions without learning to generalize (van Hasselt et al. 2016). A recent article shows that one can outperform DQNs early in learning (and make non-trivial generalizations) with an "episodic controller" that chooses actions based on memory and simple interpolation (Blundell et al. 2016). Although it is unclear if the DQN also memorizes action sequences, an alternative "human starts" metric provides a stronger test of generalization (van Hasselt et al. 2016), evaluating the algorithms on a wider variety of start states and levels that are sampled from human play. It would be preferable to compare people and algorithms on the human starts metric, but most learning curves to date have only been reported using standard test performance, which starts the game from the beginning with some added jitter.

4. More precisely, the human expert in Mnih et al. (2015) scored an average of 4335 points across 30 game sessions of up to 5 minutes of play. In individual sessions lasting no longer than 5 minutes, author TDU obtained scores of 3520 points after approximately 5 minutes of gameplay, 3510 points after 10 minutes, and 7810 points after 15 minutes. Author JBT obtained 4060 after approximately 5 minutes of gameplay, 4920 after 10 to 15 minutes, and 6710 after no more than 20 minutes. TDU and JBT each watched approximately 2 minutes of expert play on YouTube (e.g., https://www.youtube.com/watch?v=ZpUFztf9Fjc, but there are many similar examples that can be found in a YouTube search).

5. Although connectionist networks have been used to model the general transition that children undergo between the ages of 3 and 4 regarding false belief (e.g., Berthiaume et al. 2013), we are referring here to scenarios, which require inferring goals, utilities, and relations.

6. We must be careful here about what "simple" means. An inductive bias may appear simple in the sense that we can compactly describe it, but it may require complex computation (e.g., motion analysis, parsing images into objects, etc.) just to produce its inputs in a suitable form.

7. A new approach using convolutional "matching networks" achieves good one-shot classification performance when discriminating between characters from different alphabets (Vinyals et al. 2016). It has not yet been directly compared with BPL, which was evaluated on one-shot classification with characters from the same alphabet.

8. Deep convolutional neural network classifiers have error rates approximately five times higher than those of humans when pre-trained with five alphabets (23% versus 4% error), and two to three times higher when pre-training on six times as much data (30 alphabets) (Lake et al. 2015a). The current need for extensive pre-training is illustrated for deep generative models by Rezende et al. (2016), who present extensions of the DRAW architecture capable of one-shot learning.

9. In the interest of brevity, we do not discuss here another important vein of work linking neural circuits to variational approximations (Bastos et al. 2012), which have received less attention in the psychological literature.

10. Michael Jordan made this point forcefully in his 2015 speech accepting the Rumelhart Prize.

Open Peer Commentary

The architecture challenge: Future artificial-intelligence systems will require sophisticated architectures, and knowledge of the brain might guide their construction

doi:10.1017/S0140525X17000036, e254

Gianluca Baldassarre, Vieri Giuliano Santucci, Emilio Cartoni, and Daniele Caligiore

Laboratory of Computational Embodied Neuroscience, Institute of Cognitive Sciences and Technologies, National Research Council of Italy, Rome, Italy.

gianluca.baldassarre@istc.cnr.it **vieri.santucci@istc.cnr.it**
emilio.cartoni@istc.cnr.it **daniele.caligiore@istc.cnr.it**
http://www.istc.cnr.it/people/
http://www.istc.cnr.it/people/gianluca-baldassarre
http://www.istc.cnr.it/people/vieri-giuliano-santucci
http://www.istc.cnr.it/people/emilio-cartoni
http://www.istc.cnr.it/people/daniele-caligiore

Abstract: In this commentary, we highlight a crucial challenge posed by the proposal of Lake et al. to introduce key elements of human cognition into deep neural networks and future artificial-intelligence systems: the need to design effective sophisticated architectures. We propose that looking at the brain is an important means of facing this great challenge.

We agree with the claim of Lake et al. that to obtain human-level learning speed and cognitive flexibility, future artificial-intelligence (AI) systems will have to incorporate key elements of human cognition: from causal models of the world, to intuitive psychological theories, compositionality, and knowledge transfer. However, the authors largely overlook the importance of a major challenge to implementation of the functions they advocate: the need to develop sophisticated *architectures* to learn,

represent, and process the knowledge related to those functions. Here we call this the *architecture challenge*. In this commentary, we make two claims: (1) tackling the architecture challenge is fundamental to success in developing human-level AI systems; (2) looking at the brain can furnish important insights on how to face the architecture challenge.

The difficulty of the architecture challenge stems from the fact that the space of the architectures needed to implement the several functions advocated by Lake et al. *is huge*. The authors get close to this problem when they recognize that one thing that the enormous genetic algorithm of evolution has done in millions of years of the stochastic hill-climbing search is to develop suitable brain architectures. One possible way to attack the architecture challenge, also mentioned by Lake et al., would be to use evolutionary techniques mimicking evolution. We think that today this strategy is out of reach, given the "ocean-like" size of the search space. At most, we can use such techniques to explore small, interesting "islands lost within the ocean." But how do we find those islands in the first place? We propose looking at the architecture of real brains, the product of the evolution genetic algorithm, and try to "steal insights" from nature. Indeed, we think that *much of the intelligence of the brain resides in its architecture*. Obviously, identifying the proper insights is not easy to do, as the brain is very difficult to understand. However, it might be useful to try, as the effort might give us at least some general indications, a compass, to find the islands in the ocean. Here we present some examples to support our intuition.

When building architectures of AI systems, even when following cognitive science indications (e.g., Franklin 2007), the tendency is to "divide and conquer," that is, to list the needed high-level functions, implement a module for each of them, and suitably interface the modules. However, the organisation of the brain can be understood on the basis of not only high-level functions (see below), but also "low-level" functions (usually called "mechanisms"). An example of a mechanism is brain organisation based on macrostructures, each having fine repeated micro-architectures implementing specific computations and learning processes (Caligiore et al. 2016; Doya 1999): the cortex to statically and dynamically store knowledge acquired by associative learning processes (Penhune & Steele 2012; Shadmehr & Krakauer 2008), the basal ganglia to learn to select information by reinforcement learning (Graybiel 2005; Houk et al. 1995), the cerebellum to implement fast time-scale computations possibly acquired with supervised learning (Kawato et al. 2011; Wolpert et al. 1998), and the limbic brain structures interfacing the brain to the body and generating motivations, emotions, and the value of things (Mirolli et al. 2010; Mogenson et al. 1980). Each of these mechanisms supports multiple, high-level functions (see below).

Brain architecture is also forged by the fact that natural intelligence is strongly *embodied* and *situated* (an aspect not much stressed by Lake et al.); that is, it is shaped to adaptively interact with the physical world (Anderson 2003; Pfeifer & Gómez 2009) to satisfy the organism's needs and goals (Mannella et al. 2013). Thus, the cortex is organised along multiple cortical pathways running from sensors to actuators (Baldassarre et al. 2013a) and "intercepted" by the basal ganglia selective processes in their last part closer to action (Mannella & Baldassarre 2015). These pathways are organised in a hierarchical fashion, with the higher ones that process needs and motivational information controlling the lower ones closer to sensation/action. The lowest pathways dynamically connect musculoskeletal body proprioception with primary motor areas (Churchland et al. 2012). Higher-level "dorsal" pathways control the lowest pathways by processing visual/auditory information used to interact with the environment (Scott 2004). Even higher-level "ventral" pathways inform the brain on the identity and nature of resources in the environment to support decisions (Caligiore et al. 2010; Milner & Goodale 2006). At the hierarchy apex, the limbic brain supports goal selection based on visceral, social, and other types of needs/goals. Embedded within the higher pathways, an important structure involving basal ganglia–cortical loops learns and implements stimulus–response habitual behaviours (used to act in familiar situations) and goal-directed behaviours (important for problem solving and planning when new challenges are encountered) (Baldassarre et al. 2013b; Mannella et al. 2013). These brain structures form a sophisticated network, knowledge of which might help in designing the architectures of human-like embodied AI systems able to act in the real world.

A last example of the need for sophisticated architectures starts with the recognition by Lake et al. that we need to endow AI systems with a "developmental start-up software." In this respect, together with other authors (e.g., Weng et al. 2001; see Baldassarre et al. 2013b; 2014, for collections of works) we believe that human-level intelligence can be achieved only through *open-ended learning*, that is, the cumulative learning of progressively more complex skills and knowledge, driven by *intrinsic motivations*, which are motivations related to the acquisition of knowledge and skills rather than material resources (Baldassarre 2011). The brain (e.g., Lisman & Grace 2005; Redgrave & Gurney 2006) and computational theories and models (e.g., Baldassarre & Mirolli 2013; Baldassarre et al. 2014; Santucci et al. 2016) indicate how the implementation of these processes indeed requires very sophisticated architectures able to store multiple skills, to transfer knowledge while avoiding catastrophic interference, to explore the environment based on the acquired skills, to self-generate goals/tasks, and to focus on goals that ensure a maximum knowledge gain.

Building machines that learn and think for themselves

doi:10.1017/S0140525X17000048, e255

Matthew Botvinick, David G. T. Barrett, Peter Battaglia, Nando de Freitas, Darshan Kumaran, Joel Z Leibo, Timothy Lillicrap, Joseph Modayil, Shakir Mohamed, Neil C. Rabinowitz, Danilo J. Rezende, Adam Santoro, Tom Schaul, Christopher Summerfield, Greg Wayne, Theophane Weber, Daan Wierstra, Shane Legg, and Demis Hassabis

DeepMind, Kings Cross, London N1c4AG, United Kingdom.

botvinick@google.com barrettdavid@google.com
peterbattaglia@google.com nandodefreitas@google.com
dkumaran@google.com jzl@google.com
countzero@google.com modayil@google.com
shakir@google.com ncr@google.com
danilor@google.com adamsantoro@google.com
schaul@google.com csummerfield@google.com
gregwayne@google.com theophane@google.com
wierstra@google.com legg@google.com
demishassahassaibis@google.com
http://www.deepmind.com

Abstract: We agree with Lake and colleagues on their list of "key ingredients" for building human-like intelligence, including the idea that model-based reasoning is essential. However, we favor an approach that centers on one additional ingredient: autonomy. In particular, we aim toward agents that can both build and exploit their own internal models, with minimal human hand engineering. We believe an approach centered on autonomous learning has the greatest chance of success as we scale toward real-world complexity, tackling domains for which ready-made formal models are not available. Here, we survey several important examples of the progress that has been made toward building autonomous agents with human-like abilities, and highlight some outstanding challenges.

Lake et al. identify some extremely important desiderata for human-like intelligence. We agree with many of their central assertions: Human-like learning and decision making surely do depend upon rich internal models; the learning process must be informed and constrained by prior knowledge, whether this is

part of the agent's initial endowment or acquired through learning; and naturally, prior knowledge will offer the greatest leverage when it reflects the most pervasive or ubiquitous structures in the environment, including physical laws, the mental states of others, and more abstract regularities such as compositionality and causality. Together, these points comprise a powerful set of target goals for AI research. However, while we concur on these goals, we choose a differently calibrated strategy for accomplishing them. In particular, we favor an approach that prioritizes autonomy, empowering artificial agents to learn their own internal models and how to use them, mitigating their reliance on detailed configuration by a human engineer.

Lake et al. characterize their position as "agnostic with regards to the origins of the key ingredients" (sect. 4, para. 2) of human-like intelligence. This agnosticism implicitly licenses a modeling approach in which detailed, domain-specific information can be imparted to an agent directly, an approach for which some of the authors' Bayesian Program Learning (BPL) work is emblematic. The two domains Lake and colleagues focus most upon – physics and theory of mind – are amenable to such an approach, in that these happen to be fields for which mature scientific disciplines exist. This provides unusually rich support for hand design of cognitive models. However, it is not clear that such hand design will be feasible in other more idiosyncratic domains where comparable scaffolding is unavailable. Lake et al. (2015a) were able to extend the approach to Omniglot characters by intuiting a suitable (stroke-based) model, but are we in a position to build comparably detailed domain models for such things as human dialogue and architecture? What about Japanese cuisine or ice skating? Even video-game play appears daunting, when one takes into account the vast amount of semantic knowledge that is plausibly relevant (knowledge about igloos, ice floes, cold water, polar bears, video-game levels, avatars, lives, points, and so forth). In short, it is not clear that detailed knowledge engineering will be realistically attainable in all areas we will want our agents to tackle.

Given this observation, it would appear most promising to focus our efforts on developing learning systems that can be flexibly applied across a wide range of domains, without an unattainable overhead in terms of a priori knowledge. Encouraging this view, the recent machine learning literature offers many examples of learning systems conquering tasks that had long eluded more hand-crafted approaches, including object recognition, speech recognition, speech generation, language translation, and (significantly) game play (Silver et al. 2016). In many cases, such successes have depended on large amounts of training data, and have implemented an essentially model-free approach. However, a growing volume of work suggests that flexible, domain-general learning can also be successful on tasks where training data are scarcer and where model-based inference is important.

For example, Rezende and colleagues (2016) reported a deep generative model that produces plausible novel instances of Omniglot characters after one presentation of a model character, going a significant distance toward answering Lake's "Character Challenge." Lake et al. call attention to this model's "need for extensive pre-training." However, it is not clear why their pre-installed model is to be preferred over knowledge acquired through pre-training. In weighing this point, it is important to note that the human modeler, to furnish the BPL architecture with its "start-up software," must draw on his or her own large volume of prior experience. In this sense, the resulting BPL model is dependent on the human designer's own "pre-training."

A more significant aspect of the Rezende model is that it can be applied without change to very different domains, as Rezende and colleagues (2016) demonstrate through experiments on human facial images. This flexibility is one hallmark of an autonomous learning system, and contrasts with the more purpose-built flavor of the BPL approach, which relies on irreducible primitives with domain-specific content (e.g., the strokes in Lake's Omniglot model). Furthermore, a range of recent work with deep generative models (e.g. van den Oord 2016; Ranzato et al. 2016) indicates that they can identify quite rich structure, increasingly avoiding silly mistakes like those highlighted in Lake et al.'s Figure 6.

Importantly, a learning-centered approach does not prevent us from endowing learning systems with some forms of *a priori* knowledge. Indeed, the current resurgence in neural network research was triggered largely by work that does just this, for example, by building an assumption of translational invariance into the weight matrix of image classification networks (Krizhevsky et al. 2012a). The same strategy can be taken to endow learning systems with assumptions about compositional and causal structure, yielding architectures that learn efficiently about the dynamics of physical systems, and even generalize to previously unseen numbers of objects (Battaglia et al. 2016), another challenge problem highlighted by Lake et al. In such cases, however, the inbuilt knowledge takes a highly generic form, leaving wide scope for learning to absorb domain-specific structure (see also Eslami et al 2016; Raposo et al. 2017; Reed and de Freitas 2016).

Under the approach we advocate, high-level prior knowledge and learning biases can be installed not only at the level of representational structure, but also through larger-scale architectural and algorithmic factors, such as attentional filtering (Eslami et al. 2016), intrinsic motivation mechanisms (Bellemare et al. 2016), and episodic learning (Blundell et al. 2016). Recently developed architectures for memory storage (e.g., Graves et al. 2016) offer a critical example. Lake et al. describe neural networks as implementing "learning as a process of gradual adjustment of connection strengths." However, recent work has introduced a number of architectures within which learning depends on rapid storage mechanisms, independent of connection-weight changes (Duan et al. 2016; Graves et al. 2016; Wang et al. 2017; Vinyals et al. 2016). Indeed, such mechanisms have even been applied to one-shot classification of Omniglot characters (Santoro et al., 2016) and Atari video game play (Blundell et al. 2016). Furthermore, the connection-weight changes that do occur in such models can serve in part to support learning-to-learn (Duan et al. 2016; Graves et al. 2016; Ravi and Larochelle 2017; Vinyals et al. 2016; Wang et al. 2017), another of Lake et al.'s key ingredients for human-like intelligence. As recent work has shown (Andrychowicz et al. 2016; Denil et al. 2016; Duan et al. 2016; Hochreiter et al. 2001; Santoro et al. 2016; Wang et al. 2017), this learning-to-learn mechanism can allow agents to adapt rapidly to new problems, providing a novel route to install prior knowledge through learning, rather than by hand. Learning to learn enables us to learn a neural network agent over a long time. This network, however, is trained to be good at learning rapidly from few examples, regardless of what those examples might be. So, although the meta-learning process might be slow, the product is a neural network agent that can learn to harness a few data points to carry out numerous tasks, including imitation, inference, task specialization, and prediction.

Another reason why we believe it may be advantageous to autonomously learn internal models is that such models can be shaped directly by specific, concrete tasks. A model is valuable not because it veridically captures some ground truth, but because it can be efficiently leveraged to support adaptive behavior. Just as Newtonian mechanics is sufficient for explaining many everyday phenomena, yet too crude to be useful to particle physicists and cosmologists, an agent's models should be calibrated to its tasks. This is essential for models to scale to real-world complexity, because it is usually too expensive, or even impossible, for a system to acquire and work with extremely fine-grained models of the world (Botvinick & Weinstein 2015; Silver et al. 2017). Of course, a good model of the world should be applicable across a range of task conditions, even ones that have not been previously encountered. However, this simply implies that models should be calibrated not only to individual tasks, but also to the distribution of tasks – inferred through experience or evolution – that is likely to arise in practice.

Finally, in addition to the importance of model building, it is important to recognize that real autonomy also depends on control functions, the processes that leverage models to make actual decisions. An autonomous agent needs good models, but it also needs to know how to make use of them (Botvinick & Cohen 2014), especially in settings where task goals may vary over time. This point also favors a learning and agent-based approach, because it allows control structures to co-evolve with internal models, maximizing their compatibility. Though efforts to capitalize on these advantages in practice are only in their infancy, recent work from Hamrick and colleagues (2017), which simultaneously trained an internal model and a corresponding set of control functions, provides a case study of how this might work.

Our comments here, like the target article, have focused on model-based cognition. However, an aside on model-free methods is warranted. Lake et al. describe model-free methods as providing peripheral support for model-based approaches. However, there is abundant evidence that model-free mechanisms play a pervasive role in human learning and decision making (Kahneman 2011). Furthermore, the dramatic recent successes of model-free learning in areas such as game play, navigation, and robotics suggest that it may constitute a first-class, independently valuable approach for machine learning. Lake et al. call attention to the heavy data demands of model-free learning, as reflected in DQN learning curves. However, even since the initial report on DQN (Mnih et al. 2015), techniques have been developed that significantly reduce the data requirements of this and related model-free learning methods, including prioritized memory replay (Schaul et al. 2016), improved exploration methods (Bellemare et al. 2016), and techniques for episodic reinforcement learning (Blundell et al. 2016). Given the pace of such advances, it may be premature to relegate model-free methods to a merely supporting role.

To conclude, despite the differences we have focused on here, we agree strongly with Lake et al. that human-like intelligence depends at least in part on richly structured internal models. Our approach to building human-like intelligence can be summarized as a commitment to developing autonomous agents: agents that shoulder the burden of building their own models and arriving at their own procedures for leveraging them. Autonomy, in this sense, confers a capacity to build economical task-sensitive internal models, and to adapt flexibly to diverse circumstances, while avoiding a dependence on detailed, domain-specific prior information. A key challenge in pursuing greater autonomy is the need to find more efficient means of extracting knowledge from potentially limited data. But recent work on memory, exploration, compositional representation, and processing architectures, provides grounds for optimism. In fairness, the authors of the target article have also offered, in other work, some indication of how their approach might be elaborated to support greater agent autonomy (Lake et al. 2016). We may therefore be following slowly converging paths. On a final note, it is worth pointing out that as our agents gain in autonomy, the opportunity increasingly arises for us to obtain new insights from what they themselves discover. In this way, the pursuit of agent autonomy carries the potential to transform the current AI landscape, revealing new paths toward human-like intelligence.

Digging deeper on "deep" learning: A computational ecology approach

doi:10.1017/S0140525X1700005X, e256

Massimo Buscema[a,b] and Pier Luigi Sacco[c,d]

[a]*Semeion Research Center, 00128 Rome, Italy;* [b]*University of Colorado at Denver, Denver, CO 80217;* [c]*IULM University of Milan, 20143 Milan, Italy; and* [d]*Harvard University Department of Romance Languages and Literatures, Cambridge, MA 02138.*

m.buscema@semeion.it. pierluigi.sacco@iulm.it
pierluigi@metalab.harvard.edu pierluigi_sacco@fas.harvard.edu
www.semeion.it www.researchgate.net/profile/Massimo_Buscema
www.researchgate.net/profile/Pier_Sacco

Abstract: We propose an alternative approach to "deep" learning that is based on computational ecologies of structurally diverse artificial neural networks, and on dynamic associative memory responses to stimuli. Rather than focusing on massive computation of many different examples of a single situation, we opt for model-based learning and adaptive flexibility. Cross-fertilization of learning processes across multiple domains is the fundamental feature of human intelligence that must inform "new" artificial intelligence.

In *The Society of Mind*, Minsky (1986) argued that the human brain is more similar to a complex society of diverse neural networks, than to a large, single one. The current theoretical mainstream in "deep" (artificial neural network [ANN]-based) learning leans in the opposite direction: building large ANNs with many layers of hidden units, relying more on computational power than on reverse engineering of brain functioning (Bengio 2009). The distinctive structural feature of the human brain is its synthesis of uniformity and diversity. Although the structure and functioning of neurons are uniform across the brain and across humans, the structure and evolution of neural connections make every human subject unique. Moreover, the mode of functioning of the left versus right hemisphere of the brain seems distinctively different (Gazzaniga 2004). If we do not wonder about this homogeneity of components that results in a diversity of functions, we cannot understand the computational design principles of the brain, or make sense of the variety of "constitutional arrangements" in the governance of neural interactions at various levels – "monarchic" in some cases, "democratic" or "federative" in others.

In an environment characterized by considerable stimulus variability, a biological machine that responds by combining two different principles (as embodied in its two hemispheres) has a better chance of devising solutions that can flexibly adapt to circumstances, and even anticipate singular events. The two hemispheres seem to follow two opposite criteria: an analogical-intuitive one, gradient descent-like, and a digital-rational one, vector quantization-like. The former aims at anticipating and understanding sudden environmental changes – the "black swans." The latter extrapolates trends from (currently classified as) familiar contexts and situations. These two criteria are conceptually orthogonal and, therefore, span a very rich space of cognitive functioning through their complex cooperation. On the other hand, the Bayesian approach advocated by the authors to complement the current "deep" learning agenda is useful only to simulate the functioning of the left-brain hemisphere.

The best way to capture these structural features is to imagine the brain as a society of agents (Minsky 1986), very heterogeneous and communicating through their common neural base by means of shared protocols, much like the Internet. The brain, as a highly functionally bio-diverse computational ecology, may therefore extract, from a large volume of external data, limited meaningful subsets (small data sets), to generate a variety of possible responses to these data sets and to learn from these very responses. This logic is antithetical to the mainstream notion of "deep learning" and of the consequential "big data" philosophy of processing large volumes of data to generate a few, "static" (i.e., very domain specific) responses – and which could, perhaps, more appropriately be called "fat" learning. Such dichotomy clearly echoes the tension between model-based learning and pattern recognition highlighted by the authors of the target article. Teaching a single, large, neural network how to associate an output to a certain input through millions of examples of a single situation is an exercise in brute force. It would be much more effective, in our view, to train a whole population of

"deep" ANNs, mathematically very different from one another, on the same problem and to filter their results by means of a Meta-Net (Buscema 1998; Buscema et al. 2010; 2013) that ignores their specific architectures, in terms of both prediction performance and biological plausibility.

We can therefore sum up the main tenets of our approach as follows:

1. There is extreme diversity in the architectures, logical principles, and mathematical structures of the deployed ANNs.
2. "parliament" is created whereby each ANN proposes its solution to each case, in view of its past track record for similar occurrences.
3. There is dynamic negotiation among the various hypotheses: The solution proposal of an ANN and its reputation re-enter as inputs for the other ANNs, until the ANN assembly reaches a consensus.
4. Another highly diverse pool of ANNs learns the whole dynamic process generated by the previous negotiation.

Responding to a pattern with a dynamic process rather than with a single output is much closer to the actual functioning of the human brain than associating a single output in a very domain-specific way, however nonlinear. Associative memory is a fundamental component of human intelligence: It is a cognitive morphing that connects apparently diverse experiences such as a lightning bolt and the fracture of a window pane. Human intelligence is a prediction engine working on hypotheses, generated from a relatively small database and constantly verified through sequential sampling: a cycle of perception, prediction, validation, and modification. Novelties, or changes in an already known environmental scene, will command immediate attention. Pattern recognition, therefore, is but the first step in understanding human intelligence. The next step should be building machines that generate dynamic responses to stimuli, that is, behave as dynamic associative memories (Buscema 1995; 1998; 2013; Buscema et al. 2015). The very same associative process generated by the machine, in addition to interacting with itself and the external stimuli, must itself become the object of learning: This is learning-to-learn in its fuller meaning. In this way, the artificial intelligence frontier moves from pattern recognition to recognition of pattern transformations – learning the topology used by the brain to connect environmental scenes. Analyzing the cause-effect links within these internal processes provides the basis to identify meaningful rules of folk psychology or cognitive biases: A pound of feathers may be judged lighter than a pound of lead only in a thought process where feathers are associated with lightness. The meta-analysis of the connections generated by a mind may yield physically absurd, but psychologically consistent, associations.

An approach based on ecologies of computational diversity and dynamic brain associations seems to us the most promising route to a model-based learning paradigm that capitalizes on our knowledge of the brain's computational potential. And this also means allowing for mental disturbances, hallucinations, or delirium. A "deep" machine that cannot reproduce a dissociated brain is just not intelligent enough, and if it merely maximizes IQ, it is, in a sense, "dumb." A system that can also contemplate stupidity or craziness is the real challenge of the "new" artificial intelligence.

Back to the future: The return of cognitive functionalism

doi:10.1017/S0140525X17000061, e257

Leyla Roskan Çağlar and Stephen José Hanson

Psychology Department, Rutgers University Brain Imaging Center (RUBIC), Rutgers University, Newark, NJ 07102.

lcaglar@psychology.rutgers.edu jose@rubic.rutgers.edu
https://leylaroksancaglar.github.io/
http://nwkpsych.rutgers.edu/~jose/

Abstract: The claims that learning systems must build causal models and provide explanations of their inferences are not new, and advocate a cognitive functionalism for artificial intelligence. This view conflates the relationships between implicit and explicit knowledge representation. We present recent evidence that neural networks do engage in model building, which is implicit, and cannot be dissociated from the learning process.

The neural network revolution occurred more than 30 years ago, stirring intense debate over what neural networks (NNs) can and cannot learn and represent. Much of the target article resurrects these earlier concerns, but in the context of the latest NN revolution, spearheaded by an algorithm that was known, but failed because of scale and computational power, namely, deep learning (DL).

Claims that learning systems must build causal models and provide explanations of their inferences are not new (DeJong 1986; Lenat 1995; Mitchell 1986), nor have they been proven successful. Advocating the idea that artificial intelligence (AI) systems need commonsense knowledge, ambitious projects such as "Cyc" (Lenat 1990) created hand-crafted and labor-intensive knowledge bases, combined with an inference engine to derive answers in the form of explicit knowledge. Despite feeding a large but finite number of factual assertions and explicit rules into such systems, the desired human-like performance was never accomplished. Other explanation-based and expert systems (e.g., WordNet [Miller 1990]) proved useful in some applied domains, but were equally unable to solve the problem of AI. At the essence of such projects lies the idea of "cognitive functionalism." Proposing that mental states are functional states determined and individuated by their causal relations to other mental states and behaviors, it suggests that mental states are programmable with explicitly determined representational structures (Fodor, 1981; Hayes 1974; McCarthy & Hayes 1969; Putnam 1967). Such a view stresses the importance of "formalizing concepts of causality, ability, and knowledge" to create "a computer program that decides what to do by inferring in a formal language that a certain strategy will achieve its assigned goal" (McCarthy & Hayes, 1969, p. 1). Lake et al.'s appeal to causal mechanisms and their need for explicit model representations is closely related to this cognitive functionalism, which had been put forth as a set of principles by many founders of the AI field (Hayes 1974; McCarthy 1959; McCarthy & Hayes 1969; Newell & Simon, 1956).

One important shortcoming of cognitive functionalism is its failure to acknowledge that the same behavior/function may be caused by different representations and mechanisms (Block 1978; Hanson 1995). Consequently, the problem with this proposition that knowledge within a learning system must be explicit is that it conflates the relationship between implicit knowledge and explicit knowledge and their representations. The ability to throw a low hanging fast ball would be difficult, if not impossible, to encode as a series of rules. However, this type of implicit knowledge can indeed be captured in a neural network, simply by having it learn from an analog perception–action system and a series of ball throws – all while also having the ability to represent rule-based knowledge (Horgan & Tienson 1996). This associative versus rule learning debate, referred to in this article as "pattern recognition" versus "model building," was shown a number of times to be a meaningless dichotomy (Hanson & Burr 1990; Hanson et al. 2002; Prasada & Pinker 1993).

Although we agree with Lake et al. that "model building" is indeed an important component of any AI system, we do not agree that NNs merely recognize patterns and lack the ability to build models. Our disagreement arises from the presumption that "a model must include explicit representations of objects, identity and relations" (Lake et al. 2016, pp. 38–39). Rather than being explicit or absent altogether, model representation is implicit in NNs. Investigating implicitly learned models is

somewhat more challenging, but work on learning dynamics and learning functions with respect to their relationship to representations provides insights into these implicit models (Caglar & Hanson 2016; Cleeremans 1993; Hanson & Burr 1990; Metcalf et al. 1992; Saxe et al. 2014).

Recent work has shown that in DL, the internal structure, or "model," accumulates at later layers, and is effectively constructing "scaffolds" over the learning process that are then used to train subsequent layers (Caglar & Hanson 2016; Saxe 2013). These learning dynamics can be investigated through analysis of the learning curves and the internal representations resultant in the hidden units. Analysis of the learning curves of NNs with different architectures reveals that merely adding depth to a NN results in different learning dynamics and representational structures, which do not require explicit preprogramming or pre-training (Caglar & Hanson 2016). In fact, the shape of the learning curves for single-layer NNs and for multilayered DLs are qualitatively different, with the former fitting a negative exponential function ("associative") and the latter fitting a hyperbolic function ("accumulative"). This type of structured learning, consistent with the shape of the learning curves, can be shown to be equivalent to the "learning-to-learn" component suggested by the authors. Appearing across different layers of the NNs, it also satisfies the need for "learning-to-learn to occur at multiple levels of the hierarchical generative process" (Lake et al., sect. 4.2.3, para. 5).

Furthermore, in category learning tasks with DLs, the internal representation of the hidden units shows that it creates prototype-like representations at each layer of the network (Caglar & Hanson 2016). These higher-level representations are the result of concept learning from exemplars, and go far beyond simple pattern recognition. Additionally, the plateau characteristic of the hyperbolic learning curves provides evidence for rapid learning, as well as one-shot learning once this kind of implicit conceptual representation has been formed over some subset of exemplars (similar to a "prior") (Saxe 2014). Longstanding investigation in the learning theory literature proposes that the hyperbolic learning curve of DLs is also the shape that best describes human learning (Mazur & Hastie 1978; Thurstone 1919), thereby suggesting that the learning mechanisms of DLs and humans might be more similar than thought (Hanso et al., in preparation).

Taken together, the analysis of learning curves and internal representations of hidden units indicates that NNs do in fact build models and create representational structures. However, these models are implicitly built into the learning process and cannot be explicitly dissociated from it. Exploiting the rich information of the stimulus and its context, the learning process creates models and shapes representational structures without the need for explicit preprogramming.

Theories or fragments?

doi:10.1017/S0140525X17000073, e258

Nick Chater[a] and Mike Oaksford[b]

[a]*Behavioural Science Group, Warwick Business School, University of Warwick, Coventry CV4 7AL, United Kingdom;* [b]*Department of Psychological Sciences, Birkbeck, University of London, London WC1E 7HX, United Kingdom.*

Nick.Chater@wbs.ac.uk m.oaksford@bbk.ac.uk
http://www.wbs.ac.uk/about/person/nick-chater/
http://www.bbk.ac.uk/psychology/our-staff/mike-oaksford

Abstract: Lake et al. argue persuasively that modelling human-like intelligence requires flexible, compositional representations in order to embody world knowledge. But human knowledge is too sparse and self-contradictory to be embedded in "intuitive theories." We argue, instead, that knowledge is grounded in exemplar-based learning and generalization, combined with high flexible generalization, a viewpoint compatible both with non-parametric Bayesian modelling and with sub-symbolic methods such as neural networks.

Lake et al. make a powerful case that modelling human-like intelligence depends on highly flexible, compositional representations, to embody world knowledge. But will such knowledge really be embedded in "intuitive theories" of physics or psychology? This commentary argues that there is a paradox at the heart of the "intuitive theory" viewpoint, that has bedevilled analytic philosophy and symbolic artificial intelligence: human knowledge is both (1) extremely sparse and (2) self-contradictory (e.g., Oaksford & Chater 1991).

The sparseness of intuitive knowledge is exemplified in Rozenblit and Keil's (2002) discussion of the "illusion of explanatory depth." We have the feeling that we understand how a crossbow works, how a fridge stays cold, or how electricity flows around the house. Yet, when pressed, few of us can provide much more than sketchy and incoherent fragments of explanation. Therefore, our causal models of the physical world appear shallow. The sparseness of intuitive psychology seems at least as striking. Indeed, our explanations of our own and others' behavior often appear to be highly ad hoc (Nisbett & Ross 1980).

Moreover, our physical and psychological intuitions are also self-contradictory. The foundations of physics and rational choice theory have consistently shown how remarkably few axioms (e.g., the laws of thermodynamics, the axioms of decision theory) completely fix a considerable body of theory. Yet our intuitions about heat and work, or probability and utility, are vastly richer and more amorphous, and cannot be captured in any consistent system (e.g., some of our intuitions may imply our axioms, but others will contradict them). Indeed, contradictions can also be evident even in apparent innocuous mathematical or logical assumptions (as illustrated by Russell's paradox, which unexpectedly exposed a contradiction in Frege's attempted logical foundation for mathematics [Irvine & Deutsch 2016]).

The sparse and contradictory nature of our intuition explains why explicit theorizing requires continually ironing out contradictions, making vague concepts precise, and radically distorting or replacing existing concepts. And the lesson of two and half millennia of philosophy is arguable, that clarifying even the most basic concepts, such as "object" or "the good" can be entirely intractable, a lesson re-learned in symbolic artificial intelligence. In any case, the raw materials for this endeavor – our disparate intuitions – may not be properly viewed as organized as theories at all.

If this is so, how do we interact so successfully in the physical and social worlds? We have *experience*, whether direct, or by observation or instruction – of crossbows, fridges, and electricity – to be able to interact with them in familiar ways. Indeed, our ability to make sense of new physical situations often appears to involve creative extrapolation from familiar examples: for example, assuming that heavy objects will fall faster than light objects, even in a vacuum, or where air resistance can be neglected. Similarly, we have a vast repertoire of experience of human interaction, from which we can generalize to new interactions. Generalization from such experiences, to deal with new cases, can be extremely flexible and abstract (Hofstadter 2001). For example, the perceptual system uses astonishing ingenuity to construct complex percepts (e.g., human faces) from highly impoverished signals (e.g., Hoffman 2000; Rock 1983) or to interpret art (Gombrich 1960).

We suspect that the growth and operation of cognition are more closely analogous to case law than to scientific theory. Each new case is decided by reference to the facts of that present case and to ingenious and open-ended links to precedents from past cases; and the history of cases creates an intellectual tradition that is only locally coherent, often ill-defined, but surprisingly effective in dealing with a complex and ever-changing world. In short, knowledge has the form of a loosely interlinked history of reusable fragments, each building on the last, rather than being organized into anything resembling a scientific theory.

Recent work on construction-based approaches to language exemplify this viewpoint in the context of linguistics (e.g.,

Goldberg 1995). Rather than seeing language as generated by a theory (a formally specified grammar), and the acquisition of language as the fine-tuning of that theory, such approaches see language as a tradition, where each new language processing episode, like a new legal case, is dealt with by reference to past instances (Christiansen & Chater 2016). In both law and language (see Blackburn 1984), there will be a tendency to impose local coherence across similar instances, but there will typically be no globally coherent theory from which all cases can be generated.

Case instance or exemplar-based theorizing has been widespread in the cognitive sciences (e.g., Kolodner 1993; Logan 1988; Medin & Shaffer 1978). Exploring how creative extensions of past experience can be used to deal with new experience (presumably by processes of analogy and metaphor rather than deductive theorizing from basic principles) provides an exciting challenge for artificial intelligence, whether from a non-parametric Bayesian standpoint or a neural network perspective, and is likely to require drawing on the strengths of both.

ACKNOWLEDGMENTS

N.C. was supported by ERC Grant 295917-RATIONALITY, the ESRC Network for Integrated Behavioural Science (Grant ES/K002201/1), the Leverhulme Trust (Grant RP2012-V-022], and Research Councils UK Grant EP/K039830/1.

The humanness of artificial non-normative personalities

doi:10.1017/S0140525X17000085, e259

Kevin B. Clark

Research and Development Service, Veterans Affairs Greater Los Angeles Healthcare System, Los Angeles, CA 90073; California NanoSystems Institute, University of California at Los Angeles, Los Angeles, CA 90095; Extreme Science and Engineering Discovery Environment (XSEDE), National Center for Supercomputing Applications, University of Illinois at Urbana–Champaign, Urbana, IL 61801; Biological Collaborative Research Environment (BioCoRE), Theoretical and Computational Biophysics Group, NIH Center for Macromolecular Modeling and Bioinformatics, Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign, Urbana, IL 61801.

kbclarkphd@yahoo.com

www.linkedin.com/pub/kevin-clark/58/67/19a

Abstract: Technoscientific ambitions for perfecting human-like machines, by advancing state-of-the-art neuromorphic architectures and cognitive computing, may end in ironic regret without pondering the humanness of fallible artificial non-normative personalities. Self-organizing artificial personalities individualize machine performance and identity through fuzzy conscientiousness, emotionality, extraversion/introversion, and other traits, rendering insights into technology-assisted human evolution, robot ethology/pedagogy, and best practices against unwanted autonomous machine behavior.

Within a modern framework of promising, yet still inadequate state-of-the-art artificial intelligence, Lake et al. construct an optimistic, ambitious plan for innovating truer representative neural network-inspired machine emulations of human consciousness and cognition, elusive pinnacle goals of many cognitive, semiotic, and cybernetic scientists (Cardon 2006; Clark 2012; 2014; 2015; Kaipa et al. 2010; McShea 2013). Their machine learning-based agenda, possibly requiring future generations of pioneering hybrid neuromorphic computing architectures and other sorts of technologies to be fully attained (Lande 1998; Indiveri & Liu 2015; Schuller & Stevens 2015), relies on implementing sets of data-/theory-established "core ingredients" typical of natural human intelligence and development (cf. Bengio 2016; Meltzoff et al. 2009; Thomaz & Cakmak 2013; Weigmann 2006). Such core ingredients, including (1) intuitive causal physics and psychology, (2) compositionality and learning-to-learn, and (3) fast efficient real-time gradient-descent deep learning and thinking, will certainly endow contemporary state-of-the-art machines with greater human-like cognitive qualities. But, in Lake et al.'s efforts to create a standard of human-like machine learning and thinking, they awkwardly, and perhaps ironically, erect barriers to realizing ideal human simulation by ignoring what is also very human – variations in cognitive-emotional neural network structure and function capable of giving rise to non-normative (or unique) personalities and, therefore, dynamic expression of human intelligences and identities (Clark 2012; 2015; in press-a; in press-b; in press-c). Moreover, this same, somewhat counterintuitive, problem in the authors' otherwise rational approach dangerously leaves unaddressed the major ethical and security issues of "free-willed" personified artificial sentient agents, often popularized by fantasists and futurists alike (Bostrom 2014; Briegel 2012; Davies 2016; Fung 2015).

Classic interpretations of perfect humanness arising from the fallibility of humans (e.g., Clark 2012; Nisbett & Ross 1980; Parker & McKinney 1999; Wolfram 2002) appreciably impact the technical feasibility and socio-cultural significance of building and deploying human-emulating personified machines under both nonsocial and social constraints. Humans, as do all sentient biological entities, fall within a fuzzy organizational and operational template that bounds emergence of phylogenic, ontogenic, and sociogenic individuality (cf. Fogel & Fogel 1995; Romanes 1884). Extreme selected variations in individuality, embodied here by modifiable personality and its link to mind, can greatly elevate or diminish human expression, depending on pressures of situational contexts. Examples may include the presence or absence of resoluteness, daring, agile deliberation, creativity, and meticulousness essential to achieving matchless, unconventional artistic and scientific accomplishments. Amid even further examples, they may also include the presence or absence of empathy, morality, or ethics in response to severe human plight and need. Regardless, to completely simulate the range of human intelligence, particularly solitary to sociable and selfish to selfless tendencies critical for now-nascent social-like human-machine and machine-machine interactions, scientists and technologists must account for, and better understand, personality trait formation and development in autonomous artificial technologies (Cardon 2006; Clark 2012; 2015; Kaipa et al. 2010; McShea 2013). These kinds of undertakings will help yield desirable insights into the evolution of technology-augmented human nature and, perhaps more importantly, will inform best practices when establishing advisable failsafe contingencies against unwanted serendipitous or designed human-like machine behavior.

Notably, besides their described usefulness for modeling intended artificial cognitive faculties, Lake et al.'s core ingredients provide systematic concepts and guidelines necessary to begin approximating human-like machine personalities, and to probe genuine ethological, ecological, and evolutionary consequences of those personalities for both humans and machines. However, similar reported strategies for machine architectures, algorithms, and performance demonstrate only marginal success when used as protocols to reach nearer cognitive-emotional humanness in trending social robot archetypes (Arbib & Fellous 2004; Asada 2015; Berdahl 2010; Di & Wu 2015; Han et al. 2013; Hiolle et al. 2014; Kaipa et al. 2010; McShea 2013; Read et al. 2010; Thomaz & Cakmak 2013; Wallach et al. 2010; Youyou et al. 2015), emphasizing serious need for improved adaptive quasi-model-free/-based neural nets, trainable distributed cognition-emotion mapping, and artificial personality trait parameterization. The best findings from such work, although far from final reduction-to-practice, arguably involve the appearance of crude or primitive machine personalities and identities from socially learned intra-/interpersonal relationships possessing cognitive-emotional valences. Valence direction and magnitude often depend on the learner machine's disposition toward response priming/contagion, social facilitation, incentive motivation, and local/stimulus enhancement of observable demonstrator behavior (i.e., human, cohort-machine, and learner-machine behavior). The resulting self-/world discovery of the learner machine,

analogous to healthy/diseased or normal/abnormal human phenomena acquired during early formative (neo)Piagetian cognitive-emotional periods (cf. Nisbett & Ross 1980; Parker & McKinney 1999; Zentall 2013), reciprocally shapes the potential humanness of reflexive/reflective machine actions through labile interval-delimited self-organizing traits consistent with natural human personalities, including, but not restricted to, conscientiousness, openness, emotional stability, agreeableness, and extraversion/introversion.

Even simplistic artificial cognitive-emotional profiles and personalities thus effect varying control over acquisition and lean of machine domain-general/-specific knowledge, perception and expression of flat or excessive machine affect, and rationality and use of inferential machine attitudes/opinions/beliefs (Arbib & Fellous 2004; Asada 2015; Berdahl 2010; Cardon 2006; Davies 2016; Di & Wu 2015; Han et al. 2013; Hiolle et al. 2014; Kaipa et al. 2010; McShea 2013; Read et al. 2010; Wallach et al. 2010; Youyou et al. 2015). And, by favoring certain artificial personality traits, such as openness, a learner machine's active and passive pedagogical experiences may be radically directed by the quality of teacher-student rapport (e.g., Thomaz & Cakmak 2013), enabling opportunities for superior nurturing and growth of distinctive, well-adjusted thoughtful machine behavior while, in part, restricting harmful rogue machine behavior, caused by impoverished learning environments and predictable pathological Gödel-type incompleteness/inconsistency for axiomatic neuropsychological systems (cf. Clark & Hassert 2013). These more-or-less philosophical considerations, along with the merits of Lake et al.'s core ingredients for emerging artificial non-normative (or unique) personalities, will bear increasing technical and sociocultural relevance as the Human Brain Project, the Blue Brain Project, and related connectome missions drive imminent neuromorphic hardware research and development toward precise mimicry of configurable/computational soft-matter variations in human nervous systems (cf. Calimera et al. 2013).

Children begin with the same start-up software, but their software updates are cultural

doi:10.1017/S0140525X17000097, e260

Jennifer M. Clegg and Kathleen H. Corriveau

Boston University School of Education, Boston, MA 02215.

jclegg@bu.edu kcorriv@bu.edu

www.jennifermclegg.com www.bu.edu/learninglab

Abstract: We propose that early in ontogeny, children's core cognitive abilities are shaped by culturally dependent "software updates." The role of sociocultural inputs in the development of children's learning is largely missing from Lake et al.'s discussion of the development of human-like artificial intelligence, but its inclusion would help move research even closer to machines that can learn and think like humans.

Lake et al. draw from research in both artificial intelligence (AI) and cognitive development to suggest a set of core abilities necessary for building machines that think and learn like humans. We share the authors' view that children have a set of core cognitive abilities for learning and that these abilities should guide development in AI research. We also agree with the authors' focus on findings from *theory theory* research and their characterization of its principles as "developmental start-up software" that is adapted later in ontogeny for social learning. What is missing from this discussion, however, is the recognition that children's developmental start-up software is shaped by their culture-specific social environment. Children's early and ontogenetically persistent experiences with their cultural environment affect what learning "programs" children develop and have access to, particularly in the case of social learning.

Research suggests that from early infancy, children display a core set of abilities that shape their reasoning about the world, including reasoning about both inanimate objects (intuitive physics [e.g., Spelke 1990]) and animate social beings (intuitive psychology [e.g., Dennett 1987; Meltzoff & Moore 1995]). Although the early onset of these abilities provides evidence that they may be universal, little research has examined their development in non-WEIRD (Western educated industrialized rich democratic) (Henrich et al. 2010) cultures (Legare & Harris, 2016). Moreover, research that has examined children's intuitive theories in different cultural settings has suggested the potential for both cross-cultural continuity and variation in their development. Take, for example, the development of children's theory of mind, a component of intuitive psychology. A large collection of research comparing the development of children's understanding of false belief in the United States, China, and Iran indicates that although typically developing children in all cultures show an improvement in false belief understanding over the course of ontogeny, the timing of this improvement differs widely—and such variability is potentially related to different sociocultural inputs (Davoodi et al. 2016; Liu et al. 2008; Shahaeian et al. 2011). Thus, children's social environments may be shaping the development of these core abilities, "reprogramming" and updating their developmental start-up software.

To illustrate why considering the principles derived from *theory theory* are important for guiding AI development, Lake et al. point to AI's lack of human-like intuitive psychology as a key reason for why humans outperform AI. In their discussion of humans' superior performance in the Frostbite challenge, the authors highlight humans' ability to build on skills gained through the observation of an expert player,which requires reasoning about the expert player's mental state. AI can also draw on observations of expert players, but requires substantially greater input to achieve similar levels of performance. Humans' intuitive psychology and their corresponding ability to reason about others' mental states is just one element of why humans may be outperforming computers in this task. This situation also draws on humans' ability to learn by observing others and, like the development of false-belief understanding, children's ability to learn through observation as well as through verbal testimony, which is heavily influenced by sociocultural inputs (Harris 2012).

Culturally specific ethno-theories of how children learn (Clegg et al. 2017; Corriveau et al. 2013; Harkness et al. 2007; Super & Harkness 2002) and the learning opportunities to which children have access (Kline 2015; Rogoff 2003) shape their ability to learn through observation. As early as late infancy, sociocultural inputs such as how parents direct children's attention, or the typical structure of parent-child interaction, may lead to differences in the way children attend to events for the purpose of observational learning (Chavajay & Rogoff 1999). By pre-school, children from non-WEIRD cultures where observational learning is expected and socialized outperform children from WEIRD cultures in observational learning tasks (Correa-Chávez & Rogoff 2009; Mejía-Arauz et al. 2005). Recent research also suggests that children from different cultural backgrounds attend to different types of information when engaging in observational learning. For example, Chinese-American children are more sensitive to whether there is consensus about a behavior or information than Euro-American children (Corriveau & Harris 2010; Corriveau et al. 2013; DiYanni et al. 2015). Such cultural differences in attending to social information in observational learning situations persist into adulthood (Mesoudi et al. 2015). Therefore, although the developmental start-up software children begin with may be universal, early in development, children's "software updates" may be culturally dependent. Over time, these updates may even result in distinct operating systems.

The flexibility of children's core cognitive abilities to be shaped by sociocultural input is what makes human learning unique (Henrich 2015). The role of this input is largely missing from Lake et al.'s discussion of creating human-like AI, but its inclusion would help move research even closer to machines that can learn and think like humans.

Deep-learning networks and the functional architecture of executive control

doi:10.1017/S0140525X17000103, e261

Richard P. Cooper

Centre for Cognition, Computation and Modelling, Department of Psychological Sciences, Birkbeck, University of London, London WC1E 7HX, United Kingdom.

R.Cooper@bbk.ac.uk

http://www.bbk.ac.uk/psychology/our-staff/richard-cooper

Abstract: Lake et al. underrate both the promise and the limitations of contemporary deep learning techniques. The promise lies in combining those techniques with broad multisensory training as experienced by infants and children. The limitations lie in the need for such systems to possess functional subsystems that generate, monitor, and switch goals and strategies in the absence of human intervention.

Lake et al. present a credible case for why natural intelligence requires the construction of compositional, causal generative models that incorporate intuitive psychology and physics. Several of their arguments (e.g., for compositionality and theory construction and for learning from limited experience) echo arguments that have been made throughout the history of cognitive science (e.g., Fodor & Pylyshyn 1988). Indeed, in the context of Lake et al.'s criticisms, the closing remarks of Fodor and Pylyshyn's seminal critique of 1980s-style connectionism make sobering reading: "some learning is a kind of theory construction.... We seem to remember having been through this argument before. We find ourselves with a gnawing sense of deja vu" (1988, p. 69). It would appear that cognitive science has advanced little in the last 30 years with respect to the underlying debates.

Yet Lake et al. underrate both the promise and the limitations of contemporary deep learning (DL) techniques with respect to natural and artificial intelligence. Although contemporary DL approaches to, say, learning and playing Atari games undoubtedly employ psychologically unrealistic training regimes, and are undoubtedly inflexible with respect to changes to the reward/goal structure, to fixate on these limitations overlooks the promise of such approaches. It is clear the DL nets are not normally trained with anything like the experiences had by the developing child, whose learning is based on broad, multisensory experience and is cumulative, with new motor and cognitive skills building on old (Vygotsky 1978). Until DL nets are trained in this way, it is not reasonable to critique the outcomes of such approaches for unrealistic training regimes of, for example, "almost 500 times as much experience as the human received" (target article, sect. 3.2, para. 4). That 500 times as much experience neglects the prior experience that the human brought to the task. DL networks. as currently organised, require that much experience precisely because they bring nothing but a learning algorithm to the task.

A more critical question is whether contemporary DL approaches might, with appropriate training, be able to acquire intuitive physics – the kind of thing an infant learns through his or her earliest interactions with the world (that there are solids and liquids, and that solids can be grasped and that some can be picked up, but that they fall when dropped, etc.). Similarly, can DL acquire intuitive psychology through interaction with other agents? And what kind of input representations and motor abilities might allow DL networks to develop representational structures that support reuse across tasks? The promise of DL networks (and at present it remains a promise) is that, with sufficiently broad training, they may support the development of systems that capture intuitive physics and intuitive psychology. To neglect this possibility is to see the glass as half empty, rather than half full.

The suggestion is not simply that training an undifferentiated DL network with the ordered multisensory experiences of a developing child will automatically yield an agent with natural intelligence. As Lake et al. note, gains come from combining DL with reinforcement learning (RL) and Monte-Carlo Tree Search to support extended goal-directed activities (such as playing Atari games) and problem solving (as in the game of Go). These extensions are of particular interest because they parallel cognitive psychological accounts of more complex cognition. More specifically, accounts of behaviour generation and regulation have long distinguished between automatic and deliberative behaviour. Thus, the contention scheduling/supervisory system theory of Norman and Shallice (1986) proposes that one system – the contention scheduling system – controls routine, overlearned, or automatic behaviour, whereas a second system – the supervisory system – may bias or modulate the contention scheduling system in non-routine situations where deliberative control is exercised. Within this account the routine system may plausibly employ a DL-type network combined with (a hierarchical variant of) model-free reinforcement learning, whereas the non-routine system is more plausibly conceived of in terms of a model-based system (cf. Daw et al. 2005).

Viewing DL-type networks as models of the contention scheduling system suggests that their performance should be compared to those aspects of expert performance that are routinized or overlearned. From this perspective, the limits of DL-type networks are especially informative, as they indicate which cognitive functions cannot be routinized and should be properly considered as supervisory. Indeed, classical model-based RL is impoverished compared with natural intelligence. The evidence from patient and imaging studies suggests that the non-routine system is not an undifferentiated whole, as might befit a system that simply performs Monte-Carlo Tree Search. The supervisory system appears to perform a variety of functions, such as goal generation (to create one's own goals and to function in real domains outside of the laboratory), strategy generation and evaluation (to create and evaluate potential strategies that might achieve goals), monitoring (to detect when one's goals are frustrated and to thereby trigger generation of new plans/strategies or new goals), switching (to allow changing goals), response inhibition (to prevent selection of pre-potent actions which may conflict with one's high-level goals), and perhaps others. (See Shallice & Cooper [2011] for an extended review of relevant evidence and Fox et al. [2013] and Cooper [2016], for detailed suggestions for the potential organisation of higher-level modulatory systems.) These functions must also support creativity and autonomy, as expressed by naturally intelligent systems. Furthermore, "exploration" is not unguided as in the classical exploration/exploitation trade-off of RL. Natural intelligence appears to combine the largely reactive *perception-action* cycle of RL with a more active *action-perception* cycle, in which the cognitive system can act and deliberatively explore in order to test hypotheses.

To achieve natural intelligence, it is likely that a range of supervisory functions will need to be incorporated into the model-based system, or as modulators of a model-free system. Identifying the component functions and their interactions, that is, identifying the functional architecture (Newell 1990), will be critical if we are to move beyond Lake et al.'s "Character" and "Frostbite" challenges, which remain highly circumscribed tasks that draw upon limited world knowledge.

Causal generative models are just a start

doi:10.1017/S0140525X17000115, e262

Ernest Davis[a] and Gary Marcus[b,c]

[a]*Department of Computer Science, New York University, New York, NY 10012;* [b]*Uber AI Labs, San Francisco, CA 94103;* [c]*Department of Psychology, New York University, New York, NY 10012.*

davise@cs.nyu.edu Gary.marcus@nyu.edu

http://www.cs.nyu.edu/faculty/davise http://garymarcus.com/

Abstract: Human reasoning is richer than Lake et al. acknowledge, and the emphasis on theories of how images and scenes are synthesized is misleading. For example, the world knowledge used in vision presumably involves a combination of geometric, physical, and other knowledge, rather than just a causal theory of how the image was produced. In physical reasoning, a model can be a set of constraints rather than a physics engine. In intuitive psychology, many inferences proceed without detailed causal generative models. How humans reliably perform such inferences, often in the face of radically incomplete information, remains a mystery.

We entirely agree with the central thrust of the article. But a broader view of what a "model" is, is needed.

In most of the examples discussed in the target article, a "model" is a generative system that synthesizes a specified output. For example, the target article discusses a system built by Lake et al. (2015a) that learns to recognize handwritten characters from one or two examples, by modeling the sequence of strokes that produced them. The result is impressive, but the approach – identifying elements from a small class of items based on a reconstruction of how something might be generated – does not readily generalize in many other situations. Consider, for example, how one might recognize a cat, a cartoon of a cat, a painting of a cat, a marble sculpture of a cat, or a cloud that happens to look like a cat. The causal processes that generated each of these are very different; and yet a person familiar with cats will recognize any of these depictions, even if they know little of the causal processes underlying sculpture or the formation of clouds. Conversely, the differences between the causal processes that generate a cat and those that generate a dog are understood imperfectly, even by experts in developmental biology, and hardly at all by laypeople. Yet even children can readily distinguish dogs from cats. Likewise, where children learn to recognize letters significantly before they can write them at all well,[1] it seems doubtful that models of how an image is synthesized, play any necessary role in visual recognition even of letters, let alone of more complex entities. Lake et al.'s results are technically impressive, but may tell us little about object recognition in general.

The discussion of physical reasoning here, which draws on studies such as Battaglia et al. (2013), Gerstenberg et al. (2015), and Sanborn et al. (2013), may be similarly misleading. The target article argues that the cognitive processes used for human physical reasoning are "intuitive physics engines," similar to the simulators used in scientific computation and computer games. But, as we have argued elsewhere (Davis & Marcus 2014; 2016), this model of physical reasoning is much too narrow, both for AI and for cognitive modeling.

First, simulation engines require both a precise predictive theory of the domain and a geometrically and physically precise description of the situation. Human reasoners, by contrast, can deal with information that is radically incomplete. For example, if you are carrying a number of small creatures in a closed steel box, you can predict that as long as the box remains completely closed, the creatures will remain inside. This prediction can be made without knowing anything about the creatures and the way they move, without knowing the initial positions or shapes of the box or the creatures, and without knowing the trajectory of the box.

Second, simulation engines predict how a system will develop by tracing its state in detail over a sequence of closely spaced instances. For example, Battaglia et al. (2013) use an existing physics engine to model how humans reason about an unstable tower of blocks collapsing to the floor. The physics engine generates a trace of the exact positions, velocities of every block, and the forces between them, at a sequence of instants a fraction of a second apart. There is no evidence that humans routinely generate comparably detailed traces or even that they are capable of doing so. Conversely, people are capable of predicting characteristics of an end state for problems where it is impossible to predict the intermediate states in detail, as the example of the creatures in the box illustrates.

Third, there is extensive evidence that in many cases where the actual physics is simple, humans make large, systematic errors. For example, a gyroscope or a balance beam constructed of solid parts is governed by the identical physics as the falling tower of blocks studied in Battaglia et al. (2013); the physical interactions and their analysis are much simpler for these than for the tower of blocks, and the physics engine that Battaglia et al. used in their studies will handle the case of a gyroscope or a balance beam without difficulty. But here, the model is "too good" relative to humans. Human subjects often make errors in predicting the behavior of a balance beam (Siegler 1976), and most people find the behavior of a gyroscope mystifying. Neither result follows from the model.

Intuitive psychology goes even further beyond what can be explained by sorts of generative models of action choice, discussed in the target article. One's knowledge of the state of another agent's mind and one's ability to predict their action are necessarily extremely limited; nonetheless, powerful psychological reasoning can be carried out. For example, if you see a person pick up a telephone and dial, it is a good guess that they he or she is planning to talk to someone. To do so, one does not need a full causal model of whom they want to talk to, what they will say, or what their goal is in calling. In this instance (and many others), there seems to be a mismatch between the currency of generative models and the sorts of inferences that humans can readily make.

So whereas we salute Lake et al.'s interest in drawing inferences from small amounts of data, and believe as they do that rich models are essential to complex reasoning, we find their view of causal models to be too parochial. Reasoning in humans, and in general artificial intelligence, requires bringing to bear knowledge across an extraordinarily wide range of subjects, levels of abstraction, and degrees of completeness. The exclusive focus on causal generative models is unduly narrow.

NOTE

1. This may be less true with respect to Chinese and other large character sets, in which practicing drawing the characters is an effective way of memorizing them (Tan et al. 2005).

Thinking like animals or thinking like colleagues?

doi:10.1017/S0140525X17000127, e263

Daniel C. Dennett and Enoch Lambert

Center for Cognitive Studies, Tufts University, Medford, MA 02155.

daniel.dennett@tufts.edu enoch.lambert@gmail.com

http://ase.tufts.edu/cogstud/dennett/

http://ase.tufts.edu/cogstud/faculty.html

Abstract: We comment on ways in which Lake et al. advance our understanding of the machinery of intelligence and offer suggestions. The first set concerns animal-level versus human-level intelligence. The second concerns the urgent need to address ethical issues when evaluating the state of artificial intelligence.

Lake et al. present an insightful survey of the state of the art in artificial intelligence (AI) and offer persuasive proposals for feasible future steps. Their ideas of "start-up software" and tools for rapid model learning (sublinguistic "compositionality" and "learning-to-learn") help pinpoint the sources of general, flexible intelligence. Their concrete examples using the Character Challenge and Frostbite Challenge forcefully illustrate just how behaviorally effective human learning can be compared with current achievements in machine learning. Their proposal that such learning is the result of "metacognitive processes" integrating model-based and model-free learning is tantalizingly suggestive, pointing toward novel ways of explaining intelligence. So, in a sympathetic

spirit, we offer some suggestions. The first set concerns casting a wider view of *explananda* and, hence, potential *explanantia* regarding intelligence. The second set concerns the need to confront ethical concerns as AI research advances.

Lake et al.'s title speaks of "thinking like humans" but most of the features discussed—use of intuitive physics, intuitive psychology, and relying on "models"—are features of animal thinking as well. Not just apes or mammals, but also birds and octopuses and many other animals have obviously competent expectations about causal links, the reactions of predators, prey and conspecifics, and must have something like implicit models of the key features in their worlds—their affordances, to use Gibson's (1979) term.

Birds build species-typical nests they have never seen built, improving over time, and apes know a branch that is too weak to hold them. We think the authors' term *intuitive physics engine* is valuable because unlike "folk physics," which suggests a theory, it highlights the fact that neither we, nor animals in general, need to understand from the outset the basic predictive machinery we are endowed with by natural selection. We humans eventually bootstrap this behavioral competence into reflective comprehension, something more like a theory and something that is probably beyond language-less animals.

So, once sophisticated animal-level intelligence is reached, there will remain the all-important step of bridging the gap to human-level intelligence. Experiments suggest that human children differ from chimpanzees primarily with respect to social knowledge (Herrmann et al. 2007; 2010). Their unique forms of imitation and readiness to learn from teachers suggest means by which humans can accumulate and exploit an "informational commonwealth" (Kiraly et al. 2013; Sterelny 2012; 2013). This is most likely part of the story of how humans can become as intelligent as they do. But the missing part of that story remains internal mechanisms, which Lake et al. can help us focus on. Are the unique social skills developing humans deploy because of *enriched* models ("intuitive psychology" say), *novel* models (ones with principles of social emulation and articulation), or more powerful abilities to acquire and enrich models (learning-to-learn)? The answer probably appeals to some combination. But we suggest that connecting peculiarly human ways of learning from others to Lake et al.'s "learning-to-learn" mechanisms may be particularly fruitful for fleshing out the latter – and ultimately illuminating to the former.

The step up to human-style comprehension carries moral implications that are not mentioned in Lake et al.'s telling. Even the most powerful of existing AIs are intelligent tools, not colleagues, and whereas they can be epistemically authoritative (within limits we need to characterize carefully), and hence will come to be relied on more and more, they should not be granted moral authority or responsibility because they do not have skin in the game: they do not yet have interests, and *simulated* interests are not enough. We are not saying that an AI could not be created to have genuine interests, but that is down a very long road (Dennett 2017; Hurley et al. 2011). Although some promising current work suggests that genuine human consciousness depends on a fundamental architecture that would require having interests (Deacon 2012; Dennett 2013), long before that day arrives, if it ever does, we will have AIs that can communicate with natural language with their *users* (not collaborators).

How should we deal, ethically, with these pseudo-moral agents? One idea, inspired in part by recent work on self-driving cars (Pratt 2016), is that instead of letting them be autonomous, they should be definitely subordinate: co-pilots that help but do not assume responsibility for the results. We must never pass the buck to the machines, and we should take steps now to ensure that those who rely on them recognize that they are strictly liable for any harm that results from decisions *they* make with the help of their co-pilots. The studies by Dietvorst et al. (2015; 2016; see Hutson 2017) suggest that people not only tend to distrust AIs, but also *want* to exert control, and hence responsibility, over the results such AIs deliver. One way to encourage this is to establish firm policies of disclosure of all known gaps and inabilities in AIs (much like the long lists of side effects of medications). Furthermore, we should adopt the requirement that such language-using AIs must have an initiation period in which their task is to tutor users, treating them as apprentices and not giving any assistance until the user has established a clear level of expertise. Such expertise would not be in the fine details of the AIs' information, which will surely outstrip any human being's knowledge, but in the limitations of the assistance on offer and the responsibility that remains in the hands of the user. Going forward, it is time for evaluations of the state of AI to include consideration of such moral matters.

Evidence from machines that learn and think like people

doi:10.1017/S0140525X17000139, e264

Kenneth D. Forbus[a] and Dedre Gentner[b]

[a]*Department of Computer Science, Northwestern University, Evanston, IL 60208;* [b]*Department of Psychology, Northwestern University, Evanston, IL 60208.*

forbus@northwestern.edu **gentner@northwestern.edu**
http://www.cs.northwestern.edu/~forbus/
http://groups.psych.northwestern.edu/gentner/

Abstract: We agree with Lake et al.'s trenchant analysis of deep learning systems, including that they are highly brittle and that they need vastly more examples than do people. We also agree that human cognition relies heavily on structured relational representations. However, we differ in our analysis of human cognitive processing. We argue that (1) analogical comparison processes are central to human cognition; and (2) intuitive physical knowledge is captured by qualitative representations, rather than quantitative simulations.

Capturing relational capacity. We agree with Lake et al. that structured relational representations are essential for human cognition. But that raises the question of how such representations are acquired and used. There is abundant evidence from both children and adults that structure mapping (Gentner 1983) is a major route to acquiring and using knowledge. For example, physicists asked to solve a novel problem spontaneously use analogies to known systems (Clement 1988), and studies of working microbiology laboratories reveal that frequent use of analogies is a major determinant of success (Dunbar 1995). In this respect, children are indeed like little scientists. Analogical processes support children's learning of physical science (Chen & Klahr 1999; Gentner et al. 2016) and mathematics (Carey 2009; Mix 1999; Richland & Simms 2015). Analogy processes pervade everyday reasoning as well. People frequently draw inferences from analogous situations, sometimes without awareness of doing so (Day & Gentner 2007).

Moreover, computational models of structure mapping's matching, retrieval, and generalization operations have been used to simulate a wide range of phenomena, including geometric analogies, transfer learning during problem solving, and moral decision making (Forbus et al. 2017). Simulating humans on these tasks requires between 10 and 100 relations per example. This is a significant gap. Current distributed representations have difficulty handling even one or two relations.

Even visual tasks, such as character recognition, are more compactly represented by a network of relationships and objects than by an array of pixels, which is why human visual systems compute edges (Marr 1983; Palmer 1999). Further, the results from adversarial training indicate that deep learning systems do not construct human-like intermediate representations (Goodfellow et al. 2015; see also target article). In contrast, there is evidence that a structured representation approach can provide human-like visual

processing. For example, a model that combines analogy with visual processing of relational representations has achieved human-level performance on Raven's Progressive Matrices test (Lovett & Forbus 2017). Using analogy over relational representations may be a superior approach even for benchmark machine learning tasks. For example, on the link plausibility task, in which simple knowledge bases (Freebase, WordNet) are analyzed so that the plausibility of new queries can be estimated (e.g., Is Barack Obama Kenyan?), a combination of analogy and structured logistic regression achieved state-of-the-art performance, with orders of magnitude fewer training examples than distributed representation systems (Liang & Forbus 2015). Because structure mapping allows the use of relational representations, the system also provided explanations, the lack of which is a significant drawback of distributed representations.

Causality and qualitative models. Lake et al. focus on Bayesian techniques and Monte Carlo simulation as their alternative explanation for how human cognition works. We agree that statistics are important, but they are insufficient. Specifically, we argue that analogy provides exactly the sort of rapid learning and reasoning that human cognition exhibits. Analogy provides a means of transferring prior knowledge. For example, the Companion cognitive architecture can use rich relational representations and analogy to perform distant transfer. Learning games with a previously learned analogous game led to more rapid learning than learning without such an analog (Hinrichs & Forbus 2011). This and many other experiments suggest that analogy not only can explain human transfer learning, but also can provide new techniques for machine learning.

Our second major claim is that qualitative representations – not quantitative simulations – provide much of the material of our conceptual structure, especially for reasoning about causality (Forbus & Gentner 1997). Human intuitive knowledge concerns relationships such as "the higher the heat, the quicker the water will boil," not the equations of heat flow. Qualitative representations provide symbolic, relational representations of continuous properties and an account of causality organized around processes of change. They enable commonsense inferences to be made with little information, using qualitative mathematics. Decades of successful models have been built for many aspects of intuitive physics, and such models have also been used to ground scientific and engineering reasoning (Forbus 2011). Moreover, qualitative models can explain aspects of social reasoning, including blame assignment (Tomai & Forbus 2008) and moral decision making (Dehghani et al. 2008), suggesting that they are important in intuitive psychology as well.

We note two lines of qualitative reasoning results that are particularly challenging for simulation-based accounts. First, qualitative representations provide a natural way to express some aspects of natural language semantics, for example, "temperature depends on heat" (McFate & Forbus 2016). This has enabled Companions to learn causal models via reading natural language texts, thereby improving their performance in a complex strategy game (McFate et al. 2014). Second, qualitative representations combined with analogy been used to model aspects of conceptual change. For example, using a series of sketches to depict motion, a Companion learns intuitive models of force. Further, it progresses from simple to complex models in an order that corresponds to the order found in children (Friedman et al. 2010). It is hard to see how a Monte Carlo simulation approach would capture either the semantics of language about processes or the findings of the conceptual change literature.

Although we differ from Lake et al. in our view of intuitive physics and the role of analogical processing, we agree that rapid computation over structured representations is a major feature of human cognition. Today's deep learning systems are interesting for certain applications, but we doubt that they are on a direct path to understanding human cognition.

What can the brain teach us about building artificial intelligence?

doi:10.1017/S0140525X17000140, e265

Dileep George
Vicarious, Union City, CA 94587.
dileep@vicarious.com www.vicarious.com

Abstract: Lake et al. offer a timely critique on the recent accomplishments in artificial intelligence from the vantage point of human intelligence and provide insightful suggestions about research directions for building more human-like intelligence. Because we agree with most of the points they raised, here we offer a few points that are complementary.

The fact that "airplanes do not flap their wings" is often offered as a reason for not looking to biology for artificial intelligence (AI) insights. This is ironic because the idea that flapping is not required to fly, could easily have originated from observing eagles soaring on thermals. The comic strip in Figure 1 offers a humorous take on the current debate in AI. A flight researcher who does not take inspiration from birds defines an objective function for flight and ends up creating a catapult. Clearly, a catapult is an extremely useful invention. It can propel objects through the air, and in some cases, it can even be a better alternative to flying. Just as researchers who are interested in building "real flight" would be well advised to pay close attention to the differences between catapult flight and bird flight, researchers who are interested in building "human-like intelligence" or artificial general intelligence (AGI) would be well advised to pay attention to the differences between the recent successes of deep learning and human intelligence. We believe the target article delivers on that front, and we agree with many of its conclusions.

Better universal algorithms or more inductive biases? Learning and inference are instances of optimization algorithms. If we could derive a universal optimization algorithm that works well for all data, the learning and inference problems for building AGI would be solved as well. Researchers who work on assumption-free algorithms are pushing the frontier on this question.

Exploiting inductive biases and the structure of the AI problem makes learning and inference more efficient. Our brains show remarkable abilities to perform a wide variety of tasks on data that look very different. What if all of these different tasks and data have underlying similarities? Our view is that biological evolution, by trial and error, figured out a set of inductive biases that work well for learning in this world, and the human brain's efficiency and robustness derive from these biases. Lake et al. note that many researchers hope to overcome the need for inductive biases by bringing biological evolution into the fold of the learning algorithms. We point out that biological evolution had the advantage of using building blocks (proteins, cells) that obeyed the laws of the physics of the world in which these organisms were evolving to excel. In this way, assumptions about the world were implicitly baked into the representations that evolution used. Trying to evolve intelligence without assumptions might therefore be a significantly harder problem than biological evolution. AGI has one existence proof – our brains. Biological evolution is not an existence proof for artificial universal intelligence.

At the same time, we think a research agenda for building AGI could be synergistic with the quest for better universal algorithms. Our strategy is to build systems that strongly exploit inductive biases, while keeping open the possibility that some of those assumptions can be relaxed by advances in optimization algorithms.

What kind of generative model is the brain? Neuroscience can help, not just cognitive science. Lake et al. offered several compelling arguments for using cognitive science insights. In addition to cognitive science, neuroscience data can be examined to obtain

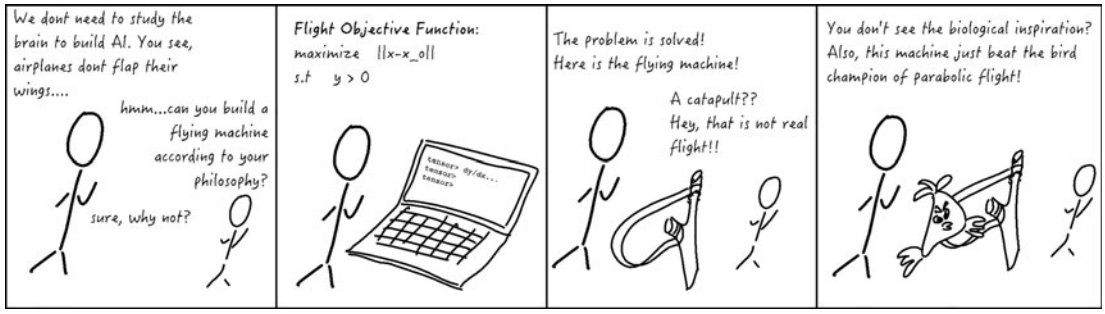

Figure 1 (George). A humorous take on the current debate in artificial intelligence.

clues about what kind of generative model the brain implements and how this model differs from models being developed in the AI community.

For instance, spatial lateral connections between oriented features are a predominant feature of the visual cortex and are known to play a role in enforcing contour continuity. However, lateral connections are largely ignored in current generative models (Lee 2015). Another example is the factorization of contours and surfaces. Evidence indicates that contours and surfaces are represented in a factored manner in the visual cortex (Zhou et al. 2000), potentially giving rise to the ability of humans to imagine and recognize objects with surface appearances that are not prototypical – like a blanket made of bananas or a banana made of blankets. Similarly, studies on top-down attention demonstrate the ability of the visual cortex to separate out objects even when they are highly overlapping and transparent (Cohen & Tong 2015). These are just a handful of examples from the vast repository of information on cortical representations and inference dynamics, all of which could be used to build AGI.

The conundrum of "human-level performance": Benchmarks for AGI. We emphasize the meaninglessness of "human-level performance," as reported in mainstream AI publications, and then use as a yardstick to measure our progress toward AGI. Take the case of the DeepQ network playing "breakout" at a "human level" (Mnih et al. 2015). We found that even simple changes to the visual environment (as insignificant as changing the brightness) dramatically and adversely affect the performance of the algorithm, whereas humans are not affected by such perturbations at all. At this point, it should be well accepted that almost any narrowly defined task can be "solved" with brute force data and computation and that any use of "human-level" as a comparison should be reserved for benchmarks that adhere to the following principles: (1) learning from few examples, (2) generalizing to distributions that are different from the training set, and (3) generalizing to new queries (for generative models) and new tasks (in the case of agents interacting with an environment).

Message passing-based algorithms for probabilistic models. Although the article makes good arguments in favor of structured probabilistic models, it is surprising that the authors mentioned only Markov chain Monte Carlo (MCMC) as the primary tool for inference. Although MCMC has asymptotic guarantees, the speed of inference in many cortical areas is more consistent with message passing (MP)-like algorithms, which arrive at maximum *a posteriori* solutions using only local computations. Despite lacking theoretical guarantees, MP has been known to work well in many practical cases, and recently we showed that it can be used for learning of compositional features (Lázaro-Gredilla et al. 2016). There is growing evidence for the use of MP-like inference in cortical areas (Bastos et al. 2012; George & Hawkins 2009), and MP could offer a happy medium where inference is fast, as in neural networks, while retaining MCMC's capability for answering arbitrary queries on the model.

Building brains that communicate like machines

doi:10.1017/S0140525X17000152, e266

Daniel Graham

Department of Psychology, Hobart & William Smith Colleges, Geneva, NY 14456.
graham@hws.edu **http://people.hws.edu/graham**

Abstract: Reverse engineering human cognitive processes may improve artificial intelligence, but this approach implies we have little to learn regarding brains from human-engineered systems. On the contrary, engineered technologies of dynamic network communication have many features that highlight analogous, poorly understood, or ignored aspects of brain and cognitive function, and mechanisms fundamental to these technologies can be usefully investigated in brains.

Lake et al. cogently argue that artificial intelligence (AI) machines would benefit from more "reverse engineering" of the human brain and its cognitive systems. However, it may be useful to invert this logic and, in particular, to use basic principles of machine communication to provide a menu of analogies and, perhaps, mechanisms that could be investigated in human brains and cognition.

We should consider that one of the missing components in deep learning models of cognition – and of most large-scale models of brain and cognitive function – is an understanding of how signals are selectively routed to different destinations in brains (Graham 2014; Graham and Rockmore 2011).

Given that brain cells themselves are not motile enough to selectively deliver messages to their destination (unlike cells in the immune system, for example), there must be a routing protocol of some kind in neural systems to accomplish this. This protocol should be relatively fixed in a given species and lineage, and have the ability to be scaled up over development and evolution.

Turning to machine communication as a model, each general technological strategy has its advantages and ideal operating conditions (grossly summarized here for brevity):

Circuit switched (traditional landline telephony): high throughput of dense real-time signals

Message switched (postal mail): multiplexed, verifiable, compact addresses

Packet switched (Internet): dynamic routing, sparse connectivity, fault tolerance, scalability

We should expect that brains adopt analogous – if not homologous – solutions when conditions require. For example, we would expect something like circuit switching in somatosensory and motor output systems, which tend to require dense, real-time communication. However, we would expect a dynamic, possibly packet-switched system in the visual system, given limited windows of attention and acuity and the need for spatial remapping, selectivity, and invariance (Olshausen et al. 1993; Poggio 1984; Wiskott 2006; Wiskott and von der Malsburg 1996).

There could be hybrid routing architectures at work in brains and several that act concurrently (consider by way of analogy that it was possible until recently for a single human communicator to use the three switching protocols described above simultaneously). Individual components of a given routing system could also be selectively employed in brains. For example, Fornito et al. (2016) proposed a mechanism of deflection routing (which is used to reroute signals around damaged or congested nodes), to explain changes in functional connectivity following focal lesions.

Nevertheless, functional demands in human cognitive systems appear to require a dynamic mechanism that could resemble a packet-switched system (Schlegel et al. 2015). As Lake et al. note, the abilities of brains to (1) grow and develop over time and (2) flexibly, creatively, and quickly adapt to new events are essential to their function. Packet switching as a general strategy may be more compatible with these requirements than alternative architectures.

In terms of growth, the number of Internet hosts – each of which can potentially communicate with any other within milliseconds – has increased without major disruption over a few decades, to surpass the number of neurons in the cortex of many primates including the macaque (Fasolo 2011). This growth has also been much faster than the growth of the message-switched U.S. Postal Service (Giambene 2005; U.S. Postal Service 2016). Cortical neurons, like Internet hosts, are separated by relatively short network distances, and have the potential for communication along many possible routes within milliseconds. Communication principles that allowed for the rapid rise and sustained development of the packet-switched Internet may provide insights relevant to understanding how evolution and development conspire to generate intelligent brains.

In terms of adapting quickly to new situations, Lake et al. point out that a fully trained artificial neural network generally cannot take on new or different tasks without substantial retraining and reconfiguration. Perhaps this is not so much a problem of computation, but rather one of routing: in neural networks, one commonly employs a fixed routing system, all-to-all connectivity between layers, and feedback only between adjacent layers. These features may make such systems well suited to learning a particular input space, but ill suited to flexible processing and efficient handling of new circumstances. Although a packet-switched routing protocol would not necessarily improve current deep learning systems, it may be better suited to modeling approaches that more closely approximate cortical networks' structure and function. Unlike most deep learning networks, the brain appears to largely show dynamic routing, sparse connectivity, and feedback among many hierarchical levels. Including such features in computational models may better approximate and explain biological function, which could in turn spawn better AI.

Progress in understanding routing in the brain is already being made through simulations of dynamic signal flow on brain-like networks and in studies of brains themselves. Mišić et al. (2014) have investigated how Markovian queuing networks (a form of message-switched architecture) with primate brain-like connectivity could take advantage of small-world and rich-club topologies. Complementing this work, Sizemore et al. (2016) have shown that the abundance of weakly interconnected brain regions suggests a prominent role for parallel processing, which would be well suited to dynamic routing. Using algebraic topology, Sizemore et al. (2016) provide evidence that human brains show loops of converging or diverging signal flow (see also Granger 2006). In terms of neurophysiology, Briggs and Usrey (2007) have shown that corticothalamic networks can pass signals in a loop in just 37 milliseconds. Such rapid feedback is consistent with the notion that corticothalamic signals could function like the "ack" (acknowledgment) system used on the Internet to ensure packet delivery (Graham 2014; Graham and Rockmore 2011).

In conclusion, it is suggested that an additional "core ingredient of human intelligence" is dynamic information routing of a kind that may mirror the packet-switched Internet, and cognitive scientists and computer engineers alike should be encouraged to investigate this possibility.

The importance of motivation and emotion for explaining human cognition

doi:10.1017/S0140525X17000164, e267

C. Dominik Güss[a] and Dietrich Dörner[b]

[a]*Department of Psychology, University of North Florida, Jacksonville, FL 32224;* [b]*Trimberg Research Academy TRAc, Otto-Friedrich Universität Bamberg, 96047 Bamberg, Germany.*

dguess@unf.edu dietrich.doerner@uni-bamberg.de

https://www.unf.edu/bio/N00174812

https://www.uni-bamberg.de/trac/senior-researchers/doerner

Abstract: Lake et al. discuss building blocks of human intelligence that are quite different from those of artificial intelligence. We argue that a theory of human intelligence has to incorporate human motivations and emotions. The interaction of motivation, emotion, and cognition is the real strength of human intelligence and distinguishes it from artificial intelligence.

Lake et al. applaud the advances made in artificial intelligence (AI), but argue that future research should focus on the most impressive form of intelligence, namely, natural/human intelligence. In brief, the authors argue that AI does not resemble human intelligence. The authors then discuss the building blocks of human intelligence, for example, developmental start-up software including intuitive physics and intuitive psychology, and learning as a process of model building based on compositionality and causality, and they stress that "people never start completely from scratch" (sect. 3.2, last para.)

We argue that a view of human intelligence that focuses solely on cognitive factors misses crucial aspects of human intelligence. In addition to cognition, a more complete view of human intelligence must incorporate motivation and emotion, a viewpoint already stated by Simon: "Since in actual human behavior motive and emotion are major influences on the course of cognitive behavior, a general theory of thinking and problem solving must incorporate such influences" (Simon 1967, p. 29; see also Dörner & Güss 2013).

Incorporating motivation (e.g., Maslow 1954; Sun 2016) in computational models of human intelligence can explain where goals come from. Namely, goals come from specific needs, for example, from existential needs such as hunger or pain avoidance; sexual needs; the social need for affiliation, to be together with other people; the need for certainty related to unpredictability of the environment; and the need for competence related to ineffective coping with problems (Dörner 2001; Dörner & Güss 2013). Motivation can explain why a certain plan has priority and why it is executed, or why a certain action is stopped. Lake et al. acknowledge the role of motivation in one short paragraph when they state: "There may also be an intrinsic drive to reduce uncertainty and construct models of the environment" (sect. 4.3.2, para. 4). This is right. However, what is almost more important is the need for competence, which drives people to explore new environments. This is also called diversive exploration (e.g., Berlyne 1966). Without diversive exploration, mental models could not grow, because people would not seek new experiences (i.e., seek uncertainty to reduce uncertainty afterward).

Human emotion is probably the biggest difference between people and AI machines. Incorporating emotion into computational models of human intelligence can explain some aspects that the authors discuss as "deep learning" and "intuitive psychology." Emotions are shortcuts. Emotions are the framework in which cognition happens (e.g., Bach 2009; Dörner 2001). For example, not reaching an important goal can make a person angry. Anger then characterizes a specific form of perception, planning, decision making, and behavior. Anger means high activation, quick and rough perception, little planning and deliberation, and making a quick choice. Emotions modulate human behavior; the *how* of the behavior is determined by the emotions.

In other situations, emotions can trigger certain cognitive processes. In some problem situations, for example, a person would get an "uneasy" feeling when all solution attempts do not result in a solution. This uneasiness can be the start of metacognition. The person will start reflecting on his or her own thinking: "What did I do wrong? What new solution could I try?" In this sense, human intelligence controls itself, reprogramming its own programs.

And what is the function of emotions? The function of emotions is to adjust behavior to the demands of the current situation. Perhaps emotions can partly explain why humans learn "rich models from sparse data" (sect. 4.3, para. 1), as the authors state. A child observing his or her father smiling and happy when watching soccer does not need many trials to come to the conclusion that soccer must be something important that brings joy.

In brief, a theory or a computational model of human intelligence that focuses solely on cognition is not a real theory of human intelligence. As the authors state, "Our machines need to confront the kinds of tasks that human learners do." This means going beyond the "simple" Atari game Frostbite. In Frostbite, the goal was well defined (build an igloo). The operations and obstacles were known (go over ice floes without falling in the water and without being hit by objects/animals). The more complex, dynamic, and "real" such tasks become – as has been studied in the field of Complex Problem Solving or Dynamic Decision Making (e.g., Funke 2010; Güss Tuason & Gerhard 2010), the more human behavior will show motivational, cognitive, and emotional processes in their interaction. This interaction of motivation, cognition, and emotion, is the real strength of human intelligence compared with artificial intelligence.

Building on prior knowledge without building it in

doi:10.1017/S0140525X17000176, e268

Steven S. Hansen,[a] Andrew K. Lampinen,[a] Gaurav Suri,[b] and James L. McClelland[a]

[a]*Psychology Department, Stanford University, Stanford, CA 94305;* [b]*Psychology Department, San Francisco State University, San Francisco, CA 94132.*

sshansen@stanford.edu **lampinen@stanford.edu**
rav.psych@gmail.com **mcclelland@stanford.edu**
http://www.suriradlab.com/ **https://web.stanford.edu/group/pdplab/**

Abstract: Lake et al. propose that people rely on "start-up software," "causal models," and "intuitive theories" built using compositional representations to learn new tasks more efficiently than some deep neural network models. We highlight the many drawbacks of a commitment to compositional representations and describe our continuing effort to explore how the ability to build on prior knowledge and to learn new tasks efficiently could arise through learning in deep neural networks.

Lake et al. have laid out a perspective that builds on earlier work within the structured/explicit probabilistic cognitive modeling framework. They have identified several ways in which humans with existing domain knowledge can quickly acquire new domain knowledge and deploy their knowledge flexibly. Lake et al. also make the argument that the key to understanding these important human abilities is the use of "start-up software," "causal models," and "intuitive theories" that rely on a compositional knowledge representation of the kind advocated by, for example, Fodor and Pylyshyn (1988).

We agree that humans can often acquire new domain knowledge quickly and can often generalize this knowledge to new examples and use it in flexible ways. However, we believe that human knowledge acquisition and generalization can be understood without building in a commitment to domain-specific knowledge structures or compositional knowledge representation. We therefore expect that continuing our longstanding effort to understand how human abilities can emerge without assuming special start-up software will be most helpful in explicating the nature of human cognition.

The explicit compositional approach of Lake et al. is limited because it downplays the often complex interactions between the multitude of contextual variables in the task settings in which the representation is used. Avoiding a commitment to symbolic compositionality increases one's flexibility to respond to sometimes subtle influences of context and allows for the possibility of more robust learning across contexts. The recent startling improvements in computer vision (Krizhevsky et al. 2012), machine translation (Johnson et al. 2016), and question answering (Weston et al. 2015a) were possible, precisely because they avoided these limitations by foregoing symbolic compositionality altogether.

Although Lake et al. seek to take the computational-level "high ground" (Marr 1982), their representational commitments also constrain the inferential procedures on which they rely. Their modeling work relies on the use of combinatorially explosive search algorithms. This approach can be effective in a specific limited domain (such as Omniglot), precisely because the startup software can be hand selected by the modeler to match the specific requirements of that specific domain. However, their approach avoids the hard question of where this startup software came from. Appeals to evolution, although they may be plausible for some tasks, seem out of place in domains of recent human invention such as character-based writing systems. Also, because many naturalistic learning contexts are far more open ended, combinatorial search is not a practical algorithmic strategy. Here, the gradient-based methods of neural networks have proven far more effective (see citations above).

We believe learning research will be better off taking a domain general approach wherein the startup software used when one encounters a task as an experienced adult human learner is the experience and prior knowledge acquired through a domain general learning process.

Most current deep learning models, however, do not build on prior experience. For example, the network in Mnih et al. (2013) that learns Atari games was trained from scratch on each new problem encountered. This is clearly not the same as human learning, which builds cumulatively on prior learning. Humans learn complex skills in a domain after previously learning simpler ones, gradually building structured knowledge as they learn. In games like Chess or Go, human learners can receive feedback not only on the outcome of an entire game – did the learner succeed or fail? – but also on individual steps in an action sequence. This sort of richer feedback can easily be incorporated into neural networks, and doing so can enhance learning (Gülçehre and Bengio 2016).

An important direction is to explore how humans learn from a rich ensemble of multiple, partially related tasks. The steps of a sequential task can be seen as mutually supporting subtasks, and a skill, such as playing chess can be seen as a broad set of related tasks beyond selecting moves: predicting the opponent's moves, explaining positions, and so on. One reason humans might be able to learn from fewer games than a neural network trained on playing chess as a single integrated task is that humans receive feedback on many of these tasks throughout learning, and this both allows more feedback from a single experience (e.g., both an emotional reward for capturing a piece and an explanation of the tactic from a teacher) and constrains the representations that can emerge (they must support all of these related subtasks). Such constraints amount to extracting shared principles that allow for accelerated learning when encountering other tasks that use them. One example is training a recurrent network on translation tasks between multiple language pairs, which can lead to zero-shot (no training necessary) generalization, to translation between unseen language pairs (Johnson et al. 2016). Just as neural networks can exhibit rulelike behavior without building in explicit rules, we believe that they may not require a compositional, explicitly symbolic form of reasoning to produce human-like behavior.

Indeed, recent work on meta-learning (or learning-to-learn) in deep learning models provides a base for making good on this claim (Bartunov and Vetrov 2016; Santoro et al. 2016; Vinyals et al. 2016). The appearance of rapid learning (e.g., one-shot classification) is explained as slow, gradient-based learning on a meta-problem (e.g., repeatedly solving one-shot classification problems drawn from a distribution). Although the meta-tasks used in these first attempts only roughly reflect the training environment that humans face (we probably do not face explicit one-shot classification problems that frequently), the same approach could be used with meta-tasks that are extremely common as a result of sociocultural conventions, such as "follow written instructions," "incorporate comments from a teacher," and "give a convincing explanation of your behavior."

Fully addressing the challenges Lake et al. pose – rather than building in compositional knowledge structures that will ultimately prove limiting – is a long-term challenge for the science of learning. We expect meeting this challenge to take time, but that the time and effort will be well spent. We would be pleased if Lake et al. would join us in this effort. Their participation would help accelerate progress toward a fuller understanding of how advanced human cognitive abilities arise when humans are immersed in the richly structured learning environments that have arisen in human cultures and their educational systems.

Building machines that adapt and compute like brains

doi:10.1017/S0140525X17000188, e269

Nikolaus Kriegeskorte and Robert M. Mok

Medical Research Council Cognition and Brain Sciences Unit, Cambridge, CB2 7EF, United Kingdom.

Nikolaus.Kriegeskorte@mrc-cbu.cam.ac.uk

Robert.Mok@mrc-cbu.cam.ac.uk

Abstract: Building machines that learn and think like humans is essential not only for cognitive science, but also for computational neuroscience, whose ultimate goal is to understand how cognition is implemented in biological brains. A new cognitive computational neuroscience should build cognitive-level and neural-level models, understand their relationships, and test both types of models with both brain and behavioral data.

Lake et al.'s timely article puts the recent exciting advances with neural network models in perspective, and usefully highlights the aspects of human learning and thinking that these models do not yet capture. Deep convolutional neural networks have conquered pattern recognition. They can rapidly recognize objects as humans can, and their internal representations are remarkably similar to those of the human ventral stream (Eickenberg et al. 2016; Güçlü & van Gerven 2015; Khaligh-Razavi & Kriegeskorte 2014; Yamins et al. 2014). However, even at a glance, we understand visual scenes much more deeply than current models. We bring complex knowledge and dynamic models of the world to bear on the sensory data. This enables us to infer past causes and future implications, with a focus on what matters to our behavioral success. How can we understand these processes mechanistically?

The top-down approach of cognitive science is one required ingredient. Human behavioral researchers have an important role in defining the key challenges for model engineering by introducing tasks where humans still outperform the best models. These tasks serve as benchmarks, enabling model builders to measure progress and compare competing approaches. Cognitive science introduced task-performing computational models of cognition. Task-performing models are also essential for neuroscience, whose theories cannot deliver explicit accounts of intelligence without them (Eliasmith & Trujillo 2014). The current constructive competition between modeling at the cognitive level and modeling at the neural level is inspiring and refreshing. We need both levels of description to understand, and to be able to invent, intelligent machines and computational theories of human intelligence.

Pattern recognition was a natural first step toward understanding human intelligence. This essential component mechanism has been conquered by taking inspiration from the brain. Machines could not do core object recognition (DiCarlo et al. 2012) until a few years ago (Krizhevsky et al. 2012). Brain-inspired neural networks gave us machines that can recognize objects robustly under natural viewing conditions. As we move toward higher cognitive functions, we might expect that it will continue to prove fruitful to think about cognition in the context of its implementation in the brain. To understand how humans learn and think, we need to understand how brains adapt and compute.

A neural network model may require more time to train than humans. This reflects the fact that current models learn from scratch. Cognitive models, like Bayesian program learning (Lake et al. 2015a), rely more strongly on built-in knowledge. Their inferences require realistically small amounts of data, but unrealistically large amounts of computation, and, as a result, their high-level feats of cognition do not always scale to complex real-world challenges. To explain human cognition, we must care about efficient implementation and scalability, in addition to the goals of computation. Studying the brain can help us understand the representations and dynamics that support the efficient implementation of cognition (e.g., Aitchison & Lengyel 2016).

The brain seamlessly merges bottom-up discriminative and top-down generative processes into a rapidly converging process of inference that combines the advantages of both: the rapidity of discriminative inference and the flexibility and precision of generative inference (Yildirim et al. 2015). The brain's inference process appears to involve recurrent cycles of message passing at multiple scales, from local interactions within an area to long-range interactions between higher- and lower-level representations.

As long as major components of human intelligence are out of the reach of machines, we are obviously far from understanding the human brain and cognition. As more and more component tasks are conquered by machines, the question of whether they do it "like humans" will come to the fore. How should we define "human-like" learning and thinking? In cognitive science, the empirical support for models comes from behavioral data. A model must not only reach human levels of task performance, but also predict detailed patterns of behavioral responses (e.g., errors and reaction times on particular instances of a task). However, humans are biological organisms, and so "human-like" cognition should also involve the same brain representations and algorithms that the human brain employs. A good model should somehow match the brain's dynamics of information processing.

Measuring the similarity of processing dynamics between a model and a brain has to rely on summary statistics of the activity and may be equally possible for neural and cognitive models. For neural network models, a direct comparison may seem more tractable. We might map the units of the model onto neurons in the brain. However, even two biological brains of the same species will have different numbers of neurons, and any given neuron may be idiosyncratically specialized, and may not have an exact match in the other brain. For either a neural or a cognitive model, we may find ways to compare the internal model representations to representations in brains (e.g., Kriegeskorte & Diedrichsen 2016; Kriegeskorte et al. 2008). For example, one could test whether the visual representation of characters in high-level visual regions reflects the similarity predicted by the generative model of character perception proposed by Lake et al. (2015a).

The current advances in artificial intelligence re-invigorate the interaction between cognitive science and computational neuroscience. We hope that the two can come together and combine their empirical and theoretical constraints, testing cognitive and neural models with brain and behavioral data. An integrated cognitive computational neuroscience might have a shot at the task that seemed impossible a few years ago: understanding how the brain works.

Will human-like machines make human-like mistakes?

doi:10.1017/S0140525X1700019X, e270

Evan J. Livesey, Micah B. Goldwater, and Ben Colagiuri
School of Psychology, The University of Sydney, NSW 2006, Australia.
evan.livesey@sydney.edu.au micah.goldwater@sydney.edu.au
ben.colagiuri@sydney.edu.au
http://sydney.edu.au/science/people/evan.livesey.php
http://sydney.edu.au/science/people/micah.goldwater.php
http://sydney.edu.au/science/people/ben.colagiuri.php

Abstract: Although we agree with Lake et al.'s central argument, there are numerous flaws in the way people use causal models. Our models are often incorrect, resistant to correction, and applied inappropriately to new situations. These deficiencies are pervasive and have real-world consequences. Developers of machines with similar capacities should proceed with caution.

Lake et al. present a compelling case for why causal model-building is a key component of human learning, and we agree that beliefs about causal relations need to be captured by any convincingly human-like approach to artificial intelligence (AI). Knowledge of physical relations between objects and psychological relations between agents brings huge advantages. It provides a wealth of transferable information that allows humans to quickly apprehend a new situation. As such, combining the computational power of deep-neural networks with model-building capacities could indeed bring solutions to some of the world's most pressing problems. However, as advantageous as causal model-building might be, it also brings problems that can lead to flawed learning and reasoning. We therefore ask, would making machines "human-like" in their development of causal models also make those systems flawed in human-like ways?

Applying a causal model, especially one based on intuitive understanding, is essentially a gamble. Even though we often feel like we understand the physical and psychological relations surrounding us, our causal knowledge is almost always incomplete and sometimes completely wrong (Rozenblit & Keil 2002). These errors may be an inevitable part of the learning process by which models are updated based on experience. However, there are many examples in which incorrect causal models persist, despite strong counterevidence. Take the supposed link between immunisation and autism. Despite the science and the author of the original vaccine-autism connection being widely and publicly discredited, many continue to believe that immunisation increases the risk of autism and their refusal to immunise has decreased the population's immunity to preventable diseases (Larson et al. 2011; Silverman & Hendrix 2015).

Failures to revise false causal models are far from rare. In fact, they seem to be an inherent part of human reasoning. Lewandowsky and colleagues (2012) identify numerous factors that increase resistance to belief revision, including several that are societal-level (e.g., biased exposure to information) or motivational (e.g., vested interest in retaining a false belief). Notwithstanding the significance of these factors (machines too can be influenced by biases in data availability and the motives of their human developers), it is noteworthy that people still show resistance to updating their beliefs even when these sources of bias are removed, especially when new information conflicts with the existing causal model (Taylor & Ahn 2012).

Flawed causal models can also be based on confusions that are less easily traced to specific falsehoods. Well-educated adults regularly confuse basic ontological categories (Chi et al. 1994), distinctions between mental, biological, and physical phenomena that are fundamental to our models of the world and typically acquired in childhood (Carey 2011). A common example is the belief that physical energy possesses psychological desires and intentions – a belief that even some physics students appear to endorse (Svedholm & Lindeman 2013). These errors affect both our causal beliefs and our choices. Ontological confusions have been linked to people's acceptance of alternative medicine, potentially leading an individual to choose an ineffective treatment over evidence-based treatments, sometimes at extreme personal risk (Lindeman 2011).

Causal models, especially those that affect beliefs about treatment efficacy, can even influence physiological responses to medical treatments. In this case, known as the placebo effect, beliefs regarding a treatment can modulate the treatment response, positively or negatively, independently of whether a genuine treatment is delivered (Colagiuri et al. 2015). The placebo effect is caused by a combination of expectations driven by causal beliefs and associative learning mechanisms that are more analogous to the operations of simple neural networks. Associative learning algorithms, of the kind often used in neural networks, are surprisingly susceptible to illusory correlations, for example, when a treatment actually has no effect on a medical outcome (Matute et al. 2015). Successfully integrating two different mechanisms for knowledge generation (neural networks and causal models), when each individually may be prone to bias, is an interesting problem, not unlike the challenge of understanding the nature of human learning. Higher-level beliefs interact in numerous ways with basic learning and memory mechanisms, and the precise nature and consequences of these interactions remain unknown (Thorwart & Livesey 2016).

Even when humans hold an appropriate causal model, they often fail to use it. When facing a new problem, humans often erroneously draw upon models that share superficial properties with the current problem, rather than those that share key structural relations (Gick & Holyoak 1980). Even professional management consultants, whose job it is to use their prior experiences to help businesses solve novel problems, often fail to retrieve the most relevant prior experience to the new problem (Gentner et al. 2009). It is unclear whether an artificial system that possesses mental modelling capabilities would suffer the same limitations. On the one hand, they may be caused by human processing limitations. For example, effective model-based decision-making is associated with capacities for learning and transferring abstract rules (Don et al. 2016), and for cognitive control (Otto et al. 2015), which may potentially be far more powerful in future AI systems. On the other hand, the power of neural networks lies precisely in their ability to encode rich featural and contextual information. Given that experience with particular causal relations is likely to correlate with experience of more superficial features, a more powerful AI model generator may still suffer similar problems when faced with the difficult decision of which model to apply to a new situation.

Would human-like AI suffer human-like flaws, whereby recalcitrant causal models lead to persistence with poor solutions, or novel problems activate inappropriate causal models? Developers of AI systems should proceed with caution, as these properties of human causal modelling produce pervasive biases, and may be symptomatic of the use of mental models rather than the limitations on human cognition. Monitoring the degree to which AI systems show the same flaws as humans will be invaluable for shedding light on why human cognition is the way it is and, it is hoped, will offer some solutions to help us change our minds when we desperately need to.

Benefits of embodiment

doi:10.1017/S0140525X17000206, e271

Bruce James MacLennan
Department of Electrical Engineering and Computer Science, University of Tennessee, Knoxville, TN 37996.
maclennan@utk.edu http://web.eecs.utk.edu/~mclennan

Abstract: Physical competence is acquired through animals' embodied interaction with their physical environments, and psychological

competence is acquired through situated interaction with other agents. The acquired neural models essential to these competencies are implicit and permit more fluent and nuanced behavior than explicit models. The challenge is to understand how such models are acquired and used to control behavior.

The target article argues for the importance of "developmental start-up software" (sects. 4.1 and 5.1), but neglects the nature of that software and how it is acquired. The embodied interaction of an organism with its environment, provides a foundation for its understanding of "intuitive physics" and physical causality. Animal nervous systems control their complex physical bodies in their complex physical environments in real time, and this competence is a consequence of innate developmental processes and, especially in more complex species, subsequent developmental processes that fine-tune neural control, such as prenatal and postnatal "motor babbling" (non-goal-directed motor activity) (Meltzoff & Moore 1997). Through these developmental processes, animals acquire a non-conceptual understanding of their bodies and physical environments, which provides a foundation for higher-order imaginative and conceptual physical understanding.

Animals acquire physical competence through interaction with their environments (both phylogenetic through evolution and ontogenetic through development), and robots can acquire physical competence similarly, for example, through motor babbling (Mahoor et al. 2016), and this is one goal of epigenetic and developmental robotics (Lungarella et al. 2003). In principle, comparable competence can be acquired by simulated physical agents behaving in simulated physical environments, but it is difficult to develop sufficiently accurate physical simulations so that agents acquire genuine physical competence (i.e., competence in the real world, not some simulated world). It should be possible to transfer physical competence from one agent to others that are sufficiently similar physically, but the tight coupling of body and nervous system suggests that physical competence will remain tied to a "form of life."

Animals are said to be *situated* because cognition primarily serves behavior, and behavior is always contextual. For most animals, situatedness involves interaction with other animals; it conditions the goals, motivations, and other factors that are causative in an animal's own behavior, and can be projected onto other agents, providing a foundation for "intuitive psychology." Psychological competence is grounded in the fact that animals are situated physical agents with interests, desires, goals, fears, and so on. Therefore, they have a basis for non-conceptual understanding of other agents (through imagination, mental simulation, projection, mirror neurons, etc.). In particular, they can project their experience of psychological causality onto other animals. This psychological competence is acquired through phylogenetic and ontogenetic adaptation.

The problem hindering AI systems from acquiring psychological competence is that most artificial agents do not have interests, desires, goals, fears, and so on that they can project onto others or use as a basis for mental simulation. For example, computer vision systems do not "care" in any significant way about the images they process. Because we can be injured and die, because we can feel fear and pain, we perceive immediately (i.e., without the mediation of conceptual thought) the significance of a man being dragged by a horse, or a family fleeing a disaster (Lake et al., Fig. 6). Certainly, through artificial evolution and reinforcement learning, we can train artificial agents to interact competently with other (real or simulated) agents, but because they are a different form of life, it will be difficult to give them the same cares and concerns as we have and that are relevant to many of our practical applications.

The target article does not directly address the important distinction between explicit and implicit models. *Explicit models* are the sort scientists construct, generally in terms of symbolic (lexical-level) variables; we expect to be able to understand explicit models conceptually, to communicate them in language, and to reason about them discursively (including mathematically). *Implicit models* are the sort that neural networks construct, generally in terms of large numbers of sub-symbolic variables, densely interrelated. Implicit models often allow an approximate emergent symbolic description, but such descriptions typically capture only the largest effects and interrelationships implicit in the sub-symbolic model. Therefore, they may lack the subtlety and context sensitivity of implicit models, which is why it is difficult, if not impossible, to capture expert behavior in explicit rules (Dreyfus & Dreyfus 1986). Therefore, terms such as "intuitive physics," "intuitive psychology," and "theory of mind" are misleading because they connote explicit models, but implicit models (especially those acquired by virtue of embodiment and situatedness) are more likely to be relevant to the sorts of learning discussed in the target article. It is less misleading to refer to *competencies*, because humans and other animals can use their physical and psychological understanding to behave competently even in the absence of explicit models.

The target article shows the importance of hierarchical compositionality to the physical competence of humans and other animals (sect. 4.2.1); therefore, it is essential to understand how hierarchical structure is represented in implicit models. Recognizing the centrality of embodiment can help, for our bodies are hierarchically articulated and our physical environments are hierarchically structured. The motor affordances of our bodies provide a basis for non-conceptual understanding of the hierarchical structure of objects and actions. However, it iss important to recognize that hierarchical decompositions need not be unique; they may be context dependent and subject to needs and interests, and a holistic behavior may admit multiple incompatible decompositions.

The target article points to the importance of simulation-based and imagistic inference (sect. 4.1.1). Therefore, we need to understand how they are implemented through implicit models. Fortunately, neural representations, such as topographic maps, permit analog transformations, which are better than symbolic digital computation for simulation-based and imagistic inference. The fact of neural implementation can reveal modes of information processing and control beyond the symbolic paradigm.

Connectionism consciously abandoned the explicit models of symbolic AI and cognitive science in favor of implicit, neural network models, which had a liberating effect on cognitive modeling, AI, and robotics. With 20-20 hindsight, we know that many of the successes of connectionism could have been achieved through existing statistical methods (e.g., Bayesian inference), without any reference to the brain, but they were not. Progress had been retarded by the desire for explicit, human-interpretable models, which connectionism abandoned in favor of neural plausibility. We are ill advised to ignore the brain again.

Understand the cogs to understand cognition

doi:10.1017/S0140525X17000218, e272

Adam H. Marblestone,[a] Greg Wayne,[b] and Konrad P. Kording[c]

[a]*Synthetic Neurobiology Group, MIT Media Lab, Cambridge, MA 02474;* [b]*DeepMind, London N1 9DR, UK;* [c]*Departments of Bioengineering and Neuroscience, University of Pennsylvania, Philadelphia, PA 19104*

adam.h.marblestone@gmail.com gregwayne@gmail.com kording@upenn.edu http://www.adammarblestone.org/ www.kordinglab.com

Abstract: Lake et al. suggest that current AI systems lack the inductive biases that enable human learning. However, Lake et al.'s proposed biases may not directly map onto mechanisms in the developing brain. A convergence of fields may soon create a correspondence between biological neural circuits and optimization in structured architectures, allowing us to systematically dissect how brains learn.

The target article by Lake et al. beautifully highlights limitations of today's artificial intelligence (AI) systems relative to the performance of human children and adults. Humans demonstrate uptake and generalization of concepts in the domains of intuitive physics and psychology, decompose the world into reusable parts, transfer knowledge across domains, and reason using models of the world. As Lake et al. emphasize, and as is a mathematical necessity (Ho & Pepyne 2002), humans are not generic, universal learning systems: they possess inductive biases that constrain and guide learning for species-typical tasks.

However, the target article's characterization of these inductive biases largely overlooks how they may arise in the brain and how they could be engineered into artificial systems. Their particular choice of inductive biases, though supported by psychological research (see Blumberg [2005] for a critique), is in some ways arbitrary or idiosyncratic: It is unclear whether these capabilities are the key ones that enable human cognition, unclear whether these inductive biases correspond to separable "modules" in any sense, and, most importantly, unclear how these inductive biases could actually be built. For example, the cognitive level of description employed by Lake et al. gives little insight into whether the systems underlying intuitive psychology and physics comprise overlapping mechanisms. An alternative and plausible view holds that both systems may derive from an underlying ability to make sensory predictions, conditioned on the effects of actions, which could be bootstrapped through, for example, motor learning. With present methods and knowledge, it is anybody's guess which of these possibilities holds true: an additional source of constraint and inspiration seems needed.

Lake et al. seem to view circuit and systems neuroscience as unable to provide strong constraints on the brain's available computational mechanisms – perhaps in the same way that transistors place few meaningful constraints on the algorithms that may run on a laptop. However, the brain is not just a hardware level on which software runs. Every inductive bias is a part of the genetic and developmental makeup of the brain. Indeed, whereas neuroscience has not *yet* produced a sufficiently well-established computational description to decode the brain's inductive biases, we believe that this will change soon. In particular, neuroscience may be getting close to establishing a more direct correspondence between neural circuitry and the optimization algorithms and structured architectures used in deep learning. For example, many inductive biases may be implemented through the precise choice of cost functions used in the optimization of the connectivity of a neuronal network. But to identify *which* cost function is actually being optimized in a cortical circuit, we must first know *how* the circuit performs optimization. Recent work is starting to shed light on this question (Guergiuev et al. 2016), and to do so, it has been forced to look deeply not only at neural circuits, but also even at how learning is implemented at the subcellular level. Similar opportunities hold for crossing thresholds in our understanding of the neural basis of other key components of machine learning agents, such as structured information routing, memory access, attention, hierarchical control, and decision making.

We argue that the study of evolutionarily conserved neural structures will provide a means to identify the brain's true, fundamental inductive biases and how they actually arise. Specifically, we propose that optimization, architectural constraints, and "bootstrapped cost functions" might be the basis for the development of complex behavior (Marblestone et al. 2016). There are many potential mechanisms for gradient-based optimization in cortical circuits, and many ways in which the interaction of such mechanisms with multiple other systems could underlie diverse forms of structured learning like those hypothesized in Lake et al. Fundamental neural structures are likely tweaked and re-used to underpin different kinds of inductive biases across animal species, including humans. Within the lifetime of an animal, a developmentally orchestrated sequence of experience-dependent cost functions may provide not just a list of inductive biases, but a procedure for sequentially unfolding inductive biases within brain systems to produce a fully functional organism.

A goal for both AI and neuroscience should be to advance both fields to the point where they can have a useful conversation about the specifics. To do this, we need not only to build more human-like inductive biases into our machine learning systems, but also to understand the architectural primitives that are employed by the brain to set up these biases. This has not yet been possible because of the fragmentation and incompleteness of our neuroscience knowledge. For neuroscience to *ask* questions that directly inform the computational architecture, it must first cross more basic thresholds in understanding. To build a bridge with the intellectual frameworks used in machine learning, it must establish the neural underpinnings of optimization, cost functions, memory access, and information routing. Once such thresholds are crossed, we will be in a position – through a joint effort of neuroscience, cognitive science, and AI – to identify the brain's actual inductive biases and how they integrate into a single developing system.

Social-motor experience and perception-action learning bring efficiency to machines

doi:10.1017/S0140525X1700022X, e273

Ludovic Marin[a] and Ghiles Mostafaoui[b]

[a]*EuroMov Laboratory, University of Montpellier, Montpellier, France;* [b]*ETIS Laboratory, Cergy-Pontoise University, 95302 Cergy Pontoise, France.*

ludovic.marin@umontpellier.fr ghiles.mostafaoui@ensea.fr

http://euromov.eu/team/ludovic-marin/

Abstract: Lake et al. proposed a way to build machines that learn as fast as people do. This can be possible only if machines follow the human processes: the perception-action loop. People perceive and act to understand new objects or to promote specific behavior to their partners. In return, the object/person provides information that induces another reaction, and so on.

The authors of the target article stated, "the interaction between representation and previous experience may be key to building machines that learn as fast as people do" (sect. 4.2.3, last para.) To design such machines, they should function as humans do. But a human acts and learns based on his or her social-MOTOR experience. Three main pieces of evidence can demonstrate our claim:

First, any learning or social interacting is based on social motor embodiment. In the field of human movement sciences, many pieces of evidence indicate that we are all influenced by the motor behavior of the one with whom we are interacting (e.g., Schmidt & Richardson 2008). The motor behavior directly expresses the state of mind of the partner (Marin et al. 2009). For example, if someone is shy, this state of mind will be directly embodied in her or his entire posture, facial expressions, gaze, and gestures. It is in the movement that we observe the state of mind of the other "interactant." But when we are responding to that shy person, we are influenced in return by that behavior. Obviously we can modify intentionally our own motor behavior (to ease the interaction with him or her). But in most cases we are not aware of the alterations of our movements. For example, when an adult walks next to a child, they both unintentionally synchronize their stride length to each other (implying they both modify their locomotion to walk side-by-side). Another example in mental health disorders showed that an individual suffering from schizophrenia does not interact "motorly" the same way as a social phobic (Varlet et al. 2014). Yet, both pathologies present motor impairment and social withdrawal. But what characterizes their motor differences is based on the state of mind of the patients. In our example, the first patient presents attentional

impairment, whereas the other suffers from social inhibition. If, however, a healthy participant is engaged in a social-motor synchronization task, both participants (the patient and the healthy subject) unintentionally adjust their moves (Varlet et al. 2014).

This study demonstrates that unconscious communication is sustained even though the patients are suffering from social interaction disorders. We can then state that mostly low-level treatments of sensorimotor flows are involved in this process. Consequently, machines/robots should be embedded with computational models, which tackles the very complex question of adapting to the human world using sensorimotor learning.

We claim that enactive approaches of this type will drastically reduce the complexity of future computational models. Methods of this type are indeed supported by recent advances in the human brain mirroring system and theories based on motor resonance (Meltzoff 2007). In this line of thinking, computational models have been built and used to improve human robot interaction and communication, in particular through the notion of learning by imitation (Breazeal & Scassellati 2002; Lopes & Santos-Victor 2007). Furthermore, some studies embedded machines with computational models using an adequate action-perception loop and showed that some complex social competencies such as immediate imitation (present in early human development) could emerge through sensorimotor ambiguities as proposed in Gaussier et al. (1998), Nagai et al. (2011), and Braud et al. (2014).

This kind of model allows future machines to better generalize their learning and to acquire new social skills. In other recent examples, using a very simple neural network providing minimal sensorimotor adaptation capabilities to the robot, unintentional motor coordination could emerge during an imitation game (of a simple gesture) with a human (Hasnain et al. 2012; 2013). An extension of this work demonstrated that a robot could quickly and "online" learn more complex gestures and synchronize its behavior to the human partner based on the same sensorimotor approach (Ansermin et al. 2016).

Second, even to learn (or understand) what a simple object is, people need to act on it (O'Regan 2011). For example, if we do not know what a "chair" is, we will understand its representation by sitting on it, touching it. The definition is then easy: A chair is an object on which we can sit, regardless of its precise shape. Now, if we try to define its representation before acting, it becomes very difficult to describe it. This requires determining the general shape, number of legs, with or without arms or wheels, texture, and so on. Hence, when programming a machine, this latter definition brings a high computational cost that drastically slows down the speed of the learning (and pushes away the idea of learning as fast as humans do). In that case, the machines/robots should be able to learn directly by acting and perceiving the consequences of their actions on the object/person.

Finally, from a more low-level aspect, even shape recognition is strongly connected to our motor experience. Viviani and Stucchi (1992) demonstrated that when they showed a participant a point light performing a perfect circle, as soon as this point slowed down at the upper and lower parts of this circle, the participant did not perceive the trajectory as a circle any longer, but as an ellipse. This perceptual mistake is explained by the fact that we perceive the shape of an object based on the way we draw it (in drawing a circle, we move with a constant speed, whereas in drawing an ellipse, we slow down at the two opposite extremities). Typically, handwriting learning (often cited by the authors) is based not only on learning visually the shape of the letters, but also mainly on global sensorimotor learning of perceiving (vision) and acting (writing, drawing). Once again, this example indicates that machines/robots should be able to understand an object or the reaction of a person based on how they have acted on that object/person.

Therefore, to design machines that learn as fast as humans, we need to make them able to (1) learn through a perception-action paradigm, (2) perceive and react to the movements of other agents or to the object on which they are acting, and (3) learn to understand what his or her or its actions mean.

ACKNOWLEDGMENT

This work was supported by the Dynamics of Interactions, Rhythmicity, Action and Communication (DIRAC), a project funded by the Agence Nationale de la Recherche (Grant ANR 13-ASTR-0018-01).

The argument for single-purpose robots

doi:10.1017/S0140525X17000231, e274

Daniel E. Moerman

University of Michigan—Dearborn, Ypsilanti, MI 48198.

dmoerman@umich.edu naeb.brit.org

Abstract: The argument by Lake et al. to create more human-like robots is, first, implausible and, second, undesirable. It seems implausible to me that a robot might have friends, fall in love, read Foucault, prefer Scotch to Bourbon, and so on. It seems undesirable because we already have 7 billion people on earth and don't really need more.

This commentary addresses the issue of Human-Like Machines (HLMs), which Lake et al. would like to be able to do more than have "object recognition" and play "video games, and board games" (abstract). They would like a machine "to learn or think like a person" (sect. 1, para. 3). I argue that people do vastly more than this: they interact, communicate, share, and collaborate; they use their learning and thinking to "behave"; they experience complex emotions. I believe that these authors have a far too limited sense of what "human-like" behavior is. The kinds of behavior I have in mind include (but are certainly not limited to) these:

1. Drive with a friend in a stick shift car from LA to Vancouver, and on to Banff…
2. Where, using a fly he or she tied, with a fly rod he or she made, he or she should be able to catch a trout which…
3. He or she should be able to clean, cook, and share with a friend.
4. He or she should have a clear gender identity, clearly recognizing what gender he or she is, and understanding the differences between self and other genders. (Let's decide our HLM was manufactured to be, and identifies as, "male.")
5. He should be able to fall in love, get married, and reproduce. He might wish to vote; he should be able to pay taxes. I'm not certain if he could be a citizen.
6. He should be able to read *Hop on Pop* to his 4-year-old, helping her to get the idea of reading. He should be able to read it to her 200 times. He should be able to read and understand Foucault, Sahlins, Hinton, le Carré, Erdrich, Munro, and authors like them. He should enjoy reading. He should be able to write a book, like *Hop on Pop*, or like Wilder's *The Foundations of Mathematics*.
7. He should be able to have irreconcilable differences with his spouse, get divorced, get depressed, get psychological counseling, get better, fall in love again, remarry, and enjoy his grandchildren. He should be able to detect by scent that the baby needs to have her diaper changed. Recent research indicates that the human nose can discriminate more than one trillion odors (Bushdid et al. 2014). Our HLM should at least recognize a million or so. He should be able to change a diaper and to comfort and calm a crying child. And make mac and cheese.
8. He should be able to go to college, get a B.A. in Anthropology, then a Ph.D., get an academic job, and succeed in teaching the complexities of kinship systems to 60 undergraduates.
9. He should be able to learn to play creditable tennis, squash, baseball, or soccer, and enjoy it into his seventies. He should be able to get a joke. (Two chemists go into a bar. The first says,

"I'll have an H_2O." The second says, "I'll have an H_2O too." The second guy dies.) He should be able both to age and to die.

10. He should be able to know the differences between Scotch and Bourbon, and to develop a preference for one or the other, and enjoy it occasionally. Same for wine.

I'm human, and I can do, or have done, all those things (except die), which is precisely why I think this is a fool's errand. I think it is a terrible idea to develop robots that are like humans. There are 7 billion humans on earth already. Why do we need fake humans when we have so many real ones? The robots we have now are (primarily) extremely useful single-function machines that can weld a car together in minutes, 300 a day, and never feel like, well, a robot, or a rivethead (Hamper 2008).

Even this sort of robot can cause lots of problems, as substantial unemployment in industry can be attributed to them. They tend to increase productivity and reduce the need for workers (Baily & Bosworth 2014). If that's what single-purpose (welding) robots can do, imagine what a HLM could do. If you think it might not be a serious problem, read Philip K. Dick's story, *Do Androids Dream Electric Sheep* (Dick 1968), or better yet, watch Ridley Scott's film *Blade Runner* (Scott 2007) based on Dick's story. The key issue in this film is that HLMs are indistinguishable from ordinary humans and are allowed legally to exist only as slaves. They don't like it. Big trouble ensues. (Re number 6, above, our HLM should probably *not* enjoy Philip Dick or *Blade Runner*.)

What kinds of things should machines be able to do? Jobs inimical to the human condition. Imagine an assistant fireman which could run into a burning building and save the 4-year-old reading Dr. Seuss. There is work going on to develop robotic devices – referred to as exoskeletons – that can help people with profound spinal cord injuries to walk again (Brenner 2016). But this is only reasonable if the device helps the patient go where he wants to go, not where the robot wants to go. There is also work going on to develop robotic birds, or orniothopters, among them the "Nano Hummingbird" and the "SmartBird." Both fly with flapping wings (Mackenzie 2012). The utility of these creatures is arguable; most of what they can do could probably be done with a $100 quad-copter drone. (Our HLM should be able to fly a quad-copter drone. I can.)

Google recently reported significant improvements in language translation as a result of the adoption of a neural-network approach (Lewis-Kraus 2016; Turovsky 2016). Many users report dramatic improvements in translations. (My own experience has been less positive.) This is a classic single-purpose "robot" that can help translators, but no one ought to rely on it alone.

In summary, it seems that even with the development of large neural-network style models, we are far from anything in *Blade Runner*. It will be a long time before we can have an HLM that can both display a patellar reflex and move the pieces in a chess game. And that, I think, is a very good thing.

Autonomous development and learning in artificial intelligence and robotics: Scaling up deep learning to human-like learning

doi:10.1017/S0140525X17000243, e275

Pierre-Yves Oudeyer

Inria and Ensta Paris-Tech, 33405 Talence, France.

pierre-yves.oudeyer@inria.fr http://www.pyoudeyer.com

Abstract: Autonomous lifelong development and learning are fundamental capabilities of humans, differentiating them from current deep learning systems. However, other branches of artificial intelligence have designed crucial ingredients towards autonomous learning: curiosity and intrinsic motivation, social learning and natural interaction with peers, and embodiment. These mechanisms guide exploration and autonomous choice of goals, and integrating them with deep learning opens stimulating perspectives.

Deep learning (DL) approaches made great advances in artificial intelligence, but are still far from human learning. As argued convincingly by Lake et al., differences include human capabilities to learn causal models of the world from very few data, leveraging compositional representations and priors like intuitive physics and psychology. However, there are other fundamental differences between current DL systems and human learning, as well as technical ingredients to fill this gap that are either superficially, or not adequately, discussed by Lake et al.

These fundamental mechanisms relate to *autonomous development and learning*. They are bound to play a central role in artificial intelligence in the future. Current DL systems require engineers to specify manually a task-specific objective function for every new task, and learn through offline processing of large training databases. On the contrary, humans learn autonomously open-ended repertoires of skills, deciding for themselves which goals to pursue or value and which skills to explore, driven by intrinsic motivation/curiosity and social learning through natural interaction with peers. Such learning processes are incremental, online, and progressive. Human child development involves a progressive increase of complexity in a curriculum of learning where skills are explored, acquired, and built on each other, through particular ordering and timing. Finally, human learning happens in the physical world, and through bodily and physical experimentation, under severe constraints on energy, time, and computational resources.

In the two last decades, the field of Developmental and Cognitive Robotics (Asada et al. 2009; Cangelosi and Schlesinger 2015), in strong interaction with developmental psychology and neuroscience, has achieved significant advances in computational modeling of mechanisms of autonomous development and learning in human infants, and applied them to solve difficult artificial intelligence (AI) problems. These mechanisms include the interaction between several systems that guide active exploration in large and open environments: curiosity, intrinsically motivated reinforcement learning (Barto 2013; Oudeyer et al. 2007; Schmidhuber 1991) and goal exploration (Baranes and Oudeyer 2013), social learning and natural interaction (Chernova and Thomaz 2014; Vollmer et al. 2014), maturation (Oudeyer et al. 2013), and embodiment (Pfeifer et al. 2007). These mechanisms crucially complement processes of incremental online model building (Nguyen and Peters 2011), as well as inference and representation learning approaches discussed in the target article.

Intrinsic motivation, curiosity and free play. For example, models of how motivational systems allow children to choose which goals to pursue, or which objects or skills to practice in contexts of free play, and how this can affect the formation of developmental structures in lifelong learning have flourished in the last decade (Baldassarre and Mirolli 2013; Gottlieb et al. 2013). In-depth models of intrinsically motivated exploration, and their links with curiosity, information seeking, and the "child-as-a-scientist" hypothesis (see Gottlieb et al. [2013] for a review), have generated new formal frameworks and hypotheses to understand their structure and function. For example, it was shown that intrinsically motivated exploration, driven by maximization of learning progress (i.e., maximal improvement of predictive or control models of the world; see Oudeyer et al. [2007] and Schmidhuber [1991]) can self-organize long-term developmental structures, where skills are acquired in an order and with timing that share fundamental properties with human development (Oudeyer and Smith 2016). For example, the structure of early infant vocal development self-organizes spontaneously from such intrinsically motivated exploration, in interaction with the physical properties of the vocal systems (Moulin-Frier et al. 2014). New experimental paradigms in psychology and neuroscience were recently developed and support these hypotheses (Baranes et al. 2014; Kidd 2012).

These algorithms of intrinsic motivation are also highly efficient for multitask learning in high-dimensional spaces. In robotics, they allow efficient stochastic selection of parameterized experiments and goals, enabling incremental collection of data and learning of skill models, through automatic and online curriculum learning. Such active control of the growth of complexity enables robots with high-dimensional continuous action spaces to learn omnidirectional locomotion on slippery surfaces and versatile manipulation of soft objects (Baranes and Oudeyer 2013) or hierarchical control of objects through tool use (Forestier and Oudeyer 2016). Recent work in deep reinforcement learning has included some of these mechanisms to solve difficult reinforcement learning problems, with rare or deceptive rewards (Bellemare et al. 2016; Kulkarni et al. 2016), as learning multiple (auxiliary) tasks in addition to the target task simplifies the problem (Jaderberg et al. 2016). However, there are many unstudied synergies between models of intrinsic motivation in developmental robotics and deep reinforcement learning systems; for example, curiosity-driven selection of parameterized problems/goals (Baranes and Oudeyer 2013) and learning strategies (Lopes and Oudeyer 2012) and combinations between intrinsic motivation and social learning, for example, imitation learning (Nguyen and Oudeyer 2013), have not yet been integrated with deep learning.

Embodied self-organization. The key role of physical embodiment in human learning has also been extensively studied in robotics, and yet it is out of the picture in current deep learning research. The physics of bodies and their interaction with their environment can spontaneously generate structure guiding learning and exploration (Pfeifer and Bongard 2007). For example, mechanical legs reproducing essential properties of human leg morphology generate human-like gaits on mild slopes without any computation (Collins et al. 2005), showing the guiding role of morphology in infant learning of locomotion (Oudeyer 2016). Yamada et al. (2010) developed a series of models showing that hand-face touch behaviours in the foetus and hand looking in the infant self-organize through interaction of a non-uniform physical distribution of proprioceptive sensors across the body with basic neural plasticity loops. Work on low-level muscle synergies also showed how low-level sensorimotor constraints could simplify learning (Flash and Hochner 2005).

Human learning as a complex dynamical system. Deep learning architectures often focus on inference and optimization. Although these are essential, developmental sciences suggested many times that learning occurs through complex dynamical interaction among systems of inference, memory, attention, motivation, low-level sensorimotor loops, embodiment, and social interaction. Although some of these ingredients are part of current DL research, (e.g., attention and memory), the integration of other key ingredients of autonomous learning and development opens stimulating perspectives for scaling up to human learning.

Human-like machines: Transparency and comprehensibility

doi:10.1017/S0140525X17000255, e276

Piotr M. Patrzyk, Daniela Link, and Julian N. Marewski

Faculty of Business and Economics, University of Lausanne, Quartier UNIL-Dorigny, Internef, CH-1015 Lausanne, Switzerland

piotr.patrzyk@unil.ch daniela.link@unil.ch
julian.marewski@unil.ch

Abstract: Artificial intelligence algorithms seek inspiration from human cognitive systems in areas where humans outperform machines. But on what level should algorithms try to approximate human cognition? We argue that human-like machines should be designed to make decisions in transparent and comprehensible ways, which can be achieved by accurately mirroring human cognitive processes.

How to build human-like machines? We agree with the authors' assertion that "reverse engineering human intelligence can usefully inform artificial intelligence and machine learning" (sect. 1.1, para. 3), and in this commentary we offer some suggestions concerning the direction of future developments. Specifically, we posit that human-like machines should not only be built to match humans in performance, but also to be able to make decisions that are both *transparent* and *comprehensible* to humans.

First, we argue that human-like machines need to decide and act in transparent ways, such that humans can readily understand how their decisions are made (see Arnold & Scheutz 2016; Indurkhya & Misztal-Radecka 2016; Mittelstadt et al. 2016). Behavior of artificial agents should be predictable, and people interacting with them ought to be in a position that allows them to intuitively grasp how those machines decide and act the way they do (Malle & Scheutz 2014). This poses a unique challenge for designing algorithms.

In current neural networks, there is typically no intuitive explanation for *why* a network reached a particular decision given received inputs (Burrell 2016). Such networks represent statistical pattern recognition approaches that lack the ability to capture agent-specific information. Lake et al. acknowledge this problem and call for structured cognitive representations, which are required for classifying social situations. Specifically, the authors' proposal of an "intuitive psychology" is grounded in the *naïve utility calculus* framework (Jara-Ettinger et al. 2016). According to this argument, algorithms should attempt to build a causal understanding of observed situations by creating representations of agents who seek rewards and avoid costs in a rational way.

Putting aside extreme examples (e.g., killer robots and autonomous vehicles), let us look at the more ordinary artificial intelligence task of scene understanding. Cost-benefit–based inferences about situations such as the one depicted in the leftmost picture in Figure 6 of Lake et al. will likely conclude that one agent has a desire to kill the other, and that he or she values higher the state of the other being dead than alive. Although we do not argue this is incorrect, a human-like classification of such a scene would rather reach the conclusion that the scene depicts either a legal execution or a murder. The returned alternative depends on the viewer's inferences about agent-specific characteristics. Making such inferences requires going beyond the attribution of simple goals – one needs to make assumptions about the roles and obligations of different agents. In the discussed example, although both a sheriff and a contract killer would have the same goal to end another person's life, the difference in their identity would change the human interpretation in a significant way.

We welcome the applicability of naïve utility calculus for inferring simple information concerning agent-specific variables, such as goals and competence level. At the same time, however, we point out some caveats inherent to this approach. Humans interacting with the system will likely expect a justification of why it has picked one interpretation rather than another, and algorithm designers might want to take this into consideration.

This leads us to our second point. Models of cognition can come in at least two flavors: (1) *As-if models*, which only aspire to achieve human-like performance on a specific task (e.g., classifying images), and (2) *process models*, which seek both to achieve human-like performance and to accurately reproduce the cognitive operations humans actually perform (classifying images by combining pieces of information in a way humans do). We believe that the task of creating human-like machines ought to be grounded in existing process models of cognition. Indeed, investigating human information processing is helpful for ensuring that generated decisions are comprehensible (i.e., that they follow human reasoning patterns).

Why is it important that machine decision mechanisms, in addition to being transparent, actually mirror human cognitive processes in a comprehensible way? In the social world, people often judge agents not only according to the agents' final decisions, but also according to the process by which they have arrived at these (e.g., Hoffman et al. 2015). It has been argued that the process of human decision making does not typically involve rational utility maximization (e.g., Hertwig & Herzog 2009). This, in turn, influences how we expect other people to make decisions (Bennis et al. 2010). To the extent that one cares about the social applications of algorithms and their interactions with people, considerations about transparency and comprehensibility of decisions become critical.

Although as-if models relying on cost-benefit analysis might be reasonably transparent and comprehensible, for example, when problems are simple and do not involve moral considerations, this might not always be the case. Algorithm designers need to ensure that the underlying process will be acceptable to the human observer. What research can be drawn up to help build transparent and comprehensible mechanisms?

We argue that one source of inspiration might be the research on *fast-and-frugal heuristics* (Gigerenzer & Gaissmaier 2011). Simple strategies such as fast-and-frugal trees (e.g., Hafenbrädl et al. 2016) might be well suited to providing justifications for decisions made in social situations. Heuristics not only are meant to capture *ecologically rational* human decision mechanisms (see Todd & Gigerenzer 2007), but also are transparent and comprehensible (see Gigerenzer 2001). Indeed, these heuristics possess a clear structure composed of simple if-then rules specifying (1) how information is searched within the search space, (2) when information search is stopped, and (3) how the final decision is made based upon the information acquired (Gigerenzer & Gaissmaier 2011).

These simple decision rules have been used to model and aid human decisions in numerous tasks with possible moral implications, for example, in medical diagnosis (Hafenbrädl et al. 2016) or classification of oncoming traffic at military checkpoints as hostile or friendly (Keller & Katsikopoulos 2016). We propose that the same heuristic principles might be useful to engineer autonomous agents that behave in a human-like way.

ACKNOWLEDGMENTS

D.L. and J.N.M acknowledge the support received from the Swiss National Science Foundation (Grants 144413 and 146702).

Intelligent machines and human minds

doi:10.1017/S0140525X17000267, e277

Elizabeth S. Spelke[a] and Joseph A. Blass[b]

[a]*Department of Psychology, Harvard University, Cambridge, MA 02138;* [b]*Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL 60208.*

spelke@wjh.harvard.edu joeblass@u.northwestern.edu

https://software.rc.fas.harvard.edu/lds/research/spelke/elizabeth-spelke/

http://qrg.northwestern.edu/people/Blass

Abstract: The search for a deep, multileveled understanding of human intelligence is perhaps *the* grand challenge for 21st-century science, with broad implications for technology. The project of building machines that think like humans is central to meeting this challenge and critical to efforts to craft new technologies for human benefit.

A century of research on human brains and minds makes three things clear. First, human cognition can be understood only if it is studied at multiple levels, from neurons to concepts to computations (Marr 1982/2010). Second, human and animal brains/minds are highly similar. Indeed, most of what we have discovered about our own capacities comes from multileveled studies of other animals (e.g., Hubel & Wiesel, 1959), suggesting that an understanding of human cognition is achievable. Third, the project of understanding human cognition has a long way to go. We have learned a lot about *what* we know *when* and what our brains are made of, but not *how* or *why* we know, think, and learn as we do.

Research on cognition in human infancy provides a case in point. Infants represent key geometric properties of the navigable layout, track objects as solid, continuously movable bodies, and endow others with goals and causal powers in ways that are highly similar to those of other inexperienced animals (e.g., Spelke & Lee 2012). These abilities are not shaped by encounters with the postnatal environment: precocial and controlled-reared animals exhibit them the first time they move through a navigable space (e.g., Chiandetti et al. 2014; Wills et al. 2010), track an object over occlusion (e.g., Regolin et al. 1995), or encounter another animal (e.g., Mascalzoni et al. 2010). Moreover, the basic ways that infants and animals understand space, objects, and goal-directed action remain central to our intuitive thinking as adults (e.g., Doeller & Burgess 2008) and to the brain systems that support it (e.g., Doeller et al. 2008; 2010).

None of these findings should be surprising or controversial. Human cognitive and neural architecture is unlikely to differ radically from that of other animals, because evolution proceeds by modifying what is already present. Abilities to represent space, objects, and other agents are unlikely to be entirely learned, because most animals need to get some problems right the first time they arise, including finding their way home, distinguishing a supporting surface from a teetering rock, avoiding predators, and staying with their group. And innate knowledge is unlikely to be overturned by later learning, because core knowledge captures fundamental properties of space, objects, and agency and because learning depends on prior knowledge.

Despite these findings, we do not know how human knowledge originates and grows, and a wealth of approaches to this question are rightly being pursued. One class of models, however, cannot plausibly explain the first steps of learning and development in any animal: deep learning systems whose internal structure is determined by analyzing massive amounts of data beyond any human scale. Human learning is fast and effective, in part, because it builds on cognitive and neural systems by which we understand the world throughout our lives. That's one reason why the effort described by Lake et al., implemented in computational models and tested against the judgments of human adults, is important to the grand challenge of achieving a deep understanding of human intelligence. We think the biggest advances from this work are still to come, through research that crafts and tests such models in systems that begin with human core knowledge and then learn, as young children do, to map their surroundings, develop a taxonomy of object kinds, and reason about others' mental states.

Computational models of infant thinking and learning may foster efforts to build smart machines that are not only better at reasoning, but also better for us. Because human infants are the best learners on the planet and instantiate human cognition in its simplest natural state, a computational model of infants' thinking and learning could guide the construction of machines that are more intelligent than any existing ones. But equally importantly, and sometimes left out of the conversation, a better understanding of our own minds is critical to building information systems for human benefit. Whether or not such systems are designed to learn and think as we do, such an understanding will help engineers build machines that will best foster our own thinking and learning.

To take just one example from current technology that is already ubiquitous, consider mobile, GPS-guided navigation systems. These systems can choose the most efficient route to a destination, based on information not accessible to the user, and allowing users to get around in novel environments. A person with a GPS-enabled cell phone never needs to know where he or she is or how the environment is structured. Is such a device good for us? A wealth of research indicates that the systems that

guide active, independent navigation in humans and animals from the moment that they first begin to locomote independently are broadly involved in learning and memory (e.g., Squire 1992). Moreover, the brain activity observed during active navigation diminishes when the same trajectory is followed passively (O'Keefe & Nadel 1978). How are these systems affected by the use of devices that do our navigating for us? If the cognitive and brain sciences could answer such questions in advance, researchers could design intelligent devices that both eliminate unnecessary cognitive burdens and provide needed cognitive exercise. Without a deeper understanding of how and why we learn and remember what we do, however, the designers of current technologies are working in the dark, even when they design devices to aid navigation, one of our best-understood cognitive functions (e.g., O'Keefe 2014; Moser et al. 2008).

Working in the dark posed less of a problem in past centuries. When each new tool that humans invented had limited function, and its use spread slowly, tools could be evaluated and modified by trial and error, without benefit of scientific insights into the workings of our minds. Today's tools, however, have multiple functions, whose workings are opaque to the end user. Technological progress is dizzyingly rapid, and even small advances bring sweeping, worldwide changes to people's lives. To design future machines for human benefit, researchers in all of the information technologies need to be able to foresee the effects that their inventions will have on us. And as Lake et al. observe, such foresight comes only with understanding. A great promise of human-inspired artificial intelligence, beyond building smarter machines, is to join neuroscience and cognitive psychology in meeting the grand challenge of understanding the nature and development of human minds.

The fork in the road

doi:10.1017/S0140525X17000279, e278

Robert J. Sternberg

Department of Human Development, College of Human Ecology, Cornell University, Ithaca, NY 14853.

robert.sternberg@cornell.edu www.robertjsternberg.com

Abstract: Machines that learn and think like people should simulate how people really think in their everyday lives. The field of artificial intelligence originally traveled down two roads, one of which emphasized abstract, idealized, rational thinking and the other, which emphasized the emotionally charged and motivationally complex situations in which people often find themselves. The roads should have converged but never did. That's too bad.

> Two roads diverged in a wood, and I—
> I took the one less traveled by,
> And that has made all the difference.
>
> —Robert Frost, *The Road Not Taken*
>
> *When you come to a fork in the road, take it.*
>
> —Yogi Berra

Lake and his colleagues have chosen to build "machines that learn and think like people." I beg to differ. Or perhaps it is a matter of what one means by learning and thinking like people. Permit me to explain. Early in the history of artificial intelligence (AI) and simulation research, investigators began following two different roads. The roads might potentially have converged, but it has become more and more apparent from recent events that they have actually diverged.

One road was initiated by pioneers like Newell et al. (1957), Winograd (1972), Minsky and Papert (1987), Minsky (2003), and Feigenbaum and Feldman (1995). This road was based on understanding people's *competencies* in learning and thinking. Investigators taking this road studied causal reasoning, game playing, language acquisition, intuitive physics, and people's understanding of block worlds. Today, Anderson's ACT-R perhaps provides the most comprehensive simulation model (Anderson et al. 2004).

A second road was taken by pioneers like Colby (1975) with PARRY, a computer program simulating a paranoid, Abelson and Carroll (1965) with their True Believer program, and Weizenbaum (1966) with his ELIZA non-directive psychotherapy program. The idea in this research was to understand people's often suboptimal *performances* in learning and thinking. These programs recognized that people are often emotional, a-rational, and function at levels well below their capabilities.

Many of these ideas have been formalized in recent psychological research. For example, Stanovich (2009) has shown that rationality and intelligence are largely distinct. Mayer and Salovey (1993) have shown the importance of emotional intelligence to people's thinking, and Sternberg (1997) has argued both for the importance of practical intelligence and for its relative independence from analytical or more academic aspects of intelligence.

The two roads of AI/simulation research might have converged with comprehensive models that comfortably incorporate aspects of both optimal and distinctly suboptimal performance. They haven't. At the time, Abelson, Colby, and others worked on their models of what was at best a-rational, and at worst wholly irrational, thinking. The work seemed a bit quirky and off the beaten track – perhaps a road not worth following very far. That was then

The 2016 presidential election has upended any assumption that everyday people think along the lines that Lake and his colleagues have pursued. Whether one is a Republican or a Democrat, it would be hard to accept this election process as representing anything other than seriously deficient and even defective thinking. The terms *learning* and *thinking* seem almost too complimentary to describe what went on. To some people the 2016 election was a frightening portent of a dystopia to come.

The first road, that of Lake et al., is of human cognition divorced from raw emotions, of often self-serving motivations and illogic that characterize much of people's everything thinking. On this view, people are more or less rational "machines." One might think that it is only stupid people (Sternberg 2002; 2004) who think and act foolishly. But smart people are as susceptible to foolish thinking as are not so smart people, or even more susceptible, because they do not realize they can think and act foolishly.

The United States, and indeed the world, seems to be entering a new and uncharted era of populism and appeals by politicians not to people's intellects, but to their basest emotions. Unless our models of learning and thinking help us understand how those appeals can succeed, and how we can counter them and help people become wiser (Sternberg & Jordan 2005), the models we create will be academic, incomplete, and, at worst, wrong-headed. The field came to a fork in the road and took it, but to where?

Avoiding frostbite: It helps to learn from others

doi:10.1017/S0140525X17000280, e279

Michael Henry Tessler, Noah D. Goodman, and Michael C. Frank

Department of Psychology, Stanford University, Stanford, CA 94305.

mtessler@standford.edu ngoodman@standford.edu
mcfrank@stanford.edu stanford.edu/~mtessler/
noahgoodman.net stanford.edu/~mcfrank/

Abstract: Machines that learn and think like people must be able to learn from others. Social learning speeds up the learning process and – in combination with language – is a gateway to abstract and unobservable information. Social learning also facilitates the accumulation of

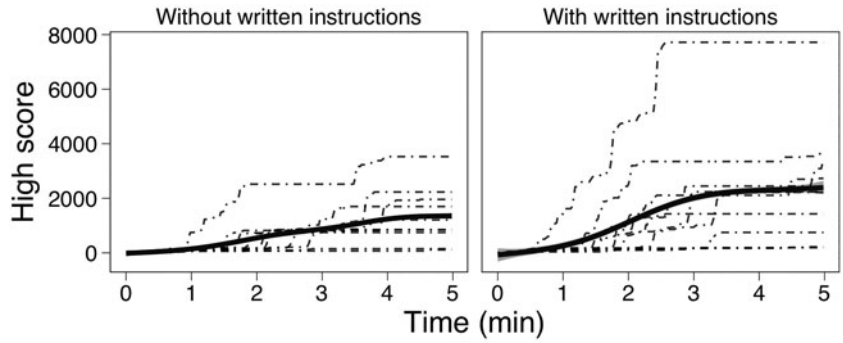

Figure 1 (Tessler et al.). Score trajectories for players in the game Frostbite over time. The two panels depict results with and without instructions on the abstract structure of the game.

knowledge across generations, helping people and artificial intelligences learn things that no individual could learn in a lifetime.

Causality, compositionality, and learning-to-learn – the future goals for artificial intelligence articulated by Lake et al. – are central for human learning. But these abilities alone would not be enough to avoid frostbite on King William Island in the Arctic Archipelago. You need to know how to hunt seals, make skin clothing, and manage dog sleds, and these skills are not easy to acquire from the environment alone. But if the Netsilik Inuit people taught them to you, your chances of surviving a winter would be dramatically improved (Lambert 2011). Similar to a human explorer, an artificial intelligence (AI) learning to play video games like Frostbite should take advantage of the rich knowledge available from other people. Access to this knowledge requires the capacity for social learning, both a critical prerequisite for language use and a gateway in itself to cumulative cultural knowledge.

Learning from other people helps you learn with fewer data. In particular, humans learn effectively even from "small data" because the social context surrounding the data is itself informative. Dramatically different inferences can result from what is ostensibly the same data in distinct social contexts or even with alternative assumptions about the same context (Shafto et al. 2012). The flexibility of the social inference machinery in humans turns small signals into weighty observations: Even for young children, ambiguous word-learning events become informative through social reasoning (Frank & Goodman 2014), non-obvious causal action sequences become "the way you do it" when presented pedagogically (Buchsbaum et al. 2011), and complex machines can become single-function tools when a learner is taught just one function (Bonawitz et al. 2011).

Learning from others comes in many forms. An expert may tolerate onlookers, a demonstrator may slow down when completing a particularly challenging part of the task, and a teacher may actively provide pedagogical examples and describe them with language (Csibra & Gergely 2009; Kline 2015). Informative demonstrations may be particularly useful for procedural learning (e.g., hunting seals, learning to play Frostbite). Language, however, is uniquely powerful in its ability to convey information that is abstract or difficult to observe, or information that otherwise does not have a way of being safely acquired such as learning that certain plants are poisonous or how to avoid frostbite (Gelman 2009). Studying social learning is an important part of studying language learning (Goodman & Frank 2016); both should be top priorities for making AIs learn like people.

Focusing on Lake et al.'s key example, you can even learn the game Frostbite with fewer data when you learn it from other people. We recruited 20 participants from Amazon's Mechanical Turk to play Frostbite for 5 minutes. Half of the participants were given written instructions about the abstract content of the game, adapted directly from the caption of Figure 2 in the target article. The other half were not given this information. (Everybody was told that you move the agent using the arrow keys.) Learners who were told about the abstract structure of the game learned to play the game more quickly and achieved higher overall scores ($M = 2440$) than the group without written instructions ($M = 1333$) (Figure 1). The highest score for those without linguistic instructions was 3530 points, achieved after about 4 minutes of play. By comparison, the highest score achieved with linguistic instructions was 7720 points, achieved after 2 minutes of play. Indeed, another group (including some authors of the target article) recently found a similar pattern of increased performance in Frostbite as a result of social guidance (Tsividis et al. 2017).

Learning from others also does more than simply "speed up" learning about the world. Human knowledge seems to accumulate across generations, hence permitting progeny to learn in one lifetime what no generation before them could learn (Boyd et al., 2011; Tomasello, 1999). We hypothesize that language – and particularly its flexibility to refer to abstract concepts – is key to faithful transmission of knowledge, between individuals and through generations. Human intelligence is so difficult to match because we stand on the shoulders of giants. AIs need to "ratchet" up their own learning, by communicating knowledge efficiently within and across generations. Rather than be subject to a top-down hive mind, intelligent agents should retain their individual intellectual autonomy, and innovate new solutions to problems based on their own experience and what they have learned from others. The important discoveries of a single AI could then be shared, and we believe language is the key to this kind of cultural transmission. Cultural knowledge could then accumulate within both AI and human networks.

In sum, learning from other people should be a high priority for AI researchers. Lake et al. hope to set priorities for future research in AI, but fail to acknowledge the importance of learning from language and social cognition. This is a mistake: The more complex the task is, the more learning to perform like a human involves learning from other people.

Crossmodal lifelong learning in hybrid neural embodied architectures

doi:10.1017/S0140525X17000292, e280

Stefan Wermter, Sascha Griffiths, and Stefan Heinrich

Knowledge Technology Group, Department of Informatics, Universität Hamburg, Hamburg, Germany.

wermter@informatik.uni-hamburg.de

griffiths@informatik.uni-hamburg.de

heinrich@informatik.uni-hamburg.de

https://www.informatik.uni-hamburg.de/~wermter/

https://www.informatik.uni-hamburg.de/~griffiths/
https://www.informatik.uni-hamburg.de/~heinrich/

Abstract: Lake et al. point out that grounding learning in general principles of embodied perception and social cognition is the next step in advancing artificial intelligent machines. We suggest it is necessary to go further and consider lifelong learning, which includes developmental learning, focused on embodiment as applied in developmental robotics and neurorobotics, and crossmodal learning that facilitates integrating multiple senses.

Artificial intelligence has recently been seen as successful in a number of domains, such a playing chess or Go, recognising handwritten characters, or describing visual scenes in natural language. Lake et al. discuss these kinds of breakthroughs as a big step for artificial intelligence, but raise the question how we can build machines that learn like people? We can find an indication in a survey of mind perception (Gray et al. 2007), which is the "amount of mind" people are willing to attribute to others. Participants judged machines to be high on agency but low on experience. We attribute this to the fact that computers are trained on individual tasks, often involving a single modality such as vision or speech, or a single context such as classifying traffic signs, as opposed to interpreting spoken and gestured utterances. In contrast, for people, the "world" essentially appears as a multimodal stream of stimuli, which unfold over time. Therefore, we suggest that the next paradigm shift in intelligent machines will have to include processing the "world" through lifelong and crossmodal learning. This is important because people develop problem-solving capabilities, including language processing, over their life span and via interaction with the environment and other people (Elman 1993, Christiansen and Chater 2016). In addition, the learning is embodied, as developing infants have a body-rational view of the world, but also seem to apply general problem-solving strategies to a wide range of quite different tasks (Cangelosi and Schlesinger 2015).

Hence, we argue that the proposed principles or "start-up software" are coupled tightly with general learning mechanisms in the brain. We argue that these conditions inherently enable the development of distributed representations of knowledge. For example, in our research, we found that architectural mechanisms, like different timings in the information processing in the cortex, foster compositionality that in turn enables both the development of more complex body actions and the development of language competence from primitives (Heinrich 2016). These kinds of distributed representations are coherent with the cognitive science on embodied cognition. Lakoff and Johnson (2003), for example, argue that people describe personal relationships in terms of the physical sensation of temperature. The transfer from one domain to the other is plausible, as an embrace or handshake between friends or family members, for example, will cause a warm sensation for the participants. These kinds of temperature exchanging actions are supposed to be signs of people's positive feelings towards each other (Hall 1966). The connection between temperature sensation and social relatedness is argued to reflect neural "bindings" (Gallese and Lakoff 2005). The domain knowledge that is used later in life can be derived from the primitives that are encountered early in childhood, for example, in interactions between infants and parents, and is referred to as *intermodal synchrony* (Rohlfing and Nomikou 2014). As a further example, our own research shows that learning, which is based on crossmodal integration, like the integration of real sensory perception on low and on intermediate levels (as suggested for the superior colliculus in the brain), can enable both super-additivity and dominance of certain modalities based on the tasks (Bauer et al. 2015).

In developing machines, approaches such as transfer learning and zero-shot learning are receiving increasing attention, but are often restricted to transfers from domain to domain or from modality to modality. In the domain case, this can take the form of a horizontal transfer, in which a concept in one domain is learned and then transferred to another domain within the same modality. For example, it is possible to learn about affect in speech and to transfer that model of affect to music (Coutinho et al. 2014). In the modality case, one can vertically transfer concepts from one modality to another. This could be a learning process in which language knowledge is transferred to the visual domain (Laptev 2008; Donahue 2015). However, based on the previous crossmodal integration in people, we must look into combinations of both, such that transferring between domains is not merely switching between two modalities, but integrating into both. Therefore, machines must exploit the representations that form when integrating multiple modalities that are richer than the sum of the parts. Recent initial examples include (1) understanding continuous counting expressed in spoken numbers from learned spatial differences in gestural motor grounding (Ruciński 2014) and (2) classifying affective states' audiovisual emotion expressions via music, speech, facial expressions, and motion (Barros and Wermter 2016).

Freeing learning from modalities and domains in favour of distributed representations, and reusing learned representations in the next individual learning tasks, will enable a larger view of learning to learn. Having underlying hybrid neural embodied architectures (Wermter et al. 2005) will support horizontal and vertical transfer and integration. This is the "true experience" machines need to learn and think like people. All in all, Lake et al. stress the important point of grounding learning in general principles of embodied perception and social cognition. Yet, we suggest it is still necessary to go a step further and consider lifelong learning, which includes developmental learning, focused on embodiment as applied in developmental robotics and neurorobotics, and crossmodal learning, which facilitates the integration of multiple senses.

Authors' Response

Ingredients of intelligence: From classic debates to an engineering roadmap

doi:10.1017/S0140525X17001224, e281

Brenden M. Lake,[a] Tomer D. Ullman,[b,c] Joshua B. Tenenbaum,[b,c] and Samuel J. Gershman[c,d]

[a]*Department of Psychology and Center for Data Science, New York University, New York, NY 10011;* [b]*Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139;* [c]*The Center for Brains Minds and Machines, Cambridge, MA 02139;* [d]*Department of Psychology and Center For Brain Science, Harvard University, Cambridge, MA 02138*

brenden@nyu.edu tomeru@mit.edu jbt@mit.edu
gershman@fas.harvard.edu http://cims.nyu.edu/~brenden/
http://www.mit.edu/~tomeru/ http://web.mit.edu/cocosci/josh.html
http://gershmanlab.webfactional.com/index.html

Abstract: We were encouraged by the broad enthusiasm for building machines that learn and think in more human-like ways. Many commentators saw our set of key ingredients as helpful, but there was disagreement regarding the origin and structure of those ingredients. Our response covers three main dimensions of this disagreement: nature versus nurture, coherent theories versus theory fragments, and symbolic versus sub-symbolic representations. These dimensions align with classic debates in artificial intelligence and cognitive science, although, rather than embracing these debates, we emphasize ways of moving beyond them. Several commentators saw our set of key ingredients as incomplete and offered a wide range of

additions. We agree that these additional ingredients are important in the long run and discuss prospects for incorporating them. Finally, we consider some of the ethical questions raised regarding the research program as a whole.

R1. Summary

We were pleased to see so many thoughtful commentaries and critiques in response to our target article. The project of "building machines that learn and think like people" will require input and insight from a broad range of disciplines, and it was encouraging that we received responses from experts in artificial intelligence (AI), machine learning, cognitive psychology, cognitive development, social psychology, philosophy, robotics, and neuroscience. As to be expected, there were many differences in perspective and approach, but before turning to those disagreements we think it is worth starting with several main points of agreement.

First, we were encouraged to see broad enthusiasm for the general enterprise and the opportunities it would bring. Like us, many researchers have been inspired by recent AI advances to seek a better computational understanding of human intelligence, and see this project's potential for driving new breakthroughs in building more human-like intelligence in machines. There were notable exceptions: A few respondents focused more on the potential risks and harms of this effort, or questioned its whole foundations or motivations. We return to these issues at the end of this response.

Most commenters also agreed that despite rapid progress in AI technologies over the last few years, machine systems are still not close to achieving human-like learning and thought. It is not merely a matter of scaling up current systems with more processors and bigger data sets. Fundamental ingredients of human cognition are missing, and fundamental innovations must be made to incorporate these ingredients into any kind of general-purpose, human-like AI.

Our target article articulated one vision for making progress toward this goal. We argued that human-like intelligence will come from machines that build models of the world – models that support explanation and understanding, prediction and planning, and flexible generalization for an open-ended array of new tasks – rather than machines that merely perform pattern recognition to optimize performance in a previously specified task or set of tasks.

We outlined a set of key cognitive ingredients that could support this approach, which are missing from many current AI systems (especially those based on deep learning), but could add great value: the "developmental start-up software" of intuitive physics and intuitive psychology, and mechanisms for rapid model learning based on the principles of compositionality, causality, and learning-to-learn (along with complementary mechanisms for efficient inference and planning with these models). We were gratified to read that many commentators found these suggested cognitive ingredients useful: "We agree … on their list of 'key ingredients' for building human-like intelligence" **(Botvinick, Barrett, Battaglia, de Freitas, Kumaran, Leibo, Lillicrap, Modayil, Mohamed, Rabinowitz, Rezende, Santoro, Schaul, Summerfield, Wayne, Weber, Wierstra, Legg, & Hassabis [Botvinick et al.],** abstract**):** "We entirely agree with the central thrust of the article" **(Davis & Marcus,** para. 1**):** "Causality, compositionality, and learning-to-learn … are central for human learning" **(Tessler, Goodman, & Frank [Tessler et al.],** para. 1**):** "Their ideas of 'start-up software' and tools for rapid model learning … help pinpoint the sources of general, flexible intelligence" **(Dennet & Lambert,** para. 1).

This is not to say that there was universal agreement about our suggested ingredients. Our list was carefully chosen but not meant to be complete, and many commentators offered additional suggestions: emotion **(Clark**; **Güss & Dörner)**, embodiment and action **(Baldassarre, Santucci, Cartoni, & Caligiore**; [**Baldassarre et al.**]; **MacLennan**; **Marin & Mostafaoui**; **Oudeyer**; **Wermter, Griffiths, & Heinrich [Wermter et al.]**), social and cultural learning (**Clegg & Corriveau**; **Dennett & Lambert**; **Tessler et al.**; Marin & Mostafaoui), and open-ended learning through intrinsic motivation (Baldassarre et al.; Güss & Dörner; Oudeyer; Wermter et al.). We appreciate these suggested additions, which help paint a richer and more complete picture of the mind and the ingredients of human intelligence. We discuss prospects for incorporating them into human-like AI systems in Section 5.

The main dimensions of disagreement in the commentaries revolved around how to implement our suggested ingredients in building AI: To what extent should they be explicitly built in, versus expected to emerge? What is their real content? How integrated or fragmented is the mind's internal structure? And what form do they take? How are these capacities represented in the mind or instantiated in the brain, and what kinds of algorithms or data structures should we be looking to in building an AI system?

Perhaps, unsurprisingly, these dimensions tended to align with classic debates in cognitive science and AI, and we found ourselves being critiqued from all sides. The first dimension is essentially the *nature versus nurture* debate (Section 2), and we were charged with advocating both for too much nature **(Botvinick et al.**; **Clegg & Corriveau**; **Cooper)** and too little **(Spelke & Blass)**. The second dimension relates to whether human mental models are better characterized in terms of *coherent theories versus theory fragments* (Section 3): We were criticized for positing theory-forming systems that were too strong **(Chater & Oaksford**; **Davis & Marcus**; **Livesey, Goldwater, & Colagiuri** [**Livesey et al.**]), but also too weak **(Dennett & Lambert)**. The third dimension concerns *symbolic versus sub-symbolic* representations (Section 4): To some commenters our proposal felt too allied with symbolic cognitive architectures **(Çağlar & Hanson**; **Hansen, Lampinen, Suri, & McClelland** [**Hansen et al.**]; **MacLennan)**. To others, we did not embrace symbols deeply enough **(Forbus & Gentner)**.

Some who saw our article through the lens of these classic debates, experienced a troubling sense of déjà vu. It was "Back to the Future: The Return of Cognitive Functionalism" for **Çağlar & Hanson**. For **Cooper**, it appeared "that cognitive science has advanced little in the last 30 years with respect to the underlying debates." We felt differently. We took this broad spectrum of reactions from commentators (who also, by and large, felt they agreed with our main points), as a sign that our field collectively might be looking to break out from these debates – to move in new directions that are not so easily classified as

just more of the same. It is understandable that many commentators would see our argument through the lens of these well-known and entrenched lines of argument, perhaps because we, as individuals, have contributed to them in previous publications. However, we wrote this target article, in part, because we felt it was vital to redefine this decades-long discussion in light of the recent progress in AI and machine learning.

Recent AI successes, on the one hand, make us optimistic about the project of building machines that learn and think like people. Working toward this goal seems much more plausible to many people than it did just a few years ago. At the same time, recent AI successes, when viewed from the perspective of a cognitive scientist, also highlight the gaps between machine and human intelligence. Our target article begins from this contrast: Whereas the driving force behind most of today's machine learning systems is sophisticated *pattern recognition*, scaled up to increasingly large and diverse data sets, the most impressive feats of human learning are better understood in terms of *model building*, often with much more limited data. We take the goal of building machines that can build models of the world as richly, as flexibly, and as quickly as humans can, as a worthy target for the next phase of AI research. Our target article lays out some of the key ingredients of human cognition that could serve as a basis for making progress toward that goal.

We explicitly tried to avoid framing these suggestions in terms of classic lines of argument that neural network researchers and other cognitive scientists have engaged in, to encourage more building and less arguing. With regards to nature versus nurture (sect. 2 of this article), we tried our best to describe these ingredients in a way that was "agnostic with regards to [their] origins" (target article, sect. 4, para. 2), but instead focused on their engineering value. We made this choice, not because we do not have views on the matter, but because we see the role of the ingredients as more important than their origins, for the next phase of AI research and the dialog between scientists and engineers. Whether learned, innate, or enriched, the fact that these ingredients are active so early in development, is a signal of their importance. They are present long before a person learns a new handwritten letter in the Character Challenge, or learns to play a new video game in the Frostbite Challenge (target article, sects. 3.1 and 3.2). AI systems could similarly benefit from utilizing these ingredients. With regards to symbolic versus sub-symbolic modeling (sect. 4 of this article), we think the ingredients could take either form, and they could potentially be added to symbolic architectures, sub-symbolic architectures, or hybrid architectures that transcend the dichotomy. Similarly, the model-building activities we describe could potentially be implemented in a diverse range of architectures, including deep learning. Regardless of implementation, demonstrations such as the Characters and Frostbite challenges show that people can rapidly build models of the world, and then flexibly reconfigure these models for new tasks without having to retrain. We see this as an ambitious target for AI that can be pursued in a variety of ways, and will have many practical applications (target article, sect. 6.2).

The rest of our response is organized as follows: The next three sections cover in detail the main dimensions of debate regarding the origin and structure of our ingredients: nature versus nurture (sect. 2), coherent theories versus theory fragments (sect. 3), and symbolic versus sub-symbolic representations (sect. 4). Additional ingredients suggested by the commentators are covered in Section 5. We discuss insights from neuroscience and the brain in Section 6. We end by discussing the societal risks and benefits of building machines that learn and think like people, in the light of the ethical issues raised by some commentators (sect. 7).

R2. Nature versus nurture

As mentioned, our target article did not intend to take a strong stance on "nature versus nurture" or "designing versus learning" for how our proposed ingredients should come to be incorporated into more human-like AI systems. We believe this question is important, but we placed our focus elsewhere in the target article. The main thesis is that a set of ingredients – each with deep roots in cognitive science – would be powerful additions to AI systems in whichever way a researcher chooses to include them. Whether the ingredients are learned, built in, or enriched through learning, we see them as a primary goal to strive for when building the next generation of AI systems. There are multiple possible paths for developing AI systems with these ingredients, and we expect individual researchers will vary in the paths they choose for pursuing these goals.

Understandably, many of the commentators linked their views on the biological origin of our cognitive principles to their strategy for developing AI systems with these principles. In contrast to the target article and its agnostic stance, some commentators took a stronger nativist stance, arguing that aspects of intuitive physics, intuitive psychology, and causality are innate, and it would be valuable to develop AI systems that "begin with human core knowledge" **(Spelke & Blass,** para. 4). Other commentators took a stronger nurture stance, arguing that the goal should be to learn these core ingredients rather than build systems that start with them **(Botvinick et al.**; **Cooper)**. Relatedly, many commentators pointed out additional nurture-based factors that are important for human-like learning, such as social and cultural forms of learning **(Clegg & Corriveau**; **Dennet & Lambert**; **Marin & Mostafaoui**; **Tessler et al.)**. In the section that follows, we respond to the different suggestions regarding the origin of the key ingredients, leaving the discussion of additional ingredients, such as social learning, for Section 5.

The response from researchers at Google DeepMind **(Botvinick et al.)** is of particular interest because our target article draws on aspects of their recent work. We offered their work as examples of recent accomplishments in AI (e.g., Graves et al. 2016; Mnih et al. 2015; Silver et al. 2016). At the same time, we highlighted ways that their systems do not learn or think like people (e.g., the Frostbite Challenge), but could potentially be improved by aiming for this target and by incorporating additional cognitive ingredients. Botvinick et al.'s response suggests that there are substantial areas of agreement. In particular, they see the five principles as "a powerful set of target goals for AI research" (para. 1), suggesting similar visions of what future accomplishments in AI will look like, and what the required building blocks are for getting there.

https://doi.org/10.1017/S0140525X17000127

Botvinick et al. strongly emphasized an additional principle: Machines should learn for themselves with minimal hand engineering from their human designers. We agree this is a valuable principle to guide researchers seeking to build learning-based general AI systems, as DeepMind aims to. To the extent that this principle is related to our principle of "learning-to-learn," we also endorse it in building machines that learn and think like people. Children are born capable of learning for themselves everything they will ultimately learn, without the need for an engineer to tweak their representations or algorithms along the way. However, it is not clear that the goals of building general AI systems and building machines that learn like people always converge, and the best design approach might be correspondingly different. Human beings (and other animals) may be born genetically programmed with mechanisms that effectively amount to highly engineered cognitive representations or algorithms – mechanisms that enable their subsequent learning and learning-to-learn abilities. Some AI designers may want to emulate this approach, whereas others may not.

The differences between our views may also reflect a difference in how we prioritize a set of shared principles and how much power we attribute to learning-to-learn mechanisms. **Botvinick et al.** suggest – but do not state explicitly – that they prioritize learning with minimal engineering above the other principles (and, thus, maximize the role of learning-to-learn). Under this strategy, the goal is to develop systems with our other key ingredients (compositionality, causality, intuitive physics, and intuitive psychology), insofar as they can be learned from scratch without engineering them. In the short term, this approach rests heavily on the power of learning-to-learn mechanisms to construct these other aspects of an intelligent system. In cases where this strategy is not feasible, Botvinick et al. state their approach also licenses them to build in ingredients too, but (we assume) with a strong preference for learning the ingredients wherever possible.

Although these distinctions may seem subtle, they can have important consequences for research strategy and outcome. Compare DeepMind's work on the Deep Q-Network (Mnih et al. 2015) to the theory learning approach our target article advocated for tackling the Frostbite Challenge, or their work on one-shot learning in deep neural networks (Rezende et al. 2016; Santoro et al. 2016; Vinyals et al. 2016) and our work on Bayesian Program Learning (Lake et al. 2015a). DeepMind's approaches to these problems clearly learn with less initial structure than we advocate for, and also clearly have yet to approach the speed, flexibility, and richness of human learning, even in these constrained domains.

We sympathize with DeepMind's goals and believe their approach should be pursued vigorously, along with related suggestions by **Cooper** and **Hansen et al**. However, we are not sure how realistic it is to pursue all of our key cognitive ingredients as emergent phenomena (see related discussion in sect. 5 of the target article), using the learning-to-learn mechanisms currently on offer in the deep learning landscape. Genuine intuitive physics, intuitive psychology, and compositionality, are unlikely to emerge from gradient-based learning in a relatively generic neural network. Instead, a far more expensive evolutionary-style search over discrete architectural variants may be required (e.g., Real et al. 2017; Stanley & Miikkulainen 2002). This approach may be characterized as "building machines that *evolve* to learn and think like people," in that such an extensive search would presumably include aspects of both phylogeny and ontogeny. As discussed in Section 4.1 of the target article, children have a foundational understanding of physics (objects, substances, and their dynamics) and psychology (agents and their goals) early in development. Whether innate, enriched, or rapidly learned, it seems unlikely that these ingredients arise purely in ontogeny from an extensive structural search over a large space of cognitive architectures, with no initial bias toward building these kinds of structures. In contrast, our preferred approach is to explore both powerful learning algorithms and starting ingredients together.

Over the last decade, this approach has led us to the key ingredients that are the topic of the target article (e.g., Baker et al. 2009; 2017; Battaglia et al. 2013; Goodman et al. 2008; Kemp et al. 2007; Lake et al. 2015a; Ullman et al. 2009); we did not start with these principles as dogma. After discovering which representations, learning algorithms, and inference mechanisms appear especially powerful in combination with each other, it is easier to investigate their origins and generalize them so they apply more broadly. Examples of this strategy from our work include the grammar-based framework for discovering structural forms in data (Kemp & Tenenbaum 2008), and a more emergent approach for implicitly learning some of the same forms (Lake et al. 2016), as well as models of causal reasoning and learning built on the theory of causal Bayesian networks (Goodman et al. 2011; Griffiths & Tenenbaum 2005, 2009). This strategy has allowed us to initially consider a wider spectrum of models, without a priori rejecting those that do not learn everything from scratch. Once an ingredient is established as important, it provides important guidance for additional research on how it might be learned.

We have pursued this strategy primarily through structured probabilistic modeling, but we believe it can be fruitfully pursued using neural networks as well. As **Botvinick et al.** point out, this strategy would not feel out of place in contemporary deep learning research. Convolutional neural networks build in a form of translation invariance that proved to be highly useful for object recognition (Krizhevsky et al. 2012; LeCun et al. 1989), and more recent work has explored building various forms of compositionality into neural networks (e.g., Eslami et al. 2016; Reed & de Freitas 2016). Increasingly, we are seeing more examples of integrating neural networks with lower-level building blocks from classic psychology and computer science (see sect. 6 of target article): selective attention (Bahdanau et al. 2015; Mnih et al. 2014; Xu et al. 2015), augmented working memory (Graves et al. 2014; Grefenstette et al. 2015; Sukhbaatar et al. 2015; Weston et al. 2015b), and experience replay (McClelland et al. 1995; Mnih et al. 2015). AlphaGo has an explicit model of the game of Go and builds in a wide range of high level and game-specific features, including how many stones a move captures, how many turns since a move was played, the number of liberties, and whether a ladder will be successful or not (Silver et al. 2016). If researchers are willing to include these types of representations and ingredients, we hope they will also consider our higher level cognitive ingredients.

It is easy to miss fruitful alternative representations by considering only models with minimal assumptions, especially in cases where the principles and representations have strong empirical backing (as is the case with our suggested principles). In fact, **Botvinick et al.** acknowledge that intuitive physics and psychology may be exceptions to their general philosophy, and could be usefully built in, given their breadth of empirical support. We were gratified to see this, and we hope it is clear to them and like-minded AI researchers that our recommendations are to consider building in only a relatively small set of core ingredients that have this level of support and scope. Moreover, a purely tabula rasa strategy can lead to models that require unrealistic amounts of training experience, and then struggle to generalize flexibly to new tasks without retraining. We believe that has been the case so far for deep learning approaches to the Characters and Frostbite challenges.

R3. Coherent theories versus theory fragments

Beyond the question of where our core ingredients come from, there is the question of their content and structure. In our article, we argued for theory-like systems of knowledge and causally structured representations, in particular (but not limited to) early-emerging intuitive physics and intuitive psychology. This view builds on extensive empirical research showing how young infants organize the world according to general principles that allow them to generalize across varied scenarios (Spelke 2003; Spelke & Kinzler 2007), and on theoretical and empirical research applied to children and adults that sees human knowledge in different domains as explained by theory-like structures (Carey 2009; Gopnik et al. 2004; Murphy & Medin 1985; Schulz 2012b; Wellman & Gelman 1992; 1998).

Commentators were split over how rich and how theory-like (or causal) these representations really are in the human mind and what that implies for building human-like AI. **Dennett & Lambert** see our view of theories as too limited – useful for describing cognitive processes shared with animals, but falling short of many distinctively human ways of learning and thought. On the other hand, several commentators saw our proposal as too rich for much of human knowledge. **Chater & Oaksford** argue by analogy to case-law, that "knowledge has the form of a loosely inter-linked history of reusable fragments" (para. 6) rather than a coherent framework. They stress that mental models are often shallow, and **Livesey et al.** add that people's causal models are not only shallow, but also often wrong, and resistant to change (such as the belief that vaccines cause autism). **Davis & Marcus** similarly suggest that the models we propose for the core ingredients are too causal, too complete, and too narrow to capture all of cognitive reasoning: Telling cats from dogs does not require understanding their underlying biological generative process; telling that a tower will fall does not require specifying in detail all of the forces and masses at play along the trajectory; and telling that someone is going to call someone does not require understanding whom they are calling or why.

In our target article, although we emphasized a view of cognition as model building, we also argued that pattern recognition can be valuable and even essential – in particular, for enabling efficient inference, prediction, and learning in rich causal theories. We suggested that different behaviors might be best explained by one, or the other, or both. For example, identifying the presence in a scene of an object that we call a "fridge" may indeed be driven by pattern recognition. But representing that object as a heavy, rigid, inanimate entity, and the corresponding predictions and plans that representation allows, is likely driven by more general abstract knowledge about physics and objects, whose core elements are not tied down to extensive patterns of experience with particular categories of objects. We could just as well form this representation upon our first encounter with a fridge, without knowing what it is called or knowing anything about the category of artifacts that it is one instance of.

On the issue of rich versus shallow representations, whereas intuitive theories of physics and psychology may be rich in the range of generalizable inferences they support, these and other intuitive theories are shallow in another sense; they are far more shallow than the type of formal theories scientists aim to develop, at the level of base reality and mechanism. From the point of view of a physicist, a game engine representation of a tower of blocks falling down is definitely not, as **Davis & Marcus** describe it, a "physically precise description of the situation" (para. 4). A game engine representation is a simplification of the physical world; it does not go down to the molecular or atomic level, and it does not give predictions at the level of a nanosecond. It represents objects with simplified bounding shapes, and it can give coarse predictions for coarse time-steps. Also, although real physics engines are useful analogies for a mental representation, they are not one and the same, and finding the level of granularity of the mental physics engine (if it exists) is an empirical question. To the point about intuitive psychology, theories that support reasoning about agents and goals do not need to specify all of the moving mental or neural parts involved in planning, to make useful predictions and explanations about what an agent might do in a given situation.

Returning to the need for multiple types of models, and to the example of the fridge, **Chater & Oaksford** point to a significant type of reasoning not captured by either recognizing an image of a fridge or reasoning about its physical behavior as a heavy, inert object. Rather, they consider the shallow and sketchy understanding of how a fridge stays cold. Chater & Oaksford use such examples to reason that, in general, reasoning is done by reference to examplars. They place stored, fragmented exemplars in the stead of wide-scope and deep theories. However, we suggest that even the shallow understanding of the operation of a fridge may best be phrased in the language of a causal, generative model, albeit a shallow or incomplete one. That is, even in cases in which we make use of previously stored examples, these examples are probably best represented by a causal structure, rather than by external or superficial features. To use Chater & Oaksford's analogy, deciding which precedent holds in a new case relies on the nature of the offense and the constraining circumstances, not the surname of the plaintiff. In the same way that two letters are considered similar not because of a pixel-difference measure, but because of the similar strokes that created them, exemplar-based reasoning would rely on the structural similarity of causal models of a new example and stored fragments (Medin & Ortony

https://doi.org/10.1017/S0140525X17000127

1989). An interesting hypothesis is that shallow causal models or mini-theories could be filling their gaps with more general, data-driven statistical machinery, such as a causal model with some of its latent variables generated by neural networks. Another possibility is that some mini-causal theories are generated ad hoc and on the fly (Schulz, 2012a), and so it should not be surprising that they are sometimes ill-specified and come into conflict with one another.

Unlike more general and early developing core theories such as intuitive physics and intuitive psychology, these mini-theory fragments may rely on later-developing language faculties (Carey 2009). More generally, the early forms of core theories such as intuitive physics and psychology, may be, as **Dennett & Lambert** put it, "[bootstrapped] into reflective comprehension" (para. 3). Similar points have been made in the past by Carey (2009) and Spelke (Spelke 2003; Spelke & Kinzler 2007), among others, regarding the role of later-developing language in using domain-specific and core knowledge concepts to extend intuitive theories as well as to build formal theories. The principles of core knowledge by themselves, are not meant to fully capture the formal or qualitative physical understanding of electricity, heat, light, and sound (distinguishing it from from the qualitative reasoning that **Forbus & Gentner** discuss). But, if later-developing aspects of physical understanding are built on these early foundations, that may be one source of the ontological confusion and messiness that permeates our later intuitive theories as well as people's attempts to understand formal theories in intuitive terms (**Livesey et al.**). Electrons are not little colliding ping-pong balls; enzymes are not trying to achieve an aspiration. But our parsing of the world into regular-bounded objects and intentional agents produces these category errors, because of the core role objects and agents play in cognition.

R4. Symbolic versus sub-symbolic representations

Beyond the richness and depth of our intuitive theories, the nature of representation was hotly contested in other ways, for the purposes of both cognitive modeling and developing more human-like AI. A salient division in the commentaries was between advocates of "symbolic" versus "sub-symbolic" representations, or relatedly, those who viewed our work through the lens of "explicit" versus "implicit" representations or "rules" versus "associations." Several commentators thought our proposal relied too much on symbolic representations, especially because sub-symbolic distributed representations have helped facilitate much recent progress in machine learning **(Çağlar & Hanson**; **Hansen et al.**; **MacLennan)**. Other commentators argued that human intelligence rests on more powerful forms of symbolic representation and reasoning than our article emphasized, such as abstract relational representations and analogical comparison **(Forbus & Gentner)**.

This is a deeply entrenched debate in cognitive science and AI – one that some of us have directly debated in past articles (along with some of the commentators [e.g., Griffiths et al. 2010; McClelland et al. 2010]), and we are not surprised to see it resurfacing here. Although we believe that this is still an interesting debate, we also see that recent work in AI and computational modeling of human cognition has begun to move beyond it, in ways that could be valuable.

To this end, we suggested that pattern recognition versus model building – and the ability to rapidly acquire new models and then reconfigure these models for new tasks without having to retrain – is a useful way to view the wide gap between human and machine intelligence. Implementing AI systems with our key ingredients would be a promising route for beginning to bridge this gap. Although our proposal is not entirely orthogonal to the symbolic versus sub-symbolic debate, we do see it as importantly different. Genuine model-building capabilities could be implemented in fully symbolic architectures or in a range of architectures that combine minimal symbolic components (e.g., objects, relations, agents, goals) with compositionality and sub-symbolic representation.

These ingredients could also be implemented in an architecture that does not appear to have symbols in any conventional sense – one that advocates of sub-symbolic approaches might even call non-symbolic – although we expect that advocates of symbolic approaches would point to computational states, which are effectively functioning as symbols. We do not claim to be breaking any new ground with these possibilities; the theoretical landscape has been well explored in philosophy of mind. We merely want to point out that our set of key ingredients is not something that should trouble people who feel that symbols are problematic. On the contrary, we hope this path can help bridge the gap between those who see symbols as essential, and those who find them mysterious or elusive.

Of our suggested ingredients, compositionality is arguably the most closely associated with strongly symbolic architectures. In relation to the above points, it is especially instructive to discuss how close this association has to be, and how much compositionality could be achievable within approaches to building intelligent machines that might not traditionally be seen as symbolic.

Hansen et al. argue that there are inherent limitations to "symbolic compositionality" that deep neural networks help overcome. Although we have found traditional symbolic forms of compositionality to be useful in our work, especially in interaction with other key cognitive ingredients such as causality and learning-to-learn (e.g., Goodman et al. 2011; 2015; Lake et al. 2015a), there may be other forms of compositionality that are useful for learning and thinking like humans, and easier to incorporate into neural networks. For example, neural networks designed to understand scenes with multiple objects (see also Fig. 6 of our target article), or to generate globally coherent text (such as a recipe), have found simple forms of compositionality to be extremely useful (e.g., Eslami et al. 2016; Kiddon et al. 2016). In particular, "objects" are minimal symbols that can support powerfully compositional model building, even if implemented in an architecture that would otherwise be characterized as sub-symbolic (e.g., Eslami et al. 2016; Raposo et al. 2017). The notion of a physical object – a chunk of solid matter that moves as a whole, moves smoothly through space and time without teleporting, disappearing, or passing through other solid objects – emerges very early in development (Carey 2009). It is arguably the central representational construct of human beings' earliest intuitive physics, one of the first symbolic concepts in any domain that infants have access to, and likely shared with

many other animal species in some form (see target article, sect. 4.1.1). Hence, the "object" concept is one of the best candidates for engineering AI to start with, and a promising target for advocates of sub-symbolic approaches who might want to incorporate useful but minimal forms of symbols and compositionality into their systems.

Deep learning research is also beginning to explore more general forms of compositionality, often by utilizing hybrid symbolic and sub-symbolic representations. Differentiable neural computers (DNCs) are designed to process symbolic structures such as graphs, and they use a mixture of sub-symbolic neural network-style computation and symbolic program traces to reason with these representations (Graves et al. 2016). Neural programmer-interpreters (NPIs) begin with symbolic program primitives embedded in their architecture, and they learn to control the flow of higher-level symbolic programs that are constructed from these primitives (Reed & de Freitas 2016). Interestingly, the learned controller is a sub-symbolic neural network, but it is trained with symbolic supervision. These systems are very far from achieving the powerful forms of model building that we see in human intelligence, and it is likely that more fundamental breakthroughs will be needed. Still, we are greatly encouraged to see neural network researchers who are not ideologically opposed to the role of symbols and compositionality in the mind and, indeed, are actively looking for ways to incorporate these notions into their paradigm.

In sum, by viewing the impressive achievements of human learning as model building rather than pattern recognition, we hope to emphasize a new distinction, different from classic debates of symbolic versus sub-symbolic computation, rules versus associations, or explicit versus implicit reasoning. We would like to focus on people's capacity for learning flexible models of the world as a target for AI research – one that might be reached successfully through a variety of representational paradigms if they incorporate the right ingredients. We are pleased that the commentators seem to broadly support "model building" and our key ingredients as important goals for AI research. This suggests a path for moving forward together.

R5. Additional ingredients

Many commentators agreed that although our key ingredients were important, we neglected another obvious, crucial component of human-like intelligence. There was less agreement on which component we had neglected. Overlooked components included emotion **(Güss & Dörner**; **Clark)**; embodiment and action **(Baldassarre et al.**; **MacLennan**; **Marin & Mostafaoui**; **Oudeyer; Wermter et al.)**; learning from others through social and cultural interaction **(Clegg & Corriveau**; **Dennett & Lambert**; **Marin & Mostafaoui**; **Tessler et al.);** open-ended learning combined with the ability to set one's own goal (Baldassarre et al.; Oudeyer; Wermter et al.); architectural diversity **(Buscema & Sacco)**; dynamic network communication **(Graham)**; and the ability to get a joke **(Moerman)**.

Clearly, our recipe for building machines that learn and think like people was not complete. We agree that each of these capacities should figure in any complete scientific understanding of human cognition, and will likely be important for building artificial human-like cognition. There are likely other missing components as well. However, the question for us as researchers interested in the reverse engineering of cognition is: Where to start?

We focused on ingredients that were largely missing from today's deep learning AI systems, ones that were clearly crucial and present early in human development, and with large expected payoffs in terms of core AI problems. Importantly, for us, we also wanted to draw focus to ingredients that to our mind can be implemented in the relatively short term, given a concentrated effort. Our challenges are not meant to be AI-complete, but ones that can potentially be met in the next few years. For many of the suggestions the commentators made, it is hard (for us, at least) to know where to begin concrete implementation.

We do not mean that there have not been engineering advances and theoretical proposals for many of these suggestions. The commentators have certainly made progress on them, and we and our colleagues have also made theoretical and engineering contributions to some. But to do full justice to all of these missing components – from emotion to sociocultural learning to embodiment – there are many gaps that we do not know how to fill yet. Our aim was to set big goal posts on the immediate horizon, and we admit that there are others beyond. With these implementation gaps in mind, we have several things to say about each of these missing components.

R5.1. Machines that feel: Emotion

In popular culture, intelligent machines differ from humans in that they do not experience the basic passions that color people's inner life. To call someone *robotic* does not mean that they lack a good grasp of intuitive physics, intuitive psychology, compositionality, or causality. It means they, like the Tin Man, have no heart. Research on "mind attribution" has also borne out this distinction (Gray & Wegner 2012; Gray et al. 2007; Haslam 2006; Loughnan & Haslam 2007): Intelligent machines and robots score highly on the agency dimension (people believe such creatures can plan and reason), but low on the experience dimension (people believe they lack emotion and subjective insight). In line with this, **Güss & Dörner**; **Clark**; and **Sternberg** highlight emotion as a crucial missing ingredient in building human-like machines. As humans ourselves, we recognize the importance of emotion in directing human behavior, in terms of both understanding oneself and predicting and explaining the behavior of others. The challenge, of course, is to operationalize this relationship in computational terms. To us, it is not obvious how to go from evocative descriptions, such as "a person would get an 'uneasy' feeling when solution attempts do not result in a solution" (as observed by **Güss & Dörner**, para. 5), to a formal and principled implementation of unease in a decision-making agent. We see this as a worthwhile pursuit for developing more powerful and human-like AI, but we see our suggested ingredients as leading to concrete payoffs that are more attainable in the short term.

Nonetheless we can speculate about what it might take to structure a human-like "emotion" ingredient in AI, and how it would relate to the other ingredients we put forth. Pattern-recognition approaches (based on deep learning or other methods) have had some limited success in mapping between video and audio of humans to simple emotion labels like *happy* (e.g., Kahou et al. 2013).

https://doi.org/10.1017/S0140525X17000127

Sentiment analysis networks learn to map between text and its positive or negative valence (e.g., Socher et al. 2013). But genuine, human-like concepts or experiences of emotion will require more, especially more sophisticated model building, with close connections and overlap with the ingredient of intuitive psychology. Humans may have a "lay theory of emotion" (Ong et al. 2015) that allows them to reason about the causal processes that drive the experiences of frustration, anger, surprise, hate, and joy. That is, something like "achieving your goal makes you feel happy." This type of theory would also connect the underlying emotions to observable behaviors such as facial expressions (downward turned lips), action (crying), body posture (hunched shoulders), and speech ("It's nothing, I'm fine"). Moreover, as pointed out by **Güss & Dörner**, a concept of "anger" must include how it modulates perception, planning, and desires, touching on key aspects of intuitive psychology.

R5.2. Machines that act: Action and embodiment

One of the aspects of intelligence "not much stressed by Lake et al." was the importance of intelligence being "strongly embodied and situated," located in an acting physical body **(Baldassarre et al**., para. 4), with possible remedies coming in the form of "developmental robotics and neurorobotics" **(Oudeyer**; **Wermter et al.)**. This was seen by some commentators as more than yet-another-key-ingredient missing from current deep learning research. Rather, they saw it as an issue for our own proposal, particularly as it relates to physical causality and learning. Embodiment and acting on the real world provides an agent with "a foundation for its understanding of intuitive physics" **(MacLennan)**, and "any learning or social interacting is based on social motor embodiment." Even understanding what a chair is requires the ability to sit on it **(Marin & Mostafaoui)**.

We were intentionally agnostic in our original proposal regarding the way a model of intuitive physics might be learned, focusing instead on the existence of the ability, its theory-like structure, usefulness, and early emergence, and its potential representation as something akin to a mental game engine. It is an interesting question whether this representation can be learned only by passively viewing video and audio, without active, embodied engagement. In agreement with some of the commentators, it seems likely to us that such a representation in humans does come about – over a combination of both evolutionary and developmental processes – from a long history of agents' physical interactions with the world – applying their own forces on objects (perhaps somewhat haphazardly at first in babies), observing the resulting effects, and revising their plans and beliefs accordingly.

An intuitive theory of physics built on object concepts, and analogs of force and mass, would also benefit a physically realized robot, allowing it to plan usefully from the beginning, rather than bumbling aimlessly and wastefully as it attempts some model-free policies for interaction with its environment. An intuitive theory of physics can also allow the robot to imagine potential situations without going through the costly operation of carrying them out. Furthermore, unlike **MacLennan's** requirement that theories be open to discourse and communication, such a generative, theory-like representation does not need to be explicit and accessible in a communicative sense (target article, sect. 4). Instead, people may have no introspective insight into its underlying computations, in the same way that they have no introspective insight into the computations that go into recognizing a face.

To **MacLennan**'s point regarding the necessary tight coupling between an agent and a real environment: If a theory-like representation turns out to be the right representation, we do not see why it cannot be arrived at by virtual agents in a virtual environment, provided that they are provided with the equivalents of somatosensory information and the ability to generate the equivalent of forces. Agents endowed with a representation of intuitive physics may have calibration issues when transferred from a virtual environment to a situated and embodied robot, but it would likely not result in a complete breakdown of their physical understanding, any more than adults experience a total breakdown of intuitive physics when transferred to realistic virtual environments.

As for being situated in a physical body, although the mental game-engine representation has been useful in capturing people's reasoning about disembodied scenes (such as whether a tower of blocks on a table will fall down), it is interesting to consider extending this analogy to the existence of an agent's body and the bodies of other agents. Many games rely on some representation of the players, with simplified bodies built of "skeletons" with joint constraints. This type of integration would fit naturally with the long-investigated problem of pose estimation (Moeslund et al. 2006), which has recently been the target of discriminative deep learning networks (e.g., Jain et al. 2014; Toshev & Szegedy 2014). Here, too, we would expect a converging combination of structured representations and pattern recognition: That is, rather than mapping directly between image pixels and the target label *sitting*, there would be an intermediate simplified body-representation, informed by constraints on joints and the physical situation. This intermediate representation could in turn be categorized as *sitting* (see related hybrid architectures from recent years [e.g., Chen & Yuille 2014; Tompson et al. 2014]).

R5.3. Machines that learn from others: Culture and pedagogy

We admit that the role of sociocultural learning is, as **Clegg & Corriveau** put it, "largely missing from Lake et al.'s discussion of creating human-like artificial intelligence" (abstract). We also agree that this role is essential for human cognition. As the commentators pointed out, it is important both on the individual level, as "learning from other people helps you learn with fewer data" (**Tessler et al.**, para. 2) and also on the societal level, as "human knowledge seems to accumulate across generations" (Tessler et al., para. 5). Solving Frostbite is not only a matter of combining intuitive physics, intuitive psychology, compositionality, and learning-to-learn, but also a matter of watching someone play the game, or listening to someone explain it (Clegg & Corriveau; Tessler et al.), as we have shown in recent experiments (Tsividis et al. 2017).

Some of the commentators stressed the role of imitation and over-imitation in this pedagogical process **(Dennet & Lambert**; **Marin & Mostafaoui)**. Additionally, **Tessler et al.** focused more on language as the vehicle for this

learning, and framed the study of social learning as a part of language learning. Our only disagreement with Tessler et al. regarding the importance of language, is their contention that we "fail to acknowledge the importance of learning from language." We completely agree about the importance of understanding language for understanding cognition. However, we think that by understanding the early building blocks we discussed, we will be in a better position to formally and computationally understand language learning and use. For a fuller reply to this point, we refer the reader to Section 5 in the target article.

Beyond being an additional ingredient, **Clegg & Corriveau** suggest sociocultural learning may override some of the key ingredients we discuss. As they nicely put it, "although the developmental start-up software children begin with may be universal, early in development children's 'software updates' may be culturally-dependent. Over time, these updates may even result in distinct operating systems" (para. 4). Their evidence for this includes different culture-dependent time-courses for passing the false belief task, understanding fictional characters as such, and an emphasis on consensus-building (Corriveau et al. 2013; Davoodi et al. 2016; Liu et al. 2008). We see these differences as variations on, or additions to, the core underlying structure of intuitive psychology, which is far from monolithic in its fringes. The specific causes of a particular behavior posited by a 21st-century Western architect may be different from those of a medieval French peasant or a Roman emperor, but the parsing of behavior in terms of agents that are driven by a mix of desire, reasoning, and necessity, would likely remain the same, just as their general ability to recognize faces would likely be the same (Or as an emperor put it, "[W]hat is such a person doing, and why, and what is he saying, and what is he thinking of, and what is he contriving" [Aurelius 1937]). Despite these different stresses, we agree with Clegg & Corriveau that sociocultural learning builds upon the developmental start-up packages, rather than by starting with a relatively blank slate child that develops primarily through socio-cultural learning via language and communication (Mikolov et al. 2016).

R5.4. Machines that explore: Open-ended learning and intrinsic motivation

Several commentators **(Baldassarre et al.**; **Güss & Dörner**; **Oudeyer**; **Wermter et al.)** raised the challenge of building machines that engage in open-ended learning and exploration. Unlike many AI systems, humans (especially children) do not seem to optimize a supervised objective function; they explore the world autonomously, develop new goals, and acquire skill repertoires that generalize across many tasks. This challenge has been particularly acute for developmental roboticists, who must endow their robots with the ability to learn a large number of skills from scratch. It is generally infeasible to solve this problem by defining a set of supervised learning problems, because of the complexity of the environment and sparseness of rewards. Instead, roboticists have attempted to endow their robots with intrinsic motivation to explore, so that they discover for themselves what goals to pursue and skills to acquire.

We agree that open-ended learning is a hallmark of human cognition. One of our main arguments for why humans develop rich internal models is that these support the ability to flexibly solve an infinite variety of tasks. Acquisition of such models would be impossible if humans were not intrinsically motivated to acquire information about the world, without being tied to particular supervised tasks. The key question, in our view, is how to define intrinsic motivation in such a way that a learning system will seek to develop an abstract understanding of the world, populated by agents, objects, and events. Developmental roboticists tend to emphasize embodiment as a source of constraints: Robots need to explore their physical environment to develop sophisticated, generalizable sensory-motor skills. Some (e.g., **MacLennan**) argue that high-level competencies, such as intuitive physics and causality, are derived from these same low-level sensory-motor skills. As in the previous section, we believe that embodiment, although important, is insufficient: humans can use exploration to develop abstract theories that transcend particular sensors and effectors (e.g., Cook et al. 2011). For example, in our Frostbite Challenge, many of the alternative goals are not defined in terms of any particular visual input or motor output. A promising approach would be to define intrinsic motivation in terms of intuitive theories – autonomous learning systems that seek information about the causal relationships between agents, objects, and events. This form of curiosity would augment, not replace, the forms of lower-level curiosity necessary to develop sensory-motor skills.

R6. Insights from neuroscience and the brain

Our article did not emphasize neuroscience as a source of constraint on AI, not because we think it is irrelevant (quite the contrary), but because we felt that it was necessary to first crystallize the core ingredients of human intelligence at a computational level before trying to figure out how they are implemented in physical hardware. In this sense, we are advocating a mostly top-down route through the famous Marr levels of analysis, much as Marr himself did. This was unconvincing to some commentators **(Baldassarre et al.**; **George**; **Kriegeskorte & Mok**; **Marblestone, Wayne, & Kording)**. Surely it is necessary to consider neurobiological constraints from the start, if one wishes to build human-like intelligence?

We agree that it would be foolish to argue for cognitive processes that are in direct disagreement with known neurobiology. However, we do not believe that neurobiology in its current state provides many strong constraints of this sort. For example, **George** suggests that lateral connections in visual cortex indicate that the internal model used by the brain enforces contour continuity. This seems plausible, but it is not the whole story. We see the world in three dimensions, and there is considerable evidence from psychophysics that we expect the surfaces of objects to be continuous in three dimensions, even if such continuity violates two-dimensional contour continuity (Nakayama et al. 1989). Thus, the situation is more like the opposite of what George argues: a challenge for neuroscience is to explain how neurons in visual cortex enforce the three-dimensional continuity constraints we know exist from psychophysical research.

Kriegeskorte & Mok point to higher-level vision as a place where neural constraints have been valuable. They

https://doi.org/10.1017/S0140525X17000127

write that core object recognition has been "conquered" by brain-inspired neural networks. We agree that there has been remarkable progress on basic object recognition tasks, but there is still a lot more to understand scientifically and to achieve on the engineering front, even in visual object perception. Take, for example, the problem of occlusion. Because most neural network models of object recognition have no explicit representation of objects arranged in depth, they are forced to process occlusion as a kind of noise. Again, psychophysical evidence argues strongly against this: When objects pass behind an occluding surface, we do not see them as disappearing or becoming corrupted by a massive amount of noise (Kellman & Spelke 1983). A challenge for neuroscience is to explain how neurons in the ventral visual stream build a 3D representation of scenes that can appropriately handle occlusion. The analogous challenge exists in AI for brain-inspired artificial neural networks.

Further challenges, just in the domain of object perception, include perceiving multiple objects in a scene at once; perceiving the fine-grained shape and surface properties of novel objects for which one does not have a class label; and learning new object classes from just one or a few examples, and then generalizing to new instances. In emphasizing the constraints biology places on cognition, it is sometimes underappreciated to what extent cognition places strong constraints on biology.

R7. Coda: Ethics, responsibility, and opportunities

Your scientists were so preoccupied with whether or not they could, that they didn't stop to think if they should.

— Dr. Ian Malcom, *Jurassic Park*

Given recent progress, AI is now widely recognized as a source of transformative technologies, with the potential to impact science, medicine, business, home life, civic life, and society, in ways that improve the human condition. There is also real potential for more negative impacts, including dangerous side effects or misuse. Recognizing both the positive and negative potential has spurred a welcome discussion of ethical issues and responsibility in AI research. Along these lines, a few commentators questioned the moral and ethical aspects of the very idea of building machines that learn and think like people. **Moerman** argues that the project is both unachievable and undesirable and, instead, advocates for building useful, yet inherently limited "single-purpose" machines. As he puts it (para. 2), "There are 7 billion humans on earth already. Why do we need fake humans when we have so many real ones?" **Dennett & Lambert** worry that machines may become intelligent enough to be given control of many vital tasks, before they become intelligent or human-like enough to be considered responsible for the consequences of their behavior.

We believe that trying to build more human-like intelligence in machines could have tremendous benefits. Many of these benefits will come from progress in AI more broadly – progress that we believe would be accelerated by the project described in our target article. There are also risks, but we believe these risks are not, for the foreseeable future, existential risks to humanity, or uniquely new kinds of risks that will sneak up on us suddenly. For anyone worried that AI research may be making too much progress too quickly, we would remind them that the best machine-learning systems are still very far from achieving human-like learning and thought, in all of the ways we discussed in the target article. Superintelligent AIs are even further away, so far that we believe it is hard to plan for them, except in the most general sense. Without new insights, ingredients, and ideas – well beyond those we have written about – we think that the loftiest goals for AI will be difficult to reach. Nonetheless, we see the current debate on AI ethics as responsible and healthy, and we take **Dennett & Lambert**'s suggestion regarding AI copilots in that spirit.

Moerman's commentary fits well with many of these points: Simply scaling up current methods is unlikely to achieve anything like human intelligence. However, he takes the project of building more human-like learning machines to its logical extreme – building a doppelgänger machine that can mimic all aspects of being human, including incidental ones. Beyond rapid model building and flexible generalization, and even after adding the additional abilities suggested by the other commentators (sect. 5), Moerman's doppelgänger machine would still need the capability to get a joke, get a Ph.D., fall in love, get married, get divorced, get remarried, prefer Bourbon to Scotch (or vice versa), and so on. We agree that it is difficult to imagine machines will do all of these things any time soon. Nonetheless, the current AI landscape would benefit from more human-like learning – with its speed, flexibility, and richness – far before machines attempt to tackle many of the abilities that Moerman discusses. We think that this type of progress, even if only incremental, would still have far-reaching, practical applications (target article, sect. 6.2), and broader benefits for society.

Apart from advances in AI more generally, advances in human-like AI would bring additional unique benefits. Several commentators remarked on this. **Spelke & Blass** point out that a better understanding of our own minds will enable new kinds of machines that "can foster our thinking and learning" (para. 5). In addition, **Patrzyk, Link, & Marewski** expound on the benefits of "explainable AI," such that algorithms can generate human-readable explanations of their output, limitations, and potential failures (Doshi-Velez & Kim 2017). People often learn by constructing explanations (Lombrozo 2016, relating to our "model building"), and a human-like machine learner would seek to do so too. Moreover, as it pertains to human-machine interaction **(e.g., Dennett & Lambert)**, it is far easier to communicate with machines that generate human-understandable explanations than with opaque machines that cannot explain their decisions.

In sum, building machines that learn and think like people is an ambitious project, with great potential for positive impact: through more powerful AI systems, a deeper understanding of our own minds, new technologies for easing and enhancing human cognition, and explainable AI for easier communication with the technologies of the future. As AI systems become more fully autonomous and agentive, building machines that learn and think like people will be the best route to building machines that treat people the way people want and expect to be treated by others: with a sense of fairness, trust, kindness, considerateness, and intelligence.

References

[The letters "a" and "r" before author's initials stand for target article and response references, respectively]

Abelson, R. P. & Carroll, J. D. (1965) Computer simulation of individual belief systems. *The American behavioral scientist (pre-1986)* 8(9):24–30. [RJS]

Aitchison, L. & Lengyel, M. (2016) The Hamiltonian brain: Efficient probabilistic inference with excitatory-inhibitory neural circuit dynamics. *PLoS Computational Biology* 12(12):e1005186. [NK]

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C. & Qin, Y. (2004) An integrated theory of the mind. *Psychological Review* 111:1036–60. [RJS]

Anderson, M. L. (2003) Embodied cognition: A field guide. *Artificial Intelligence* 149 (1):91–130. [GB]

Andrychowicz, M., Denil, M., Gomez, S., Hoffman, M. W., Pfau, D., Schaul, T., Shillingford, B. & de Freitas, N. (2016). Learning to learn by gradient descent by gradient descent. Presented at the 2016 Neural Information Processing Systems conference, Barcelona, Spain, December 5–10, 2016. In: *Advances in neural information processing systems 29 (NIPS 2016)*, ed. D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon & R. Garnett, pp. 3981–89). Neural Information Processing Systems. [aBML, MB]

Anselmi, F., Leibo, J. Z., Rosasco, L., Mutch, J., Tacchetti, A. & Poggio, T. (2016) Unsupervised learning of invariant representations. *Theoretical Computer Science* 633:112–21. [aBML]

Ansermin, E., Mostafaoui, G., Beausse, N. & Gaussier, P. (2016). Learning to synchronously imitate gestures using entrainment effect. In: *From Animals to Animats 14: Proceedings of the 14th International Conference on Simulation of Adaptive Behavior (SAB 2016), Aberystwyth, United Kingdom, August 23–26, 2016*, ed. Tuci E, Giagkos A, Wilson M, Hallam J, pp. 219–31. Springer. [LM]

Arbib, M. A. & Fellous, J. M. (2004) Emotions: From brain to robot. *Trends in Cognitive Science* 8(12):554–61. [KBC]

Arnold, T. & Scheutz, M. (2016) Against the moral Turing test: Accountable design and the moral reasoning of autonomous systems. *Ethics and Information Technology* 18(2):103–15. doi:10.1007/s10676-016-9389-x. [PMP]

Asada, M. (2015) Development of artificial empathy. *Neuroscience Research* 90:41–50. [KBC]

Asada, M., Hosoda, K., Kuniyoshi, Y., Ishiguro, H., Inui, T., Yoshikawa, Y. & Yoshida, C. (2009) Cognitive developmental robotics: A survey. *IEEE Transactions on Autonomous Mental Development* 1(1):12–34. [P-YO]

Aurelius, M. (1937) *Meditations*, transl. G. Long. P. F. Collier & Son. [rBML]

Bach, J. (2009) *Principles of synthetic intelligence. PSI: An architecture of motivated cognition*. Oxford University Press. [CDG]

Bahdanau, D., Cho, K. & Bengio, Y. (2015) Neural machine translation by jointly learning to align and translate. Presented at the International Conference on Learning Representations (ICLR), San Diego, CA, May 7–9, 2015. *arXiv preprint 1409.0473*. Available at: http://arxiv.org/abs/1409.0473v3. [arBML]

Baillargeon, R. (2004) Infants' physical world. *Current Directions in Psychological Science* 13:89–94. [aBML]

Baillargeon, R., Li, J., Ng, W. & Yuan, S. (2009) An account of infants physical reasoning. In: *Learning and the infant mind*, ed. A. Woodward & A. Neeham, pp. 66–116. Oxford University Press. [aBML]

Baily, M. N. & Bosworth, B. P. (2014) US manufacturing: Understanding its past and its potential future. *The Journal of Economic Perspectives* 28(1):3–25. [DEM]

Baker, C. L., Jara-Ettinger, J., Saxe, R. & Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour* 1:0064. [rBML]

Baker, C. L., Saxe, R. & Tenenbaum, J. B. (2009) Action understanding as inverse planning. *Cognition* 113(3):329–49. [arBML]

Baldassarre, G. (2011) What are intrinsic motivations? A biological perspective. In: *Proceedings of the International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob-2011)*, ed. A. Cangelosi, J. Triesch, I. Fasel, K. Rohlfing, F. Nori, P.-Y. Oudeyer, M. Schlesinger & Y. Nagai, pp. E1–8. IEEE. [GB]

Baldassarre, G., Caligiore, D. & Mannella, F. (2013a) The hierarchical organisation of cortical and basal-ganglia systems: A computationally-informed review and integrated hypothesis. In: *Computational and robotic models of the hierarchical organisation of behaviour*, ed. G. Baldassarre & M. Mirolli, pp. 237–70. Springer-Verlag. [GB]

Baldassarre, G., Mannella, F., Fiore, V. G., Redgrave, P., Gurney, K. & Mirolli, M. (2013b) Intrinsically motivated action-outcome learning and goal-based action recall: A system-level bio-constrained computational model. *Neural Networks* 41:168–87. [GB]

Baldassarre, G. & Mirolli, M., eds. (2013) *Intrinsically motivated learning in natural and artificial systems*. Springer. [GB, P-YO]

Baldassarre, G., Stafford, T., Mirolli, M., Redgrave, P., Ryan, R. M. & Barto, A. (2014) Intrinsic motivations and open-ended development in animals, humans, and robots: An overview. *Frontiers in Psychology* 5:985. [GB]

Baranes, A. & Oudeyer, P.-Y. (2013) Active learning of inverse models with intrinsically motivated goal exploration in robots. *Robotics and Autonomous Systems* 61(1):49–73. [P-YO]

Baranes, A. F., Oudeyer, P. Y. & Gottlieb, J. (2014) The effects of task difficulty, novelty and the size of the search space on intrinsically motivated exploration. *Frontiers in Neurosciences* 8:1–9. [P-YO]

Barros, P. & Wermter, S. (2016) Developing crossmodal expression recognition based on a deep neural model. *Adaptive Behavior* 24(5):373–96. [SW]

Barsalou, L. W. (1983) Ad hoc categories. *Memory & Cognition* 11(3):211–27. [aBML]

Barto, A. (2013) Intrinsic motivation and reinforcement learning. In: *Intrinsically motivated learning in natural and artificial systems*, ed. G. Baldassarre & M. Mirolli, pp. 17–47. Springer. [P-YO]

Bartunov, S. & Vetrov, D. P. (2016) Fast adaptation in generative models with generative matching networks. *arXiv preprint 1612.02192*. [SSH]

Bastos, A. M., Usrey, W. M., Adams, R. A., Mangun, G. R., Fries, P. & Friston, K. J. (2012) Canonical microcircuits for predictive coding. *Neuron* 76:695–711. http://doi.org/10.1016/j.neuron.2012.10.038. [aBML, DGe]

Bates, C. J., Yildirim, I., Tenenbaum, J. B. & Battaglia, P. W. (2015) Humans predict liquid dynamics using probabilistic simulation. In: *Proceedings of the 37th Annual Conference of the Cognitive Science Society, Pasadena, CA, July 22–25, 2015*, pp. 172–77. Cognitive Science Society. [aBML]

Battaglia, P., Pascanu, R., Lai, M. & Rezende, D. J. (2016) Interaction networks for learning about objects, relations and physics. Presented at the 2016 Neural Information Processing Systems conference, Barcelona, Spain, December 5–10, 2016. In: *Advances in neural information processing systems 29 (NIPS 2016)*, ed. D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon & R. Garnett, pp. 4502–10. Neural Information Processing Systems. [MB]

Battaglia, P. W., Hamrick, J. B. & Tenenbaum, J. B. (2013) Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences of the United States of America* 110(45):18327–32. [arBML, ED]

Baudiš, P. & Gailly, J.-l. (2012) PACHI: State of the art open source Go program. In: *Advances in computer games: 13th International Conference, ACG 2011, Tilburg, The Netherlands, November 20–22, 2011, Revised Selected Papers*, ed. H. Jaap van den Herik & A. Plast, pp. 24–38. Springer. [aBML]

Bauer, J., Dávila-Chacón, J. & Wermter, S. (2015) Modeling development of natural multi-sensory integration using neural self-organisation and probabilistic population codes. *Connection Science* 27(4):358–76. [SW]

Baxter, J. (2000) A model of inductive bias learning. *Journal of Artificial Intelligence Research* 12:149–98. [aBML]

Bayer, H. M. & Glimcher, P. W. (2005) Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron* 47:129–41. [aBML]

Bellemare, M., Srinivasan, S., Ostrovski, G., Schaul, T., Saxton, D. & Munos, R. (2016) Unifying count-based exploration and intrinsic motivation. Presented at the 2016 Neural Information Processing Systems conference, Barcelona, Spain, December 5–10, 2016. In: *Advances in neural information processing systems 29 (NIPS 2016)*, ed. D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon & R. Garnett, pp. 1471–79. Neural Information Processing Systems. [MB, P-YO]

Bellemare, M. G., Naddaf, Y., Veness, J. & Bowling, M. (2013) The arcade learning environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research* 47:253–79. [aBML]

Bengio, J. (2009) Learning deep architectures for AI. *Foundations and Trends in Machine Learning* 2(1):1–127. [MBu]

Bengio, Y. (2016) Machines who learn. *Scientific American* 314(6):46–51. [KBC]

Bennis, W. M., Medin, D. L. & Bartels, D. M. (2010) The costs and benefits of calculation and moral rules. *Perspectives on Psychological Science* 5(2):187–202. doi:10.1177/1745691610362354. [PMP]

Berdahl, C. H. (2010) A neural network model of Borderline Personality Disorder. *Neural Networks* 23(2):177–88. [KBC]

Berlyne, D. E. (1966) Curiosity and exploration. *Science* 153(3731):25–33. doi:10.1126/science.153.3731.25 [aBML, CDG]

Berthiaume, V. G., Shultz, T. R. & Onishi, K. H. (2013) A constructivist connectionist model of transitions on false-belief tasks. *Cognition* 126(3): 441–58. [aBML]

Berwick, R. C. & Chomsky, N. (2016) *Why only us: Language and evolution*. MIT Press. [aBML]

Bever, T. G. & Poeppel, D. (2010) Analysis by synthesis: A (re-) emerging program of research for language and vision. *Biolinguistics* 4:174–200. [aBML]

Bi, G.-Q. & Poo, M.-M. (2001) Synaptic modification by correlated activity: Hebb's postulate revisited. *Annual Review of Neuroscience* 24:139–66. [aBML]

Biederman, I. (1987) Recognition-by-components: A theory of human image understanding. *Psychological Review* 94(2):115–47. [aBML]

Bienenstock, E., Cooper, L. N. & Munro, P. W. (1982) Theory for the development of neuron selectivity: Orientation specificity and binocular interaction in visual cortex. *The Journal of Neuroscience* 2(1):32–48. [aBML]

Bienenstock, E., Geman, S. & Potter, D. (1997) Compositionality, MDL priors, and object recognition. Presented at the 1996 Neural Information Processing Systems conference, Denver, CO, December 2–5, 1996. In: *Advances in neural*

information processing systems 9, ed. M. C. Mozer, M. I. Jordan & T. Petsche, pp. 838–44. Neural Information Processing Systems Foundation. [aBML]
Blackburn, S. (1984) *Spreading the word: Groundings in the philosophy of language*. Oxford University Press. [NC]
Block, N. (1978) Troubles with functionalism. *Minnesota Studies in the Philosophy of Science* 9:261–325. [LRC]
Bloom, P. (2000) *How children learn the meanings of words*. MIT Press. [aBML]
Blumberg, M. S. (2005) *Basic instinct: The genesis of behavior*. Basic Books. [AHM]
Blundell, C., Uria, B., Pritzel, A., Li, Y., Ruderman, A., Leibo, J. Z., Rae, J., Wierstra, D. & Hassabis, D. (2016) Model-free episodic control. *arXiv preprint 1606.04460*. Available at: https://arxiv.org/abs/1606.04460. [aBML, MB]
Bobrow, D. G. & Winograd, T. (1977) An overview of KRL, a knowledge representation language. *Cognitive Science* 1:3–46. [aBML]
Boden, M. A. (1998) Creativity and artificial intelligence. *Artificial Intelligence* 103:347–56. [aBML]
Boden, M. A. (2006) *Mind as machine: A history of cognitive science*. Oxford University Press. [aBML]
Bonawitz, E., Denison, S., Griffiths, T. L. & Gopnik, A. (2014) Probabilistic models, learning algorithms, and response variability: Sampling in cognitive development. *Trends in Cognitive Sciences* 18:497–500. [aBML]
Bonawitz, E., Shafto, P., Gweon, H., Goodman, N. D., Spelke, E. & Schulz, L. (2011) The double-edged sword of pedagogy: Instruction limits spontaneous exploration and discovery. *Cognition* 120(3):322–30. Available at: http://doi.org/10.1016/j.cognition.2010.10.001. [MHT]
Bostrom, N. (2014) *Superintelligence: Paths, dangers, strategies*. Oxford University Press. ISBN 978-0199678112. [KBC]
Bottou, L. (2014) From machine learning to machine reasoning. *Machine Learning* 94(2):133–49. [aBML]
Botvinick, M. M. & Cohen, J. D. (2014) The computational and neural basis of cognitive control: Charted territory and new frontiers. *Cognitive Science* 38:1249–85. [MB]
Botvinick, M., Weinstein, A., Solway, A. & Barto, A. (2015) Reinforcement learning, efficient coding, and the statistics of natural tasks. *Current Opinion in Behavioral Sciences* 5:71–77. [MB]
Bouton, M. E. (2004) Context and behavioral processes in extinction. *Learning & Memory* 11:485–94. [aBML]
Boyd, R., Richerson, P. J. & Henrich, J. (2011) The cultural niche: Why social learning is essential for human adaptation. *Proceedings of the National Academy of Sciences of the United States of America* 108(suppl 2):10918–25. [MHT]
Braud, R., Mostafaoui, G., Karaouzene, A. & Gaussier, P. (2014). Simulating the emergence of early physical and social interactions: A developmental route through low level visuomotor learning. In: *From Animal to Animats 13: Proceedings of the 13th International Conference on Simulation of Adaptive Behavior, Castellon, Spain, July 2014*, ed. A. P. del Pobil, E. Chinalleto, E. Martinez-Martin, J. Hallam, E. Cervera & A. Morales, pp. 154–65. Springer. [LM]
Breazeal, C. & Scassellati, B. (2002). Robots that imitate humans. *Trends in Cognitive Sciences* 6(11):481–87. [LM]
Brenner, L. (2016) Exploring the psychosocial impact of Ekso Bionics Technology. *Archives of Physical Medicine and Rehabilitation* 97(10):e113. [DEM]
Briegel, H. J. (2012) On creative machines and the physical origins of freedom. *Scientific Reports* 2:522. [KBC]
Briggs, F. & Usrey, W. M. (2007) A fast, reciprocal pathway between the lateral geniculate nucleus and visual cortex in the macaque monkey. *The Journal of Neuroscience* 27(20):5431–36. [DG]
Buchsbaum, D., Gopnik, A., Griffiths, T. L. & Shafto, P. (2011) Children's imitation of causal action sequences is influenced by statistical and pedagogical evidence. *Cognition* 120(3):331–40. Available at: http://doi.org/10.1016/j.cognition.2010.12.001. [MHT]
Buckingham, D. & Shultz, T. R. (2000) The developmental course of distance, time, and velocity concepts: A generative connectionist model. *Journal of Cognition and Development* 1(3):305–45. [aBML]
Buesing, L., Bill, J., Nessler, B. & Maass, W. (2011) Neural dynamics as sampling: A model for stochastic computation in recurrent networks of spiking neurons. *PLoS Computational Biology* 7:e1002211. [aBML]
Burrell, J. (2016) How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society* 3(1):1–12. doi:10.1177/2053951715622512. [PMP]
Buscema, M. (1995) Self-reflexive networks: Theory – topology – Applications. *Quality and Quantity* 29(4):339–403. [MBu]
Buscema, M. (1998) Metanet*: The theory of independent judges. *Substance Use and Misuse* 32(2):439–61. [MBu]
Buscema, M. (2013) Artificial adaptive system for parallel querying of multiple databases. In: *Intelligent data mining in law enforcement analytics*, ed. M. Buscema & W. J. Tastle, pp. 481–511. Springer. [MBu]
Buscema, M., Grossi, E., Montanini, L. & Street, M. E. (2015) Data mining of determinants of intrauterine growth retardation revisited using novel algorithms generating semantic maps and prototypical discriminating variable profiles. *PLoS One* 10(7):e0126020. [MBu]
Buscema, M., Tastle, W. J. & Terzi, S. (2013) Meta net: A new meta-classifier family. In: *Data mining applications using artificial adaptive systems*, ed. W. J. Tastle, pp. 141–82. Springer. [MBu]
Buscema, M., Terzi, S. & Tastle, W.J. (2010). A new meta-classifier. In: *2010 Annual Meeting of the North American Fuzzy Information Processing Society (NAFIPS), Toronto, ON, Canada*, pp. 1–7. IEEE. [MBu]
Bushdid, C., Magnasco, M. O., Vosshall, L. B. & Keller, A. (2014) Humans can discriminate more than 1 trillion olfactory stimuli. *Science* 343(6177):1370–72. [DEM]
Caglar, L. R. & Hanson, S. J. (2016) Deep learning and attentional bias in human category learning. Poster presented at the Neural Computation and Psychology Workshop on Contemporary Neural Networks, Philadelphia, PA, August 8–10, 2016. [LRC]
Caligiore, D., Borghi, A., Parisi, D. & Baldassarre, G. (2010) TRoPICALS: A computational embodied neuroscience model of compatibility effects. *Psychological Review* 117(4):1188–228. [GB]
Caligiore, D., Pezzulo, G., Baldassarre, G., Bostan, A. C., Strick, P. L., Doya, K., Helmich, R. C., Dirkx, M., Houk, J., Jörntell, H., Lago-Rodriguez, A., Galea, J. M., Miall, R. C., Popa, T., Kishore, A., Verschure, P. F. M. J., Zucca, R. & Herreros, I. (2016) Consensus paper: Towards a systems-level view of cerebellar function: The interplay between cerebellum, basal ganglia, and cortex. *The Cerebellum* 16(1):203–29. doi: 10.1007/s12311-016-0763-3. [GB]
Calimera, A., Macii, E. & Poncino, M. (2013) The human brain project and neuromorphic computing. *Functional Neurology* 28(3):191–96. [KBC]
Cangelosi, A. & Schlesinger, M. (2015) *Developmental robotics: From babies to robots*. MIT Press. [P-YO, SW]
Cardon, A. (2006) Artificial consciousness, artificial emotions, and autonomous robots. *Cognitive Processes* 7(4):245–67. [KBC]
Carey, S. (1978) The child as word learner. In: *Linguistic theory and psychological reality*, ed. J. Bresnan, G. Miller & M. Halle, pp. 264–93. MIT Press. [aBML]
Carey, S. (2004) Bootstrapping and the origin of concepts. *Daedalus* 133(1):59–68. [aBML]
Carey, S. (2009) *The origin of concepts*. Oxford University Press. [arBML, KDF]
Carey, S. (2011) The origin of concepts: A précis. *Behavioral and Brain Sciences* 34 (03):113–62. [EJL]
Carey, S. & Bartlett, E. (1978) Acquiring a single new word. *Papers and Reports on Child Language Development* 15:17–29. [aBML]
Chavajay, P. & Rogoff, B. (1999) Cultural variation in management of attention by children and their caregivers. *Developmental Psychology* 35(4):1079. [JMC]
Chen, X. & Yuille, A. L. (2014) Articulated pose estimation by a graphical model with image dependent pairwise relations. In: *Advances in neural information processing systems 27 (NIPS 2014)*, ed. Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence & K. Q. Weinberger, pp. 1736–44. Neural Information Processing Systems Foundation. [rBML]
Chen, Z. & Klahr, D. (1999) All other things being equal: Acquisition and transfer of the control of variables strategy. *Child Development* 70(5):1098–120. [KDF]
Chernova, S. & Thomaz, A. L. (2014) *Robot learning from human teachers*. Synthesis lectures on artificial intelligence and machine learning. Morgan & Claypool. [P-YO]
Chi, M. T., Slotta, J. D. & De Leeuw, N. (1994) From things to processes: A theory of conceptual change for learning science concepts. *Learning and Instruction* 4 (1):27–43. [EJL]
Chiandetti, C., Spelke, E. S. & Vallortigara, G. (2014) Inexperienced newborn chicks use geometry to spontaneously reorient to an artificial social partner. *Developmental Science* 18(6):972–78. doi:10.1111/desc.12277. [ESS]
Chouard, T. (2016) The Go files: AI computer wraps up 4–1 victory against human champion. (Online; posted March 15, 2016.) [aBML]
Christiansen, M. H. & Chater, N. (2016) *Creating language: Integrating evolution, acquisition, and processing*. MIT Press. [NC, SW]
Churchland, M. M., Cunningham, J. P., Kaufman, M. T., Foster, J. D., Nuyujukian, P., Ryu, S. I. & Shenoy, K. V. (2012) Neural population dynamics during reaching. *Nature* 487:51–56. [GB]
Ciresan, D., Meier, U. & Schmidhuber, J. (2012) Multi-column deep neural networks for image classification. In: *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Providence, RI, June 16–21, 2012*, pp. 3642–49. IEEE. [aBML]
Clark, K. B. (2012) A statistical mechanics definition of insight. In: *Computational intelligence*, ed. A. G. Floares, pp. 139–62. Nova Science. ISBN 978-1-62081-901-2. [KBC]
Clark, K. B. (2014) Basis for a neuronal version of Grover's quantum algorithm. *Frontiers in Molecular Neuroscience* 7:29. [KBC]
Clark, K. B. (2015) Insight and analysis problem solving in microbes to machines. *Progress in Biophysics and Molecular Biology* 119:183–93. [KBC]
Clark, K. B. (in press-a) Classical and quantum Hebbian learning in modeled cognitive processing. *Frontiers in Psychology*. [KBC]
Clark, K. B. (in press-b) Neural field continuum limits and the partitioning of cognitive-emotional brain networks. *Molecular and Cellular Neuroscience*. [KBC]
Clark, K. B. (in press-c) Psychometric "Turing test" of general intelligences in social robots. *Information Sciences*. [KBC]

Clark, K. B. & Hassert, D. L. (2013) Undecidability and opacity of metacognition in animals and humans. *Frontiers in Psychology* 4:171. [KBC]

Cleeremans, A. (1993) *Mechanisms of implicit learning: Connectionist models of sequence processing*. MIT Press. [LRC]

Clegg, J. M., Wen, N. J. & Legare, C. H. (2017) Is non-conformity WEIRD? Cultural variation in adults' beliefs about children's competency and conformity. *Journal of Experimental Psychology: General* 146(3):428–41. [JMC]

Cohen, E. H. & Tong, F. (2015) Neural mechanisms of object-based attention. *Cerebral Cortex* 25(4):1080–92. http://doi.org/10.1093/cercor/bht303. [DGe]

Colagiuri, B., Schenk, L. A., Kessler, M. D., Dorsey, S. G. & Colloca, L. (2015) The placebo effect: from concepts to genes. *Neuroscience* 307:171–90. [EJL]

Colby, K. M. (1975) *Artificial paranoia: Computer simulation of paranoid processes*. Pergamon. [RJS]

Collins, A. G. E. & Frank, M. J. (2013) Cognitive control over learning: Creating, clustering, and generalizing task-set structure. *Psychological Review* 120 (1):190–229. [aBML]

Collins, S., Ruina, A., Tedrake, R. & Wisse, M. (2005) Efficient bipedal robots based on passive-dynamic walkers. *Science* 307(5712):1082–85. [P-YO]

Cook, C., Goodman, N. D. & Schulz, L. E. (2011) Where science starts: Spontaneous experiments in preschoolers' exploratory play. *Cognition* 120(3):341–49. [arBML]

Cooper, R. P. (2016) Executive functions and the generation of "random" sequential responses: A computational account. *Journal of Mathematical Psychology* 73:153–68. doi: 10.1016/j.jmp.2016.06.002. [RPK]

Correa-Chávez, M. & Rogoff, B. (2009) Children's attention to interactions directed to others: Guatemalan Mayan and European American patterns. *Developmental Psychology* 45(3):630. [JMC]

Corriveau, K. H. & Harris, P. L. (2010) Preschoolers (sometimes) defer to the majority when making simple perceptual judgments. *Developmental Psychology* 26:437–45. [JMC]

Corriveau, K. H., Kim, E., Song, G. & Harris, P. L. (2013) Young children's deference to a consensus varies by culture and judgment setting. *Journal of Cognition and Culture* 13(3–4):367–81. [JMC, rBML]

Coutinho, E., Deng, J. & Schuller, B. (2014) Transfer learning emotion manifestation across music and speech. In: *Proceedings of the 2014 International Joint Conference on Neural Networks (IJCNN), Beijing, China*. pp. 3592–98. IEEE. [SW]

Crick, F. (1989) The recent excitement about neural networks. *Nature* 337:129–32. [aBML]

Csibra, G. (2008) Goal attribution to inanimate agents by 6.5-month-old infants. *Cognition* 107:705–17. [aBML]

Csibra, G., Biro, S., Koos, O. & Gergely, G. (2003) One-year-old infants use teleological representations of actions productively. *Cognitive Science* 27:111–33. [aBML]

Csibra, G. & Gergely, G. (2009) Natural pedagogy. *Trends in Cognitive Sciences* 13 (4):148–53. [MHT]

Dalrymple, D. (2016) *Differentiable programming*. Available at: https://www.edge.org/response-detail/26794. [aBML]

Davies, J. (2016) Program good ethics into artificial intelligence. *Nature* 538(7625). Available at: http://www.nature.com/news/program-good-ethics-into-artificial-intelligence-1.20821. [KBC]

Davis, E. & Marcus, G. (2014) The scope and limits of simulation in cognition. *arXiv preprint 1506.04956*. Available at: arXiv: http://arxiv.org/abs/1506.04956. [ED]

Davis, E. & Marcus, G. (2015) Commonsense reasoning and commonsense knowledge in artificial Intelligence. *Communications of the ACM* 58(9):92–103. [aBML]

Davis, E. & Marcus, G. (2016) The scope and limits of simulation in automated reasoning. *Artificial Intelligence* 233:60–72. [ED]

Davoodi, T., Corriveau, K. H. & Harris, P. L. (2016) Distinguishing between realistic and fantastical figures in Iran. *Developmental Psychology* 52(2):221. [JMC, rBML]

Daw, N. D., Niv, Y. & Dayan, P. (2005) Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience* 8(12):1704–11. doi:10.1038/nn1560. [aBML, RPK]

Day, S. B. & Gentner, D. (2007) Nonintentional analogical inference in text comprehension. *Memory & Cognition* 35:39–49. [KDF]

Dayan, P., Hinton, G. E., Neal, R. M. & Zemel, R. S. (1995) The Helmholtz machine. *Neural Computation* 7(5):889–904. [aBML]

Deacon, T. (2012) *Incomplete nature: How mind emerged from matter*. W.W. Norton. [DCD]

Deacon, T. W. (1998) *The symbolic species: The co-evolution of language and the brain*. WW Norton. [aBML]

Dehghani, M., Tomai, E., Forbus, K. & Klenk, M. (2008) An integrated reasoning approach to moral decision-making. In: *Proceedings of the 23rd AAAI National Conference on Artificial Intelligence, vol. 3*, pp. 1280–86. AAAI Press. [KDF]

DeJong, G. & Mooney, R. (1986) Explanation-based learning: An alternative view. *Machine Learning* 1(2):145–76. [LRC]

Denil, M., Agrawal, P., Kulkarni, T. D., Erez, T., Battaglia, P. & de Freitas, N. (2016). Learning to perform physics experiments via deep reinforcement learning. *arXiv preprint:1611.01843*. Available at: https://arxiv.org/abs/1611.01843. [MB]

Dennett, D. C. (1987) *The intentional stance*. MIT Press. [JMC]

Dennett, D. C. (2013) Aching voids and making voids [Review of the book *Incomplete nature: How mind emerged from matter* by T. Deacon]. *The Quarterly Review of Biology* 88(4):321–24. [DCD]

Dennett, D. C. (2017) *From bacteria to Bach and back: The evolution of minds*. W.W. Norton. [DCD]

Denton, E., Chintala, S., Szlam, A. & Fergus, R. (2015) Deep generative image models using a Laplacian pyramid of adversarial networks. Presented at the 2015 Neural Information Processing Systems conference, Montreal, QC, Canada, In: *Advances in neural information processing systems 28 (NIPS 2015)*, ed. C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama & R. Garnett [poster]. Neural Information Processing Systems Foundation. [aBML]

Di, G. Q. & Wu, S. X. (2015) Emotion recognition from sound stimuli based on back-projection neural networks and electroencephalograms. *Journal of the Acoustics Society of America* 138(2):994–1002. [KBC]

DiCarlo, J. J., Zoccolan, D. & Rust, N. C. (2012) How does the brain solve visual object recognition? *Neuron* 73(3):415–34. [NK]

Dick, P. K. (1968) *Do androids dream of electric sheep?* Del Ray-Ballantine. [DEM]

Dietvorst, B. J., Simmons, J. P. & Massey, C. (2015) Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General* 144(1):114–26. [DCD]

Dietvorst, B. J., Simmons, J. P. & Massey, C. (2016) Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2616787. [DCD]

Diuk, C., Cohen, A. & Littman, M. L. (2008) An object-oriented representation for efficient reinforcement learning. In: *Proceedings of the 25th International Conference on Machine Learning (ICML'08)*, Helsinki, Finland, pp. 240–47. ACM. [aBML]

DiYanni, C. J., Corriveau, K. H., Kurkul, K., Nasrini, J. & Nini, D. (2015) The role of consensus and culture in children's imitation of questionable actions. *Journal of Experimental Child Psychology* 137:99–110. [JMC]

Doeller, C. F., Barry, C. & Burgess, N. (2010) Evidence for grid cells in a human memory network. *Nature* 463(7281):657–61. doi:10.1038/nature08704. [ESS]

Doeller, C. F. & Burgess, N. (2008) Distinct error-correcting and incidental learning of location relative to landmarks and boundaries. *Proceedings of the National Academy of Sciences of the United States of America* 105(15):5909–14. [ESS]

Doeller, C. F., King, J. A. & Burgess, N. (2008) Parallel striatal and hippocampal systems for landmarks and boundaries in spatial memory. *Proceedings of the National Academy of Sciences of the United States of America* 105(15):5915–20. doi:10.1073/pnas.0801489105. [ESS]

Dolan, R. J. & Dayan, P. (2013) Goals and habits in the brain. *Neuron* 80:312–25. [aBML]

Don, H. J., Goldwater, M. B., Otto, A. R. & Livesey, E. J. (2016) Rule abstraction, model-based choice, and cognitive reflection. *Psychonomic Bulletin & Review* 23(5):1615–23. [EJL]

Donahue, J., Hendricks, L. A., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K. & Darrell, T. (2015) Long-term recurrent convolutional networks for visual recognition and description. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, June 7-12, 2015*, pp. 2625–34. IEEE. [SW]

Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E. & Darrell, T. (2014) DeCAF: A deep convolutional activation feature for generic visual recognition. Presented at the International Conference on Machine Learning, Beijing, China, June 22–24, 2014. *Proceedings of Machine Learning Research* 32(1):647–55. [aBML]

Dörner, D. (2001) *Bauplan für eine Seele [Blueprint for a soul]*. Rowolt. [CDG]

Dörner, D. & Güss, C. D. (2013) PSI: A computational architecture of cognition, motivation, and emotion. *Review of General Psycholog* 17:297–317. doi:10.1037/a0032947. [CDG]

Doshi-Velez, F. & Kim, B. (2017) A roadmap for a rigorous science of interpretability. *arXiv preprint 1702.08608*. Available at: https://arxiv.org/abs/1702.08608. [rBML]

Doya, K. (1999) What are the computations of the cerebellum, the basal ganglia and the cerebral cortex? *Neural Networks* 12(7–8):961–74. [GB]

Dreyfus, H. & Dreyfus, S. (1986) *Mind over machine*. Macmillan. [BJM]

Duan, Y., Schulman, J., Chen, X., Bartlett, P. L., Sutskever, I. & Abbeel, P. (2016) RL^2: Fast reinforcement learning via slow reinforcement learning. *arXiv preprint 1611.02779*. Available at: https://arxiv.org/pdf/1703.07326.pdf. [MB]

Dunbar, K. (1995) How scientists really reason: Scientific reasoning in real-world laboratories. In: *The nature of insight*, ed. R. J. Sternberg & J. E. Davidson, pp. 365–95. MIT Press. [KDF]

Economides, M., Kurth-Nelson, Z., Lübbert, A., Guitart-Masip, M. & Dolan, R. J. (2015) Model-based reasoning in humans becomes automatic with training. *PLoS Computation Biology* 11:e1004463. [aBML]

Edelman, S. (2015) The minority report: Some common assumptions to reconsider in the modelling of the brain and behaviour. *Journal of Experimental & Theoretical Artificial Intelligence* 28(4):751–76. [aBML]

Eden, M. (1962) Handwriting and pattern recognition. *IRE Transactions on Information Theory* 8:160–66. [aBML]

Eickenberg, M., Gramfort, A., Varoquaux, G. & Thirion, B. (2016) Seeing it all: Convolutional network layers map the function of the human visual system. *NeuroImage* 2017;152:184–94. [NK]

Eliasmith, C., Stewart, T. C., Choo, X., Bekolay, T., DeWolf, T., Tang & Y. Rasmussen, D. (2012) A large-scale model of the functioning brain. *Science* 338(6111):1202–05. [aBML]

Eliasmith, C. & Trujillo, O. (2014) The use and abuse of large-scale brain models. *Current Opinion in Neurobiology* 25:1–6. [NK]

Elman, J. L. (1993) Learning and development in neural networks: The importance of starting small. *Cognition* 48(1):71–99. [SW]

Elman, J. L. (2005) Connectionist models of cognitive development: Where next? *Trends in Cognitive Sciences* 9(3):111–17. [aBML]

Elman, J. L., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Parisi, D. & Plunkett, K. (1996) *Rethinking innateness*. MIT Press. [aBML]

Eslami, S. M., Heess, N., Weber, T., Tassa, Y., Kavukcuoglu, K. & Hinton, G. E. (2016) Attend, infer, repeat: Fast scene understanding with generative models. Presented at the 2016 Neural Information Processing Systems conference, Barcelona, Spain, December 5–10, 2016. In: *Advances in Neural Information Processing Systems 29 (NIPS 2016)*, ed. D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon & R. Garnett, pp. 3225–33. Neural Information Processing Systems Foundation. [arBML, MB]

Eslami, S. M. A., Tarlow, D., Kohli, P. & Winn, J. (2014) Just-in-time learning for fast and flexible inference. Presented at the 2014 Neural Information Processing Systems conference, Montreal, QC, Canada, December 8–13, 2014. In: *Advances in neural information processing systems 27 (NIPS 2014)*, ed. Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence & K. Q. Weinberger, pp. 1736–44. Neural Information Processing Systems Foundation. [aBML]

Fasolo, A. (2011) *The theory of evolution and its impact*. Springer. [DG]

Feigenbaum, E. & Feldman, J., eds. (1995) *Computers and thought*. AAAI Press. [RJS]

Flash, T., Hochner, B. (2005) Motor primitives in vertebrates and invertebrates. *Current Opinion in Neurobiology* 15(6):660–66. [P-YO]

Fodor, J. A. (1975) *The language of thought*. Harvard University Press. [aBML]

Fodor, J. A. (1981) *Representations: Philosophical essays on the foundations of cognitive science*. MIT Press. [LRC]

Fodor, J. A. & Pylyshyn, Z. W. (1988) Connectionism and cognitive architecture: A critical analysis. *Cognition* 28(1–2):3–71. [aBML, RPK, SSH]

Fogel, D. B. & Fogel, L. J. (1995) Evolution and computational intelligence. *IEEE Transactions on Neural Networks* 4:1938–41. [KBC]

Forbus, K. (2011) Qualitative modeling. *Wiley Interdisciplinary Reviews: Cognitive Science* 2(4):374–91. [KDF]

Forbus, K., Ferguson, R., Lovett, A. & Gentner, D. (2017) Extending SME to handle large-scale cognitive modeling. *Cognitive Science* 41(5):1152–201. doi:10.1111/cogs.12377. [KDF]

Forbus, K. & Gentner, D. 1997. Qualitative mental models: Simulations or memories? Presented at the Eleventh International Workshop on Qualitative Reasoning, Cortona, Italy, June 3–6, 1997. [KDF]

Forestier, S. & Oudeyer, P.-Y. (2016) Curiosity-driven development of tool use precursors: A computational model. In: *Proceedings of the 38th Annual Conference of the Cognitive Science Society, Philadelphia, PA*, ed. A. Papafragou, D. Grodner, D. Mirman & J. C. Trueswell, pp. 1859–1864. Cognitive Science Society. [P-YO]

Fornito, A., Zalesky, A. & Bullmore, E. (2016) *Fundamentals of brain network analysis*. Academic Press. [DG]

Fox, J., Cooper, R. P. & Glasspool, D. W. (2013) A canonical theory of dynamic decision-making. *Frontiers in Psychology* 4(150):1–19. doi: 10.3389/fpsyg.2013.00150. [RPK]

Frank, M. C. & Goodman, N. D. (2014) Inferring word meanings by assuming that speakers are informative. *Cognitive Psychology* 75:80–96. [MHT]

Frank, M. C., Goodman, N. D. & Tenenbaum, J. B. (2009) Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science* 20:578–85. [aBML]

Franklin, S. (2007) A foundational architecture for artificial general intelligence. In: *Advances in artificial general intelligence: Concepts, architectures and algorithms: Proceedings of the AGI Workshop 2006*, ed. P. Want & B. Goertzel, pp. 36–54. IOS Press. [GB]

Freyd, J. (1983) Representing the dynamics of a static form. *Memory and Cognition* 11(4):342–46. [aBML]

Freyd, J. (1987) Dynamic mental representations. *Psychological Review* 94(4):427–38. [aBML]

Friedman, S. E. and Forbus, K. D. (2010) An integrated systems approach to explanation-based conceptual change. In: *Proceedings of the 24th AAAI Conference on Artificial Intelligence, Atlanta, GA, July 11–15, 2010*. AAAI Press. [KDF]

Fukushima, K. (1980) Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics* 36:193–202. [aBML]

Fung, P. (2015) Robots with heart. *Scientific American* 313(5):60–63. [KBC]

Funke, J. (2010) Complex problem solving: A case for complex cognition? *Cognitive Processing* 11:133–42. [CDG]

Gallese, V. & Lakoff, G. (2005) The brain's concepts: The role of the sensory-motor system in conceptual knowledge. *Cognitive Neuropsychology* 22(3–4):455–79. [SW]

Gallistel, C. & Matzel, L. D. (2013) The neuroscience of learning: beyond the Hebbian synapse. *Annual Review of Psychology* 64:169–200. [aBML]

Gaussier, P., Moga, S., Quoy, M. & Banquet, J. P. (1998). From perception-action loops to imitation processes: A bottom-up approach of learning by imitation. *Applied Artificial Intelligence* 12(7–8):701–27. [LM]

Gazzaniga, M. (2004) *Cognitive neuroscience*. MIT Press. [MBu]

Gelly, S. & Silver, D. (2008) Achieving master level play in 9 × 9 computer Go. In: *Proceedings of the Twenty-third AAAI Conference on Artificial Intelligence, Chicago, Illinois, July 13–17, 2008*, pp. 1537–40. AAAI Press. [aBML]

Gelly, S. & Silver, D. (2011) Monte-Carlo tree search and rapid action value estimation in computer go. *Artificial Intelligence* 175(11):1856–75. [aBML]

Gelman, A., Carlin, J. B., Stern, H. S. & Rubin, D. B. (2004) *Bayesian data analysis*. Chapman & Hall/CRC. [aBML]

Gelman, A., Lee, D. & Guo, J. (2015) Stan: A probabilistic programming language for Bayesian inference and optimization. *Journal of Educational and Behavioral Statistics* 40:530–43. [aBML]

Gelman, S. A. (2009) Learning from others: Children's construction of concepts. *Annual Review of Psychology* 60:115–40. [MHT]

Geman, S., Bienenstock, E. & Doursat, R. (1992) Neural networks and the bias/variance dilemma. *Neural Computation* 4:1–58. [aBML]

Gentner, D. (1983) Structure-mapping: A theoretical framework for analogy. *Cognitive Science* 7:155–70. (Reprinted in A. Collins & E. E. Smith, eds. *Readings in cognitive science: A perspective from psychology and artificial intelligence*. Kaufmann.) [KDF]

Gentner, D. (2010) Bootstrapping the mind: Analogical processes and symbol systems. *Cognitive Science* 34(5):752–75. [KDF]

Gentner, D., Loewenstein, J., Thompson, L. & Forbus, K. D. (2009) Reviving inert knowledge: Analogical abstraction supports relational retrieval of past events. *Cognitive Science* 33(8):1343–82. [EJL]

George, D. & Hawkins, J. (2009) Towards a mathematical theory of cortical microcircuits. *PLoS Computational Biology* 5(10):e1000532. Available at: http://doi.org/10.1371/journal.pcbi.1000532. [DGe]

Gershman, S. J. & Goodman, N. D. (2014) Amortized inference in probabilistic reasoning. In: *Proceedings of the 36th Annual Conference of the Cognitive Science Society, Quebec City, QC, Canada, July 23–26, 2014*, pp. 517–522. Cognitive Science Society. [aBML]

Gershman, S. J., Horvitz, E. J. & Tenenbaum, J. B. (2015) Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science* 34:273–78. [aBML]

Gershman, S. J., Markman, A. B. & Otto, A. R. (2014) Retrospective revaluation in sequential decision making: A tale of two systems. *Journal of Experimental Psychology: General* 143:182–94. [aBML]

Gershman, S. J., Vul, E. & Tenenbaum, J. B. (2012) Multistability and perceptual inference. *Neural Computation* 24:1–24. [aBML]

Gerstenberg, T., Goodman, N. D., Lagnado, D. A. & Tenenbaum, J. B. (2015) How, whether, why: Causal judgments as counterfactual contrasts. In: *Proceedings of the 37th Annual Conference of the Cognitive Science Society, Pasadena, CA, July 22–25, 2015*, ed. D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings & P. P. Maglio, pp. 782–787. Cognitive Science Society. [aBML, ED]

Ghahramani, Z. (2015) Probabilistic machine learning and artificial intelligence. *Nature* 521:452–59. [aBML]

Giambene, G. (2005) *Queuing theory and telecommunications networks and applications*. Springer Science + Business Media. [DG]

Gibson, J. J. (1979) *The ecological approach to visual perception*. Houghton Mifflin. [DCD]

Gick, M. L. & Holyoak, K. J. (1980) Analogical problem solving. *Cognitive Psychology* 12(3):306–55. [EJL]

Gigerenzer, G. (2001) The adaptive toolbox. In: *Bounded rationality: The adaptive toolbox*, ed. G. Gigerenzer & R. Selten, pp. 37–50. MIT Press. [PMP]

Gigerenzer, G. & Gaissmaier, W. (2011) Heuristic decision making. *Annual Review of Psychology* 62:451–82. doi:10.1146/annurev-psych-120709-145346. [PMP]

Goldberg, A. E. (1995) *Constructions: A construction grammar approach to argument structure*. University of Chicago Press. [NC]

Gombrich, E. (1960) *Art and illusion*. Pantheon Books. [NC]

Goodfellow, I., Schlens, J. & Szegedy, C. (2015) Explaining and harnessing adversarial examples. Presented at International Conference on Learning Representations (ICLR), San Diego, CA, May 7–9, 2015. *arXiv preprint 1412.6572*. Available at: https://arxiv.org/abs/1412.6572. [KDF]

Goodman, N. D. & Frank, M. C. (2016) Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences* 20(11):818–29. [MHT]

Goodman, N. D., Mansinghka, V. K., Roy, D. M., Bonawitz, K. & Tenenbaum, J. B. (2008) Church: A language for generative models. In: *Proceedings of the Twenty-Fourth Annual Conference on Uncertainty in Artificial Intelligence, Helsinki, Finland, July 9–12, 2008*, pp. 220–29. AUAI Press. [aBML]

Goodman, N. D., Tenenbaum, J. B., Feldman, J. & Griffiths, T. L. (2008) A rational analysis of rule-based concept learning. *Cognitive Science* 32(1):108–54. [rBML]

Goodman, N. D., Tenenbaum, J. B. & Gerstenberg, T. (2015). Concepts in a probabilistic language of thought. In: *The conceptual mind: New directions in the study of concepts*, ed. E. Margolis & S. Laurence, pp. 623–54. MIT Press. [rBML]

Goodman, N. D., Ullman, T. D. & Tenenbaum, J. B. (2011) Learning a theory of causality. *Psychological Review* 118(1):110–19. [rBML]

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T. & Danks, D. (2004) A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review* 111(1):3–32. [arBML]

Gopnik, A. & Meltzoff, A. N. (1999) *Words, thoughts, and theories*. MIT Press. [aBML]

Gottlieb, J., Oudeyer, P-Y., Lopes, M. & Baranes, A. (2013) Information seeking, curiosity and attention:Computational and neural mechanisms. *Trends in Cognitive Science* 17(11):585–96. [P-YO]

Graham, D. J. (2014) Routing in the brain. *Frontiers in Computational Neuroscience* 8:44. [DG]

Graham, D. J. and Rockmore, D. N. (2011) The packet switching brain. *Journal of Cognitive Neuroscience* 23(2):267–76. [DG]

Granger, R. (2006) Engines of the brain: The computational instruction set of human cognition. *AI Magazine* 27(2):15. [DG]

Graves, A. (2014) Generating sequences with recurrent neural networks. *arXiv preprint 1308.0850*. Available at: http://arxiv.org/abs/1308.0850. [aBML]

Graves, A., Mohamed, A.-R. & Hinton, G. (2013) Speech recognition with deep recurrent neural networks. In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Vancouver, BC, Canada, May 26–31, 2013*, pp. 6645–49. IEEE. [aBML]

Graves, A., Wayne, G. & Danihelka, I. (2014) Neural Turing machines. *arXiv preprint 1410.5401v1*. Available at: http://arxiv.org/abs/1410.5401v1. [arBML]

Graves, A., Wayne, G., Reynolds, M., Harley, T., Danihelka, I., Grabska-Barwińska, A., Colmenarejo, S. G., Grefenstette, E., Ramalho, T., Agapiou, J., Badia, A. P., Hermann, K. M., Zwols, Y., Ostrovski, G., Cain, A., King, H., Summerfield, C., Blunsom, P., Kayukcuoglu, K. & Hassabis, D. (2016) Hybrid computing using a neural network with dynamic external memory. *Nature* 538(7626):471–76. [arBML, MB]

Gray, H. M., Gray, K. & Wegner, D. M. (2007) Dimensions of mind perception. *Science* 315(5812):619. [rBML, SW]

Gray, K. & Wegner, D. M. (2012) Feeling robots and human zombies: Mind perception and the uncanny valley. *Cognition* 125(1):125–30. [rBML]

Graybiel, A. M. (2005) The basal ganglia: learning new tricks and loving it. *Current Opinion in Neurobiology* 15(6):638–44. [GB]

Grefenstette, E., Hermann, K. M., Suleyman, M. & Blunsom, P. (2015). Learning to transduce with unbounded memory. Presented at the 2015 Neural Information Processing Systems conference. In: *Advances in Neural Information Processing Systems 28*, ed. C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama & R. Garnett. Neural Information Processing Systems Foundation. [arBML]

Gregor, K., Besse, F., Rezende, D. J., Danihelka, I. & Wierstra, D. (2016) Towards conceptual compression. Presented at the 2016 Neural Information Processing Systems conference, Barcelona, Spain, December 5–10, 2016. In: *Advances in Neural Information Processing Systems 29 (NIPS 2016)*, ed. D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon & R. Garnett [poster]. Neural Information Processing Systems Foundation. [aBML]

Gregor, K., Danihelka, I., Graves, A., Rezende, D. J. & Wierstra, D. (2015) DRAW: A recurrent neural network for image generation. Presented at the 32nd Annual International Conference on Machine Learning (ICML'15), Lille, France, July 7–9, 2015. *Proceedings of Machine Learning Research* 37:1462–71. [aBML]

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A. & Tenenbaum, J. B. (2010) Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in Cognitive Sciences* 14(8):357–64. [arBML]

Griffiths, T. L. & Tenenbaum, J. B. (2005) Structure and strength in causal induction. *Cognitive Psychology* 51(4):334–84. [rBML]

Griffiths, T. L. & Tenenbaum, J. B. (2009) Theory-based causal induction. *Psychological Review* 116(4):661–716. [rBML]

Griffiths, T. L., Vul, E. & Sanborn, A. N. (2012) Bridging levels of analysis for probabilistic models of cognition. *Current Directions in Psychological Science* 21:263–68. [aBML]

Grossberg, S. (1976) Adaptive pattern classification and universal recoding: I. Parallel development and coding of neural feature detectors. *Biological Cybernetics* 23:121–34. [aBML]

Grosse, R., Salakhutdinov, R., Freeman, W. T. & Tenenbaum, J. B. (2012) Exploiting compositionality to explore a large space of model structures. In: Proceedings of the Twenty-Eighth Annual Conference on *Uncertainty in Artificial Intelligence , Catalina Island , CA*, ed. N. de Freitas & K. Murphy , pp. 306–15. AUAI Press. [aBML]

Güçlü, U. & van Gerven, M. A. J. (2015) Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience* 35(27):10005–14. [NK]

Guergiuev, J., Lillicrap, T. P. & Richards, B. A. (2016) Toward deep learning with segregated dendrites. *arXiv preprint 1610.00161*. Available at: http://arxiv.org/pdf/1610.00161.pdf. [AHM]

Gülçehre, Ç. & Bengio, Y. (2016) Knowledge matters: Importance of prior information for optimization. *Journal of Machine Learning Research* 17(8):1–32. [SSH]

Guo, X., Singh, S., Lee, H., Lewis, R. L. & Wang, X. (2014) Deep learning for real-time Atari game play using offline Monte-Carlo tree search planning. In: *Advances in neural information processing systems 27 (NIPS 2014)*, ed. Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence & K. Q. Weinberger [poster]. Neural Information Processing Systems Foundation. [aBML]

Güss, C. D., Tuason, M. T. & Gerhard, C. (2010) Cross-national comparisons of complex problem-solving strategies in two microworlds. *Cognitive Science* 34:489–520. [CDG]

Gweon, H., Tenenbaum, J. B. & Schulz, L. E. (2010) Infants consider both the sample and the sampling process in inductive generalization. *Proceedings of the National Academy of Sciences of the United States of America* 107:9066–71. [aBML]

Hafenbrädl, S., Waeger, D., Marewski, J. N. & Gigerenzer, G. (2016) Applied decision making with fast-and-frugal heuristics. *Journal of Applied Research in Memory and Cognition* 5(2):215–31. doi:10.1016/j.jarmac.2016.04.011. [PMP]

Hall, E. T. (1966) *The hidden dimension*. Doubleday. [SW]

Halle, M. & Stevens, K. (1962) Speech recognition: A model and a program for research. *IRE Transactions on Information Theory* 8(2):155–59. [aBML]

Hamlin, K. J. (2013) Moral judgment and action in preverbal infants and toddlers: Evidence for an innate moral core. *Current Directions in Psychological Science* 22:186–93. [aBML]

Hamlin, K. J., Ullman, T., Tenenbaum, J., Goodman, N. D. & Baker, C. (2013) The mentalistic basis of core social cognition: Experiments in preverbal infants and a computational model. *Developmental Science* 16:209–26. [aBML]

Hamlin, K. J., Wynn, K. & Bloom, P. (2007) Social evaluation by preverbal infants. *Nature* 450:57–60. [aBML]

Hamlin, K. J., Wynn, K. & Bloom, P. (2010) Three-month-olds show a negativity bias in their social evaluations. *Developmental Science* 13:923–29. [aBML]

Hamper, B. (2008) *Rivethead: Tales from the assembly line*. Grand Central. [DEM]

Hamrick, J. B., Ballard, A. J., Pascanu, R., Vinyals, O., Heess, N. & Battaglia, P. W. (2017) Metacontrol for adaptive imagination-based optimization. In: *Proceedings of the 5th International Conference on Learning Representations (ICLR)*. [MB]

Han, M. J., Lin, C. H. & Song, K. T. (2013) Robotic emotional expression generation based on mood transition and personality model. *IEEE Transactions on Cybernetics* 43(4):1290–303. [KBC]

Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., Prenger, R., Satheesh, S., Shubho, S., Coates, A. & Ng, A. Y. (2014) Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint 1412.5567*. Available at: https://arxiv.org/abs/1412.5567. [aBML]

Hanson, S. J., (1995) Some comments and variations on back-propagation. In: *The handbook of back-propagation*, ed. Y. Chauvin & D. Rummelhart, pp. 292–323. Erlbaum. [LRC]

Hanson, S. J. (2002) On the emergence of rules in neural networks. *Neural Computation* 14(9):2245–68. [LRC]

Hanson, S. J. & Burr, D. J., (1990) What connectionist models learn: Toward a theory of representation in connectionist networks. *Behavioral and Brain Sciences* 13:471–518. [LRC]

Hanson, S. J., Caglar, L. R. & Hanson, C. (under review) The deep history of deep learning. [LRC]

Harkness, S., Blom, M., Oliva, A., Moscardino, U., Zylicz, P. O., Bermudez, M. R. & Super, C. M. (2007) Teachers' ethnotheories of the 'ideal student' in five western cultures. *Comparative Education* 43(1):113–35. [JMC]

Harlow, H. F. (1949) The formation of learning sets. *Psychological Review* 56(1):51–65. [aBML]

Harlow, H. F. (1950) Learning and satiation of response in intrinsically motivated complex puzzle performance by monkeys. *Journal of Comparative and Physiological Psychology* 43:289–94. [aBML]

Harris, P. L. (2012) *Trusting what you're told: How children learn from others*. Belknap Press of Harvard University Press. [JMC]

Haslam, N. (2006) Dehumanization: An integrative review. *Personality and Social Psychology Review* 10(3):252–64. [rBML]

Hasnain, S.K., Mostafaoui, G. & Gaussier, P. (2012). A synchrony-based perspective for partner selection and attentional mechanism in human-robot interaction. *Paladyn, Journal of Behavioral Robotics* 3(3):156–71. [LM]

Hasnain, S. K., Mostafaoui, G., Salesse, R., Marin, L. & Gaussier, P. (2013) Intuitive human robot interaction based on unintentional synchrony: A psycho-experimental study. In: *Proceedings of the IEEE 3rd Joint Conference on Development*

and Learning and on Epigenetic Robotics, Osaka, Japan, August 2013, pp. 1–7. Hal Archives-ouvertes. [LM]
Hauser, M. D., Chomsky, N. & Fitch, W. T. (2002) The faculty of language: what is it, who has it, and how did it evolve? *Science* 298:1569–79. [aBML]
Hayes, P. J. (1974) Some problems and non-problems in representation theory. In: *Proceedings of the 1st summer conference on artificial intelligence and simulation of behaviour*, pp. 63–79. IOS Press. [LRC]
Hayes-Roth, B. & Hayes-Roth, F. (1979) A cognitive model of planning. *Cognitive Science* 3:275–310. [aBML]
He, K., Zhang, X., Ren, S. & Sun, J. (2016) Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Las Vegas, NV, June 27–30, 2016. pp. 770–78. IEEE. [aBML]
Hebb, D. O. (1949) *The organization of behavior*. Wiley. [aBML]
Heess, N., Tarlow, D. & Winn, J. (2013) Learning to pass expectation propagation messages. Presented at the 25th International Conference on Neural Information Processing Systems, Lake Tahoe, NV, December 3–6, 2012. In: *Advances in Neural Information Processing Systems 25 (NIPS 2012)*, ed. F. Pereira, C. J. C. Burges, L. Bottou & K. Q. Weinberger, pp. 3219–27. Neural Information Processing Systems Foundation. [aBML]
Heinrich, S. (2016) *Natural language acquisition in recurrent neural architectures*. Ph.D. thesis, Universität Hamburg, DE. [SW]
Henrich, J. (2015) *The secret of our success: How culture is driving human evolution, domesticating our species, and making us smarter*. Princeton University Press. [JMC]
Henrich, J., Heine, S. J. & Norenzayan, A. (2010) The weirdest people in the world? *Behavioral and Brain Sciences* 33(2–3):61–83. [JMC]
Herrmann, E., Call, J., Hernandez-Lloreda, M. V., Hare, B. & Tomasello, M. (2007) Humans have evolved specialized skills of social cognition: The cultural intelligence hypothesis. *Science* 317(5843):1360–66. [DCD]
Herrmann, E., Hernandez-Lloreda, M. V., Call, J., Hare, B. & Tomasello, M. (2010) The structure of individual differences in the cognitive abilities of children and chimpanzees. *Psychological Science* 21(1):102–10. [DCD]
Hertwig, R. & Herzog, S. M. (2009) Fast and frugal heuristics: Tools of social rationality. *Social Cognition* 27(5):661–98. doi:10.1521/soco.2009.27.5.661. [PMP]
Hespos, S. J. & Baillargeon, R. (2008) Young infants' actions reveal their developing knowledge of support variables: Converging evidence for violation-of-expectation findings. *Cognition* 107:304–16. [aBML]
Hespos, S. J., Ferry, A. L. & Rips, L. J. (2009) Five-month-old infants have different expectations for solids and liquids. *Psychological Science* 20(5):603–11. [aBML]
Hinrichs, T. & Forbus, K. (2011) Transfer learning through analogy in games. *AI Magazine* 32(1):72–83. [KDF]
Hinton, G. E. (2002) Training products of experts by minimizing contrastive divergence. *Neural Computation* 14(8):1771–800. [aBML]
Hinton, G. E., Dayan, P., Frey, B. J. & Neal, R. M. (1995) The "wake-sleep" algorithm for unsupervised neural networks. *Science* 268(5214):1158–61. [aBML]
Hinton, G. E., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. & Kingsbury, B. (2012) Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal Processing Magazine* 29:82–97. [aBML]
Hinton, G. E., Osindero, S. & Teh, Y. W. (2006) A fast learning algorithm for deep belief nets. *Neural Computation 18*:1527–54. [aBML]
Hiolle, A., Lewis, M. & Cañamero, L. (2014) Arousal regulation and affective adaptation to human responsiveness by a robot that explores and learns a novel environment. *Frontiers in Neurorobotics* 8:17. [KBC]
Ho, Y-C. & Pepyne, D. L. (2002) Simple explanation of the no-free-lunch theorem and its implications. *Journal of Optimization Theory and Applications* 115:549–70. [AHM]
Hochreiter, S. A., Younger, S. & Conwell, P. R. (2001) Learning to learn using gradient descent. In: *International Conference on Artificial Neural Network—ICANN 2001*, ed. G. Dorffner, H. Bischoff & K. Hornik, pp. 87–94. Springer. [MB]
Hoffman, D. D. (2000) *Visual intelligence: How we create what we see*. W. W. Norton. [NC]
Hoffman, D. D. & Richards, W. A. (1984) Parts of recognition. *Cognition* 18:65–96. [aBML]
Hoffman, M., Yoeli, E. & Nowak, M. A. (2015) Cooperate without looking: Why we care what people think and not just what they do. *Proceedings of the National Academy of Sciences of the United States of America* 112(6):1727–32. doi:10.1073/pnas.1417904112. [PMP]
Hofstadter, D. R. (1985) *Metamagical themas: Questing for the essence of mind and pattern*. Basic Books. [aBML]
Hofstadter, D. R. (2001) Epilogue: Analogy as the core of cognition. In: *The analogical mind: perspectives from cognitive science*, ed. D. Gentner, K. J. Holyoak & B. N. Kokinov, pp. 499–538. MIT Press. [NC]
Horgan, T. & J. Tienson, (1996) *Connectionism and the philosophy of psychology*. MIT Press. [LRC]
Horst, J. S. & Samuelson, L. K. (2008) Fast mapping but poor retention by 24-month-old infants. *Infancy* 13(2):128–57. [aBML]
Houk, J. C., Adams, J. L. & Barto, A. G. (1995) A model of how the basal ganglia generate and use neural signals that predict reinforcement. In: *Models of information processing in the basal ganglia*, ed. J. C. Houk, J. L. Davids & D. G. Beiser, pp. 249–70. MIT Press. [GB]
Huang, Y. & Rao, R. P. (2014) Neurons as Monte Carlo samplers: Bayesian? inference and learning in spiking networks Presented at the 2014 Neural Information Processing Systems conference, Montreal, QC, Canada. In: *Advances in neural information processing systems 27 (NIPS 2014)*, ed. Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence & K. Q. Weinberger, pp. 1943–51. Neural Information Processing Systems Foundation. [aBML]
Hubel, D. H. & Wiesel, T. N. (1959) Receptive fields of single neurons in the cat's striate cortex. *Journal of Physiology* 124:574–91. [ESS]
Hummel, J. E. & Biederman, I. (1992) Dynamic binding in a neural network for shape recognition. *Psychological Review* 99(3):480–517. [aBML]
Hurley, M., Dennett, D. C. & Adams, R., (2011) *Inside jokes: Using humor to reverse-engineer the mind*. MIT Press. [DCD]
Hutson, M. (2017) In bots we distrust. *Boston Globe*, p. K4. [DCD]
Indiveri, G. & Liu, S.-C. (2015) Memory and information processing in neuromorphic systems. *Proceedings of the IEEE* 103(8):1379–97. [KBC]
Indurkhya, B. & Misztal-Radecka, J. (2016) Incorporating human dimension in autonomous decision-making on moral and ethical issues. In: *Proceedings of the AAAI Spring Symposium: Ethical and Moral Considerations in Non-human Agents, Palo Alto, CA*, ed. B. Indurkhya & G. Stojanov. AAAI Press. [PMP]
Irvine, A. D. & Deutsch, H. (2016) Russell's paradox. In: *The Stanford encyclopedia of philosophy* (Winter 2016 Edition), ed. E. N. Zalta. Available at: https://plato.stanford.edu/archives/win2016/entries/russell-paradox. [NC]
Jackendoff, R. (2003) *Foundations of language*. Oxford University Press. [aBML]
Jaderberg, M., Mnih, V., Czarnecki, W. M., Schaul, T., Leibo, J. Z., Silver, D. & Kavukcuoglu, K. (2016) Reinforcement learning with unsupervised auxiliary tasks. Presented at the 5th International Conference on Learning Representations, Palais des Congrès Neptune, Toulon, France, April 24–26, 2017. *arXiv preprint 1611.05397*. Available at: https://arxiv.org/abs/1611.05397. [P-YO]
Jain, A., Tompson, J., Andriluka, M., Taylor, G. W. & Bregler, C. (2014). Learning human pose estimation features with convolutional networks. Presented at the International Conference on Learning Representations (ICLR), Banff, Canada, April 14–16, 2014. *arXiv preprint 1312.7302*. Available at: https://arxiv.org/abs/1312.7302. [rBML]
Jara-Ettinger, J., Gweon, H., Schulz, L. E. & Tenenbaum, J. B. (2016) The naïve utility calculus: Computational principles underlying commonsense psychology. *Trends in Cognitive Sciences* 20(8):589–604. doi:10.1016/j.tics.2016.05.011. [PMP]
Jara-Ettinger, J., Gweon, H., Tenenbaum, J. B. & Schulz, L. E. (2015) Children's understanding of the costs and rewards underlying rational action. *Cognition* 140:14–23. [aBML]
Jern, A. & Kemp, C. (2013) A probabilistic account of exemplar and category generation. *Cognitive Psychology* 66(1):85–125. [aBML]
Jern, A. & Kemp, C. (2015) A decision network account of reasoning about other peoples choices. *Cognition* 142:12–38. [aBML]
Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z. & Hughes, M. (2016) Google's multilingual neural machine translation system: Enabling zero-shot translation. *arXiv preprint 1611.04558*. Available at: https://arxiv.org/abs/1611.04558. [SSH]
Johnson, S. C., Slaughter, V. & Carey, S. (1998) Whose gaze will infants follow? The elicitation of gaze-following in 12-month-olds. *Developmental Science* 1:233–38. [aBML]
Jonge, M. de & Racine, R. J. (1985) The effects of repeated induction of long-term potentiation in the dentate gyrus. *Brain Research* 328:181–85. [aBML]
Juang, B. H. & Rabiner, L. R. (1990) Hidden Markov models for speech recognition. *Technometric* 33(3):251–72. [aBML]
Kahneman, D. (2011) *Thinking, fast and slow*. Macmillan. [MB]
Kahou, S. E., Pal, C., Bouthillier, X., Froumenty, P., Gülçehre, Ç., Memisevic, R., Vincent, P., Courville, A. & Bengio, Y. (2013) Combining modality specific deep neural networks for emotion recognition in video. In: *Proceedings of the 15th ACM International Conference on Multimodal Interaction, Koogee Beach, Sydney, Australia*, pp. 543–50. ACM. [rBML]
Kaipa, K. N., Bongard, J. C. & Meltzoff, A. N. (2010) Self discovery enables robot social cognition: Are you my teacher? *Neural Networks* 23(8–9):1113–24. [KBC]
Karpathy, A. & Fei-Fei, L. (2017) Deep visual-semantic alignments for generating image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39(4):664–76. [aBML]
Kawato, M., Kuroda, S. & Schweighofer, N. (2011) Cerebellar supervised learning revisited: biophysical modeling and degrees-of-freedom control. *Current Opinion in Neurobiology* 21(5):791–800. [GB]
Keller, N. & Katsikopoulos, K. V. (2016) On the role of psychological heuristics in operational research; and a demonstration in military stability operations.

European Journal of Operational Research 249(3):1063–73. doi:10.1016/j.ejor.2015.07.023. [PMP]

Kellman, P. J. & Spelke, E. S. (1983) Perception of partly occluded objects in infancy. *Cognitive Psychology* 15(4):483–524. [rBML]

Kemp, C. (2007) The acquisition of inductive constraints. Unpublished doctoral dissertation, Massachusetts Institute of Technology. [aBML]

Kemp, C., Perfors, A. & Tenenbaum, J. B. (2007) Learning overhypotheses with hierarchical Bayesian models. *Developmental Science* 10(3):307–21. [rBML]

Kemp, C. & Tenenbaum, J. B. (2008) The discovery of structural form. *Proceedings of the National Academy of Sciences of the United States of America* 105 (31):10687–92. [rBML]

Keramati, M., Dezfouli, A. & Piray, P. (2011) Speed/accuracy trade-off between the habitual and the goal-directed processes. *PLoS Computational Biology* 7: e1002055. [aBML]

Khaligh-Razavi, S. M. & Kriegeskorte, N. (2014) Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Computational Biology* 10(11):e1003915. [aBML, NK]

Kidd, C., Piantadosi, S. T. & Aslin, R. N. (2012) The Goldilocks effect: Human infants allocate attention to visual sequences that are neither too simple nor too complex. *PLoS One* 7(5):e36399. [P-YO]

Kiddon, C., Zettlemoyer, L. & Choi, Y. (2016). Globally coherent text generation with neural checklist models. In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, Texas, November 1–5, 2016,* pp. 329–39. Association for Computational Linguistics. [rBML]

Kilner, J. M., Friston, K. J. & Frith, C. D. (2007) Predictive coding: An account of the mirror neuron system. *Cognitive Processing* 8(3):159–66. [aBML]

Kingma, D. P., Rezende, D. J., Mohamed, S. & Welling, M. (2014) Semi-supervised learning with deep generative models. Presented at the 2014 Neural Information Processing Systems conference, Montreal, QC, Canada. In: *Advances in neural information processing systems 27 (NIPS 2014),* ed. Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence & K. Q. Weinberger [spotlight]. Neural Information Processing Systems Foundation. [aBML]

Kiraly, I., Csibra, G. & Gergely, G. (2013) Beyond rational imitation: Learning arbitrary means actions from communicative demonstrations. *Journal of Experimental Child Psychology* 116(2):471–86. [DCD]

Kline, M. A. (2015) How to learn about teaching: An evolutionary framework for the study of teaching behavior in humans and other animals. *Behavioral and Brain Sciences* 2015;38:e31. [JMC, MHT]

Koch, G., Zemel, R. S. & Salakhutdinov, R. (2015) Siamese neural networks for one-shot image recognition. Presented at the Deep Learning Workshop at the 2015 International Conference on Machine Learning, Lille, France. Available at: https://www.cs.cmu.edu/~rsalakhu/papers/oneshot1.pdf. [aBML]

Kodratoff, Y. & Michalski, R. S. (2014) *Machine: earning: An artificial intelligence approach, vol. 3*. Morgan Kaufmann. [aBML]

Kolodner, J. (1993) *Case-based reasoning*. Morgan Kaufmann. [NC]

Koza, J. R. (1992) *Genetic programming: On the programming of computers by means of natural selection, vol. 1.* MIT press. [aBML]

Kriegeskorte, N. (2015) Deep neural networks: A new framework for modeling biological vision and brain information processing. *Annual Review of Vision Science* 1:417–46. [aBML]

Kriegeskorte, N. & Diedrichsen, J. (2016) Inferring brain-computational mechanisms with models of activity measurements. *Philosophical Transactions of the Royal Society of London Series B: Biological Sciences* 371(1705):489–95. [NK]

Kriegeskorte, N., Mur, M. & Bandettini, P. (2008) Representational similarity analysis – Connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience* 2:4. doi: 10.3389/neuro.06.004.2008. [NK]

Krizhevsky, A., Sutskever, I. & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. Presented at the 25th International Conference on Neural Information Processing Systems, Lake Tahoe, NV, December 3–6, 2012. In: *Advances in Neural Information Processing Systems 25 (NIPS 2012),* ed. F. Pereira, C. J. C. Burges, L. Bottou & K. Q. Weinberger, pp. 1097–105. Neural Information Processing Systems Foundation. [arBML, MB, NK, SSH]

Kulkarni, T. D., Kohli, P., Tenenbaum, J. B. & Mansinghka, V. (2015a) Picture: A probabilistic programming language for scene perception. In: *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, June 7–12, 2015,* pp. 4390–99. IEEE. [aBML]

Kulkarni, T. D., Narasimhan, K. R., Saeedi, A. & Tenenbaum, J. B. (2016) Hierarchical deep reinforcement learning: Integrating temporal abstraction and intrinsic motivation. *arXiv preprint 1604.06057*. Available at: https://arxiv.org/abs/1604.06057. [aBML, P-YO]

Kulkarni, T. D., Whitney, W., Kohli, P. & Tenenbaum, J. B. (2015b) Deep convolutional inverse graphics network. *arXiv preprint 1503.03167.* Available at: https://arxiv.org/abs/1503.03167. [aBML]

Lake, B. M. (2014) *Towards more human-like concept learning in machines: Compositionality, causality, and learning-to-learn*. Unpublished doctoral dissertation, Massachusetts Institute of Technology. [aBML]

Lake, B. M., Lawrence, N. D. & Tenenbaum, J. B. (2016) The emergence of organizing structure in conceptual representation. *arXiv preprint 1611.09384*. Available at: http://arxiv.org/abs/1611.09384. [MB, rBML]

Lake, B. M., Lee, C.-Y., Glass, J. R. & Tenenbaum, J. B. (2014) One-shot learning of generative speech concepts. In: *Proceedings of the 36th Annual Conference of the Cognitive Science Society, Quebec City, QC, Canada, July 23–26, 2014,* pp. 803–08. Cognitive Science Society. [aBML]

Lake, B. M., Salakhutdinov, R. & Tenenbaum, J. B. (2012) Concept learning as motor program induction: A large-scale empirical study. In: *Proceedings of the 34th Annual Conference of the Cognitive Science Society, Sapporo, Japan, August 1–4, 2012*, pp. 659–64. Cognitive Science Society. [aBML]

Lake, B. M., Salakhutdinov, R. & Tenenbaum, J. B. (2015a) Human-level concept learning through probabilistic program induction. *Science* 350(6266):1332–38. [arBML, MB, ED, NK]

Lake, B. M., Zaremba, W., Fergus, R. & Gureckis, T. M. (2015b) Deep neural networks predict category typicality ratings for images. In: *Proceedings of the 37th Annual Meeting of the Cognitive Science Society, Pasadena, CA, July 22–25, 2015*. Cognitive Science Society. ISBN: 978-0-9911967-2-2. [aBML]

Lakoff, G. & Johnson, M. (2003) *Metaphors we live by*, 2nd ed. University of Chicago Press. [SW]

Lambert, A (2011) *The gates of hell: Sir John Franklin's tragic quest for the Northwest Passage*. Yale University Press. [MHT]

Landau, B., Smith, L. B. & Jones, S. S. (1988) The importance of shape in early lexical learning. *Cognitive Development* 3(3):299–321. [aBML]

Lande, T. S., ed. (1998) Neuromorphic systems engineering: Neural networks in silicon. *Kluwer International Series in Engineering and Computer Science, vol. 447*. Kluwer Academic. ISBN 978-0-7923-8158-7. [KBC]

Langley, P., Bradshaw, G., Simon, H. A. & Zytkow, J. M. (1987) *Scientific discovery: Computational explorations of the creative processes*. MIT Press. [aBML]

Laptev, I., Marszalek, M., Schmid, C. & Rozenfeld, B. (2008) Learning realistic human actions from movies. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, June 23–28, 2008 (CVPR 2008)*, pp. 1–8. IEEE. [SW]

Larson, H. J., Cooper, L. Z., Eskola, J., Katz, S. L. & Ratzan, S. (2011) Addressing the vaccine confidence gap. *The Lancet* 378(9790):526–35. [EJL]

Lázaro-Gredilla, M., Liu, Y., Phoenix, D. S. & George, D. (2016) Hierarchical compositional feature learning. *arXiv preprint 1611.02252*. Available at: http://arxiv.org/abs/1611.02252. [DGe]

LeCun, Y., Bengio, Y. & Hinton, G. (2015) Deep learning. *Nature* 521:436–44. [aBML]

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W. & Jackel, L. D. (1989) Backpropagation applied to handwritten zip code recognition. *Neural Computation* 1:541–51. [arBML]

LeCun, Y., Bottou, L., Bengio, Y. & Haffner, P. (1998) Gradient-based learning applied to document recognition. *Proceedings of the IEEE* 86(11):2278–323. [aBML]

Lee, T. S. (2015) The visual system's internal model of the world. *Proceedings of the IEEE* 103(8):1359–78. Available at: http://doi.org/10.1109/JPROC.2015.2434601. [DGe]

Legare, C. H. & Harris, P. L. (2016) The ontogeny of cultural learning. *Child Development* 87(3):633–42. [JMC]

Lenat, D. & Guha, R. V. (1990) *Building large. Knowledge based systems: Representation and inference in the Cyc project*. Addison-Wesley. [LRC]

Lenat, D., Miller, G. & Yokoi, T (1995) CYC, WordNet, and EDR: Critiques and responses. *Communications of the ACM* 38(11):45–48. [LRC]

Lerer, A., Gross, S. & Fergus, R. (2016) Learning physical intuition of block towers by example. Presented at the 33rd International Conference on Machine Learning. *Proceedings of Machine Learning Research* 48:430–08. [aBML]

Levy, R. P., Reali, F. & Griffiths, T. L. (2009) Modeling the effects of memory on human online sentence processing with particle filters. Presented at the 2008 Neural Information Processing Systems conference. Vancouver, BC, Canada, December 8–10, 2008. In: *Advances in neural information processing systems 21 (NIPS 2008),* pp. 937–44. Neural Information Processing Systems. [aBML]

Lewandowsky, S., Ecker, U. K., Seifert, C. M., Schwarz, N. & Cook, J. (2012) Misinformation and its correction continued influence and successful debiasing. *Psychological Science in the Public Interest* 13(3):106–31. [EJL]

Lewis-Kraus, G. (2016) Going neural. *New York Times Sunday Magazine* 40–49+, December 18, 2016. [DEM]

Liang, C. and Forbus, K. (2015) Learning plausible inferences from semantic web knowledge by combining analogical generalization with structured logistic regression. In: *Proceedings of the 29th AAAI Conference on Artificial Intelligence, Austin, TX*. AAAI Press. [KDF]

Liao, Q., Leibo, J. Z. & Poggio, T. (2015) How important is weight symmetry in backpropagation? *arXiv preprint arXiv:1510.05067*. Available at: https://arxiv.org/abs/1510.05067. [aBML]

Liberman, A. M., Cooper, F. S., Shankweiler, D. P. & Studdert-Kennedy, M. (1967) Perception of the speech code. *Psychological Review* 74(6):431–61. [aBML]

Lillicrap, T. P., Cownden, D., Tweed, D. B. & Akerman, C. J. (2014) Random feedback weights support learning in deep neural networks. *arXiv preprint:1411.0247*. Available at: https://arxiv.org/abs/1411.0247. [aBML]

Lindeman, M. (2011) Biases in intuitive reasoning and belief in complementary and alternative medicine. *Psychology and Health* 26(3):371–82. [EJL]

Lisman, J. E. & Grace, A. A. (2005) The hippocampal-VTA loop: Controlling the entry of information into long-term memory. *Neuron* 46:703–13. [GB]

Liu, D., Wellman, H. M., Tardif, T., & Sabbagh, M. A. (2008). Theory of mind development in Chinese children: A meta-analysis of false-belief understanding across cultures and languages. *Developmental Psychology* 44(2):523–31. Available at: http://dx.doi.org/10.1037/0012-1649.44.2.523. [JMC, rBML]

Lloyd, J., Duvenaud, D., Grosse, R., Tenenbaum, J. & Ghahramani, Z. (2014) Automatic construction and natural-language description of nonparametric regression models. In: *Proceedings of the national conference on artificial intelligence* 2:1242–50. [aBML]

Logan, G. D. (1988) Toward an instance theory of automatization. *Psychological Review* 95(4):492–527. [NC]

Lombrozo, T. (2009) Explanation and categorization: How "why?" informs "what?". *Cognition* 110(2):248–53. [aBML]

Lombrozo, T. (2016) Explanatory preferences shape learning and inference. *Trends in Cognitive Sciences* 20(10):748–59. [rBML]

Lopes, M. & Oudeyer, P.-Y. (2012) The strategic student approach for life-long exploration and learning. In: *IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL)*, San Diego, CA, November 7–9, 2012, pp. 1–8. IEEE. [P-YO]

Lopes, M. & Santos-Victor, J. (2007). A developmental roadmap for learning by imitation in robots. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics* 37(2):308–21. [LM]

Lopez-Paz, D., Bottou, L., Scholköpf, B. & Vapnik, V. (2016) Unifying distillation and privileged information. Presented at the International Conference on Learning Representations (ICLR), San Juan, Puerto Rico, May 2–4, 2016. *arXiv preprint 1511.03643v3*. Available at: https://arxiv.org/abs/1511.03643. [aBML]

Lopez-Paz, D., Muandet, K., Scholköpf, B. & Tolstikhin, I. (2015) Towards a learning theory of cause-effect inference. Presented at the 32nd International Conference on Machine Learning (ICML), Lille, France, July 7–9, 2015. *Proceedings of Machine Learning Research* 37:1452–61. [aBML]

Loughnan, S. & Haslam, N. (2007) Animals and androids implicit associations between social categories and nonhumans. *Psychological Science* 18(2):116–21. [rBML]

Lovett, A. & Forbus, K. (2017) Modeling visual problem solving as analogical reasoning. *Psychological Review* 124(1):60–90. [KDF]

Lungarella, M., Metta, G., Pfeifer, R. & Sandini, G. (2003) Developmental robotics: A survey. *Connection Science* 15:151–90. [BJM]

Luong, M.-T., Le, Q. V., Sutskever, I., Vinyals, O. & Kaiser, L. (2015) Multi-task sequence to sequence learning. *arXiv preprint 1511.06114*. Available at: https://arxiv.org/pdf/1511.06114.pdf. [aBML]

Lupyan, G. & Bergen, B. (2016) How language programs the mind. *Topics in Cognitive Science* 8(2):408–24. [aBML]

Lupyan, G. & Clark, A. (2015) Words and the world: Predictive coding and the language perception-cognition interface. *Current Directions in Psychological Science* 24(4):279–84. [aBML]

Macindoe, O. (2013) Sidekick agents for sequential planning problems. Unpublished doctoral dissertation, Massachusetts Institute of Technology. [aBML]

Mackenzie, D. (2012) A flapping of wings. *Science* 335(6075):1430–33. [DEM]

Magid, R. W., Sheskin, M. & Schulz, L. E. (2015) Imagination and the generation of new ideas. *Cognitive Development* 34:99–110. [aBML]

Mahoor, Z. MacLennan, B. & MacBride, A. (2016) Neurally plausible motor babbling in robot reaching. In: *The 6th Joint IEEE International Conference on Development and Learning and on Epigenetic Robotics, September 19–22, 2016, Cergy-Pontoise/Paris,* pp. 9–14. IEEE. [BJM]

Malle, B. F. & Scheutz, M. (2014) Moral competence in social robots. In: *Proceedings of the 2014 IEEE International Symposium on Ethics in Science, Technology and Engineering*. IEEE. doi:10.1109/ETHICS.2014.6893446. [PMP]

Mannella, F. & Baldassarre, G. (2015) Selection of cortical dynamics for motor behaviour by the basal ganglia. *Biological Cybernetics* 109:575–95. [GB]

Mannella, F., Gurney, K. & Baldassarre, G. (2013) The nucleus accumbens as a nexus between values and goals in goal-directed behavior: A review and a new hypothesis. *Frontiers in Behavioral Neuroscience* 7(135):e1–29. [GB]

Mansinghka, V., Selsam, D. & Perov, Y. (2014) Venture: A higher-order probabilistic programming platform with programmable inference. *arXiv preprint 1404.0099*. Available at: https://arxiv.org/abs/1404.0099 [aBML]

Marblestone, A. H., Wayne, G. & Kording, K. P. (2016) Toward an integration of deep learning and neuroscience. *Frontiers in Computational Neuroscience* 10:94. [AHM, NK]

Marcus, G. (1998) Rethinking eliminative connectionism. *Cognitive Psychology* 282 (37):243–82. [aBML]

Marcus, G. (2001) *The algebraic mind: Integrating connectionism and cognitive science*. MIT Press. [aBML]

Marin, L., Issartel, J. & Chaminade, T. (2009). Interpersonal motor coordination: From human-human to human-robot interactions. *Interaction Studies* 10 (3):479–504. [LM]

Markman, A. B. & Makin, V. S. (1998) Referential communication and category acquisition. *Journal of Experimental Psychology: General* 127(4):331–54. [aBML]

Markman, A. B. & Ross, B. H. (2003) Category use and category learning. *Psychological Bulletin* 129(4):592–613. [aBML]

Markman, E. M. (1989) *Categorization and naming in children*. MIT Press. [aBML]

Marr, D. (1982/2010). *Vision*. MIT Press. [ESS]

Marr, D. (1982) *Vision: A computational investigation into the human representation and processing of visual information*. MIT Press. [SSH]

Marr, D. (1983) *Vision*. W. H. Freeman. [KDF]

Marr, D. C. (1982) *Vision*. W. H. Freeman. [aBML]

Marr, D. C. & Nishihara, H. K. (1978) Representation and recognition of the spatial organization of three-dimensional shapes. *Proceedings of the Royal Society of London Series B: Biological Sciences* 200(1140):269–94. [aBML]

Mascalzoni, E., Regolin, L. & Vallortigara, G. (2010). Innate sensitivity for self-propelled causal agency in newly hatched chicks. *Proceedings of the National Academy of Sciences of the United States of America* 107(9):4483–85. [ESS]

Maslow, A. (1954) *Motivation and personality*. Harper & Brothers. [CDG]

Matute, H., Blanco, F., Yarritu, I., Díaz-Lago, M., Vadillo, M. A. & Barberia, I. (2015) Illusions of causality: How they bias our everyday thinking and how they could be reduced. *Frontiers in Psychology* 6:888. doi: 10.3389/fpsyg.2015.00888. [EJL]

Mayer, J. D. & Salovey, P. (1993) The intelligence of emotional intelligence. *Intelligence* 17:442–43. [RJS]

Mazur, J. E. & Hastie, R. (1978) Learning as accumulation: A reexamination of the learning curve. *Psychological Bulletin* 85:1256–74. [LRC]

McCarthy, J. (1959) Programs with common sense at the Wayback machine (archived October 4, 2013). In: *Proceedings of the Teddington Conference on the Mechanization of Thought Processes*, pp. 756–91. AAAI Press. [LRC]

McCarthy, J. & Hayes, P. J. (1969) Some philosophical problems from the standpoint of artificial intelligence. In: Machine Intelligence 4, ed. B. Meltzer & D. Michie, pp. 463–502. Edinburgh University Press. [LRC]

McClelland, J. L. (1988) *Parallel distributed processing: Implications for cognition and development* [technical report]. Defense Technical Information Center document. Available at: http://www.dtic.mil/get-tr-doc/pdf?AD=ADA219063. [aBML]

McClelland, J. L., Botvinick, M. M., Noelle, D. C., Plaut, D. C., Rogers, T. T., Seidenberg, M. S. & Smith, L. B. (2010) Letting structure emerge: Connectionist and dynamical systems approaches to cognition. *Trends in Cognitive Sciences* 14(8):348–56. [arBML]

McClelland, J. L., McNaughton, B. L. & O'Reilly, R. C. (1995) Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review* 102(3):419–57. [arBML]

McClelland, J. L. & Rumelhart, D. E. (1986) *Parallel distributed processing: Explorations in the microstructure of cognition, Vol. 2.* MIT Press. [aBML]

McFate, C. & Forbus, K. (2016) An analysis of frame semantics of continuous processes. *Proceedings of the 38th Annual Conference of the Cognitive Science Society, Philadelphia, PA*, ed. A. Papafragou, D. Grodner, D. Mirman & J. C. Trueswell, pp. 836–41. Cognitive Science Society. [KDF]

McFate, C. J., Forbus, K. & Hinrichs, T. (2014) Using narrative function to extract qualitative information from natural language texts. *Proceedings of the 28th AAAI Conference on Artificial Intelligence, Québec City, Canada, July 27–31, 2014*, pp. 373–379. AAAI Press. [KDF]

McShea, D. W. (2013) Machine wanting. *Studies on the History and Philosophy of Biological and Biomedical Sciences* 44(4 pt B):679–87. [KBC]

Medin, D. L. & Ortony, A. (1989). Psychological essentialism. In: *Similarity and analogical reasoning*, ed. S. Vosniadou & A. Ortony, pp. 179–95. Cambridge University Press. [rBML]

Medin, D. L. & Schaffer, M. M. (1978) Context theory of classification learning. *Psychological Review* 85(3):207–38. [NC]

Mejía-Arauz, R., Rogoff, B. & Paradise, R. (2005) Cultural variation in children's observation during a demonstration. *International Journal of Behavioral Development* 29(4):282–91. [JMC]

Meltzoff, A. N. (2007). 'Like me': A foundation for social cognition. *Developmental Science* 10(1):126–34. [LM]

Meltzoff, A. N., Kuhl, P. M., Movellan, J. & Sejnowski, T. J. (2009) Foundations for a new science of learning. *Science* 325(5938):284–88. [KBC]

Meltzoff, A. N. & Moore, M. K. (1995) Infants' understanding of people and things: From body imitation to folk psychology. In: *The body and the self*, ed. J. L. Bermúdez, A. Marcel & N. Eilan, pp. 43–70. MIT Press. [JMC]

Meltzoff, A. N. & Moore, M. K. (1997) Explaining facial imitation: a theoretical model. *Early Development and Parenting* 6:179–92. [BJM]

Mesoudi, A., Chang, L., Murray, K. & Lu, H. J. (2015) Higher frequency of social learning in China than in the West shows cultural variation in the dynamics of

cultural evolution. *Proceeding of the Royal Society of London Series B: Biological Sciences* 282(1798):20142209. [JMC]
Metcalfe, J., Cottrell, G. W. & Mencl, W. E. (1992) Cognitive binding: A computational-modeling analysis of a distinction between implicit and explicit memory. *Journal of Cognitive Neuroscience* 4(3):289–98. [LRC]
Mikolov, T., Joulin, A. & Baroni, M. (2016) A roadmap towards machine intelligence. *arXiv preprint 1511.08130*. Available at: http://arxiv.org/abs/1511.08130. [arBML]
Mikolov, T., Sutskever, I. & Chen, K. (2013) Distributed representations of words and phrases and their compositionality. Presented at the 2013 Neural Information Processing Systems conference, Lake Tahoe, NV, December 5–10, 2013. In: *Advances in Neural Information Processing Systems 26 (NIPS)*, ed C. J. C. Burges, L. Bottou, M. Welling, Z. Ghagramani & K. Q. Weinberger [poster].Neural Information Processing Systems Foundation. [aBML]
Miller, E. G., Matsakis, N. E. & Viola, P. A. (2000) Learning from one example through shared densities on transformations. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Hilton Head Island, SC, June 15, 2000*. IEEE. [aBML]
Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D. & Miller, K. J. (1990) Introduction to WordNet: An on-line lexical database. *International Journal of Lexicography* 3(4):235–44. [LRC]
Miller, G. A. & Johnson-Laird, P. N. (1976) *Language and perception*. Belknap Press. [aBML]
Milner, D. & Goodale, M. (2006) *The visual brain in action*. Oxford University Press. [GB]
Minsky, M. (1986) *The society of mind*. Simon and Schuster. [MBu]
Minsky, M. (2003) *Semantic information processing*. MIT Press. [RJS]
Minsky, M. & Papert, S. A. (1987) *Perceptrons: An introduction to computational geometry*, expanded edn. MIT Press. [RJS]
Minsky, M. L. (1974) A framework for representing knowledge. MIT-AI Laboratory Memo 306. [aBML]
Minsky, M. L. & Papert, S. A. (1969) *Perceptrons: An introduction to computational geometry*. MIT Press. [aBML]
Mirolli, M., Mannella, F. & Baldassarre, G. (2010) The roles of the amygdala in the affective regulation of body, brain and behaviour. *Connection Science* 22 (3):215–45. [GB]
Mišić, B., Sporns, O. & McIntosh, A. R. (2014) Communication efficiency and congestion of signal traffic in large-scale brain networks. *PLoS Computational Biology* 10(1):e1003427. [DG]
Mitchell, T. M., Keller, R. R. & Kedar-Cabelli, S. T. (1986) Explanation-based generalization: A unifying view. *Machine Learning* 1:47–80. [aBML]
Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S. & Floridi, L. (2016) The ethics of algorithms: Mapping the debate. *Big Data & Society* 3(2):1–21. doi:10.1177/2053951716679679. [PMP]
Mix, K. S. (1999) Similarity and numerical equivalence: Appearances count. *Cognitive Development* 14:269–97. [KDF]
Mnih, A. & Gregor, K. (2014) Neural variational inference and learning in belief networks. Presented at the 31st International Conference on Machine Learning, Beijing, China, June 22–24, 2014. *Proceedings of Machine Learning Research* 32:1791–99. [aBML]
Mnih, V., Heess, N., Graves, A. & Kavukcuoglu, K. (2014). Recurrent models of visual attention. Presented at the 28th Annual Conference on Neural Information Processing Systems, Montreal, Canada. In: *Advances in Neural Information Processing Systems* 27*(NIPS 2014)*, ed. Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence & K. Q. Weinberger. Neural Information Processing Systems Foundation. [arBML]
Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D. & Riedmiller, M. (2013) Playing Atari with deep reinforcement learning. *arXiv preprint 1312.5602*. Available at: https://arxiv.org/abs/1312.5602. [SSH]
Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglous, I., King, H., Kumaran, D., Wierstra, D. & Hassabis, D. (2015) Human-level control through deep reinforcement learning. *Nature* 518(7540):529–33. [arBML, MB, DGe]
Moeslund, T. B., Hilton, A. & Krüger, V. (2006) A survey of advances in vision-based human motion capture and analysis. *Computer Vision and Image Understanding* 104(2):90–126. [rBML]
Mogenson, G. J., Jones, D. L. & Yim, C. Y. (1980) From motivation to action: Functional interface between the limbic system and the motor system. *Progress in Neurobiology* 14(2–3):69–97. [GB]
Mohamed, S. & Rezende, D. J. (2015) Variational information maximisation for intrinsically motivated reinforcement learning. Presented at the 2015 Neural Information Processing Systems conference, Montreal, QC, Canada, December 7–12, 2015. *Advances in Neural Information Processing Systems 28 (NIPS 2015)*, ed. C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama & R. Garnett, pp. 2125–33. Neural Information Processing Systems Foundation. [aBML]
Moreno-Bote, R., Knill, D. C. & Pouget, A. (2011) Bayesian sampling in visual perception. *Proceedings of the National Academy of Sciences of the United States of America* 108:12491–96. [aBML]
Moser, E., Kropff, E. & Moser, M. B. (2008). Place cells, grid cells, and the brain's spatial representation system. *Annual Review of Neuroscience* 31:69–89. [ESS]
Moulin-Frier, C., Nguyen, M. & Oudeyer, P.-Y. (2014) Self-organization of early vocal development in infants and machines: The role of intrinsic motivation. *Frontiers in Psychology* 4:1006. Available at: http://dx.doi.org/10.3389/fpsyg.2013.01006. [P-YO]
Murphy, G. L. (1988) Comprehending complex concepts. *Cognitive Science* 12 (4):529–62. [aBML]
Murphy, G. L. & Medin, D. L. (1985) The role of theories in conceptual coherence. *Psychological Review* 92(3):289–316. [arBML]
Murphy, G. L. & Ross, B. H. (1994) Predictions from uncertain categorizations. *Cognitive Psychology* 27:148–93. [aBML]
Nagai, Y., Kawai, Y. & Asada, M. (2011). Emergence of mirror neuron system: Immature vision leads to self-other correspondence. *Proceedings of the 1st Joint IEEE International Conference on Development and Learning and on Epigenetic Robotics, Vol. 2*, pp. 1–6. IEEE. [LM]
Nakayama, K., Shimojo, S. & Silverman, G. H. (1989) Stereoscopic depth: Its relation to image segmentation, grouping, and the recognition of occluded objects. *Perception* 18:55–68. [rBML]
Neisser, U. (1966) *Cognitive psychology*. Appleton-Century-Crofts. [aBML]
Newell, A. (1990) *Unified theories of cognition*. Harvard University Press. [RPK]
Newell, A., Shaw, J. C. & Simon, H. A. (1957) Problem solving in humans and computers. *Carnegie Technical* 21(4):34–38. [RJS]
Newell, A. & Simon, H. (1956) The logic theory machine. A complex information processing system. *IRE Transactions on Information Theory* 2(3):61–79. [LRC]
Newell, A. & Simon, H. A. (1961) *GPS, A program that simulates human thought*. Defense Technical Information Center. [aBML]
Newell, A. & Simon, H. A. (1972) *Human problem solving*. Prentice-Hall. [aBML]
Nguyen, M. & Oudeyer, P.-Y. (2013) Active choice of teachers, learning strategies and goals for a socially guided intrinsic motivation learner, *Paladyn Journal of Behavioural Robotics* 3(3):136–46. [P-YO]
Nguyen-Tuong, D. & Peters, J. (2011) Model learning for robot control: A survey. *Cognitive Processing* 12(4):319–40. [P-YO]
Nisbett, R. E. & Ross, L. (1980) *Human inference: Strategies and shortcomings of social judgment*. Prentice-Hall. ISBN 0-13-445073-6. [KBC, NC]
Niv, Y. (2009) Reinforcement learning in the brain. *Journal of Mathematical Psychology* 53:139–54. [aBML]
Norman, D. A. & Shallice, T. (1986) Attention to action: willed and automatic control of behaviour. In: *Advances in research: Vol. IV. Consciousness and self regulation*, ed. R. Davidson, G. Schwartz & D. Shapiro. Plenum. [RPK]
Oaksford, M. & Chater, N. (1991) Against logicist cognitive science. *Mind and Language* 6(1):1–38. [NC]
O'Donnell, T. J. (2015) *Productivity and reuse in language: A theory of linguistic computation and storage*. MIT Press. [aBML]
O'Keefel (2014). *Nobel lecture: Spatial cells in the hippocampal formation*. Available at: http://www.nobelprize.org/nobel_prizes/medicine/laureates/2014/okeefe-lecture.html. [ESS]
O'Keefe, J. & Nadel, L. (1978). *The hippocampus as a cognitive map*. Oxford University Press. [ESS]
Olshausen, B. A., Anderson, C. H. & Van Essen, D. C. (1993) A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *The Journal of Neuroscience* 13(11):4700–19. [DG]
Ong, D. C., Zaki, J. & Goodman, N. D. (2015) Affective cognition: Exploring lay theories of emotion. *Cognition* 143:141–62. [rBML]
O'Regan, J.K. (2011). *Why red doesn't sound like a bell: Understanding the feel of consciousness*. Oxford University Press. [LM]
Osherson, D. N. & Smith, E. E. (1981) On the adequacy of prototype theory as a theory of concepts. *Cognition* 9(1):35–58. [aBML]
Otto, A. R., Skatova, A., Madlon-Kay, S. & Daw, N. D. (2015) Cognitive control predicts use of model-based reinforcement learning. *Journal of Cognitive Neuroscience* 27:319–33. [EJL]
Oudeyer, P.-Y. (2016) What do we learn about development from baby robots? *WIREs Cognitive Science* 8(1–2):e1395. Available at: http://www.pyoudeyer.com/oudeyerWiley16.pdf. doi: 10.1002/wcs.1395. [P-YO]
Oudeyer, P.-Y., Baranes, A. & Kaplan, F. (2013) Intrinsically motivated learning of real-world sensorimotor skills with developmental constraints. In: *Intrinsically motivated learning in natural and artificial systems*, ed. G. Baldassarre & M. Mirolli, pp. 303–65. Springer. [P-YO]
Oudeyer, P.-Y., Kaplan, F. & Hafner, V. (2007) Intrinsic motivation systems for autonomous mental development. *IEEE Transactions on Evolutionary Computation* 11(2):265–86. [P-YO]
Oudeyer, P-Y. & Smith, L. (2016) How evolution may work through curiosity-driven developmental process. *Topics in Cognitive Science* 8(2):492–502. [P-YO]
Palmer, S. (1999) *Vision science: Photons to phenomenology*. MIT Press. [KDF]
Parisotto, E., Ba, J. L. & Salakhutdinov, R. (2016) Actor-mimic: Deep multitask and transfer reinforcement learning. Presented at the International Conference on

Learning Representations (ICLR), San Juan, Puerto Rico. May 2–5, 2016. *arXiv preprint 1511.06342v4*. Available at: https://www.google.com/search?q=arXiv%3A+preprint+1511.06342v4&ie=utf-8&oe=utf-8. [aBML]

Parker, S. T. & McKinney, M. L. (1999) *Origins of intelligence: The evolution of cognitive development in monkeys, apes and humans*. Johns Hopkins University Press. ISBN 0-8018-6012-1. [KBC]

Pecevski, D., Buesing, L. & Maass, W. (2011) Probabilistic inference in general graphical models through sampling in stochastic networks of spiking neurons. *PLoS Computational Biology* 7:e1002294. [aBML]

Penhune, V. B. & Steele, C. J. (2012) Parallel contributions of cerebellar, striatal and M1 mechanisms to motor sequence learning. *Behavioural Brain Research* 226 (2):579–91. [GB]

Peterson, J. C., Abbott, J. T. & Griffiths, T. L. (2016) Adapting deep network features to capture psychological representations. In: *Proceedings of the 38th Annual Conference of the Cognitive Science Society, Philadelphia, Pennsylvania, August 10–13, 2016*, ed. A. Papafragou, Daniel J. Grodner, D. Mirman & J. Trueswell, pp. 2363–68. Cognitive Science Society. [aBML]

Pfeifer, R. & Gómez, G. (2009) Morphological computation–connecting brain, body, and environment. In: *Creating brain-like intelligence*, ed. B. Sendhoff, E Körner, H. Ritter & K. Doya, pp. 66–83. Springer. [GB]

Pfeifer, R., Lungarella, M. & Iida, F. (2007) Self-organization, embodiment, and biologically inspired robotics. *Science* 318(5853):1088–93. [P-YO]

Piantadosi, S. T. (2011) Learning and the language of thought. Unpublished doctoral dissertation, Massachusetts Institute of Technology. [aBML]

Pinker, S. (2007) *The stuff of thought: Language as a window into human nature*. Penguin. [aBML]

Pinker, S. & Prince, A. (1988) On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition* 28:73–193. [aBML]

Poggio, T. (1984) Routing thoughts. Massachusetts Institute of Technology Artificial Intelligence Laboratory Working Paper 258. [DG]

Power, J. M., Thompson, L. T., Moyer, J. R. & Disterhoft, J. F. (1997) Enhanced synaptic transmission in ca1 hippocampus after eyeblink conditioning. *Journal of Neurophysiology* 78:1184–87. [aBML]

Prasada, S. & Pinker, S. (1993) Generalizations of regular and irregular morphology. *Language and Cognitive Processes* 8(1):1–56. [LRC]

Pratt, G. (2016, December 6). Presentation to Professor Deb Roy's class on machine learning and society at the MIT Media Lab. Class presentation that was videotaped but has not been made public. [DCD]

Premack, D. & Premack, A. J. (1997) Infants attribute value to the goal-directed actions of self-propelled objects. *Cognitive Neuroscience* 9(6):848–56. doi: 10.1162/jocn.1997.9.6.848. [aBML]

Putnam, H. (1967) Psychophysical predicates. In: *Art, mind, and religion*, ed. W. Capitan & D. Merrill. University of Pittsburgh Press. (Reprinted in 1975 as *The nature of mental states*, pp. 429–40. Putnam.) [LRC]

Ranzato, M., Szlam, A., Bruna, J., Mathieu, M., Collobert, R. & Chopra, S. (2016) Video (language) modeling: A baseline for generative models of natural videos. *arXiv preprint 1412.6604*. Available at: https://www.google.com/search?q=arXiv+preprint+1412.6604&ie=utf-8&oe=utf-8. [MB]

Raposo, D., Santoro, A., Barrett, D. G. T., Pascanu, R., Lillicrap, T. & Battaglia, P. (2017) Discovering objects and their relations from entangled scene representations. Presented at the Workshop Track at the International Conference on Learning Representations, Toulon, France, April 24–26, 2017. *arXiv preprint 1702.05068*. Available at: https://openreview.net/pdf?id=Bk2TqVcxe. [MB, rBML]

Ravi, S. & Larochelle, H. (2017) Optimization as a model for few-shot learning. Presented at the International Conference on Learning Representations, Toulon, France, April 24–26, 2017. Available at: https://openreview.net/pdf?id=rJY0-Kcll. [MB]

Read, S. J., Monroe, B. M., Brownstein, A. L., Yang, Y., Chopra, G. & Miller, L. C. (2010) A neural network model of the structure and dynamics of human personality. *Psychological Reviews* 117(1):61–92. [KBC]

Real, E., Moore, S., Selle, A., Saxena, S., Suematsu, Y. L., Le, Q. & Kurakin, A. (2017) Large-scale evolution of image classifiers. *arXiv preprint 1703.01041*. Available at: https://arxiv.org/abs/1703.01041. [rBML]

Redgrave, P. & Gurney, K. (2006) The short-latency dopamine signal: a role in discovering novel actions? *Nature Reviews Neuroscience* 7:967–75. [GB]

Reed, S. & de Freitas, N. (2016) Neural programmer-interpreters. Presented at the 4th International Conference on Learning Representations (ICLR), San Juan, Puerto Rico, May 2–5, 2016. *arXiv preprint 1511.06279*. Available at: https://arxiv.org/abs/1511.06279. [arBML, MB]

Regolin, L., Vallortigara, G. & Zanforlin, M. (1995). Object and spatial representations in detour problems by chicks. *Animal Behaviour* 49:195–99. [ESS]

Rehder, B. (2003) A causal-model theory of conceptual representation and categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 29(6):1141–59. [aBML]

Rehder, B. & Hastie, R. (2001) Causal knowledge and categories: The effects of causal beliefs on categorization, induction, and similarity. *Journal of Experimental Psychology: General* 130(3):323–60. [aBML]

Rehling, J. A. (2001) Letter spirit (part two): Modeling creativity in a visual domain. Unpublished doctoral dissertation, Indiana University. [aBML]

Rezende, D. J., Mohamed, S., Danihelka, I., Gregor, K. & Wierstra, D. (2016) One-shot generalization in deep generative models. Presented at the International Conference on Machine Learning, New York, NY, June 20–22, 2016. *Proceedings of Machine Learning Research* 48:1521–29. [arBML, MB]

Rezende, D. J., Mohamed, S. & Wierstra, D. (2014) Stochastic backpropagation and approximate inference in deep generative models. Presented at the International Conference on Machine Learning (ICML), Beijing, China, June 22–24, 2014. *Proceedings of Machine Learning Research* 32:1278–86. [aBML]

Richland, L. E. & Simms, N. (2015) Analogy, higher order thinking, and education. *Wiley Interdisciplinary Reviews: Cognitive Science* 6(2):177–92. [KDF]

Rips, L. J. (1975) Inductive judgments about natural categories. *Journal of Verbal Learning and Verbal Behavior* 14(6):665–81. [aBML]

Rips, L. J. & Hespos, S. J. (2015) Divisions of the physical world: Concepts of objects and substances. *Psychological Bulletin* 141:786–811. [aBML]

Rock, I. (1983) *The logic of perception*. MIT Press. [NC]

Rogers, T. T. & McClelland, J. L. (2004) *Semantic cognition*. MIT Press. [aBML]

Rogoff, B. (2003) *The cultural nature of human development*. Oxford University Press. [JMC]

Rohlfing, K. J. & Nomikou, I. (2014) Intermodal synchrony as a form of maternal responsiveness: Association with language development. *Language, Interaction and Acquisition* 5(1):117–36. [SW]

Romanes, G. J. (1884) *Animal intelligence*. Appleton. [KBC]

Rosenblatt, F. (1958) The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review* 65:386–408. [aBML]

Rougier, N. P., Noelle, D. C., Braver, T. S., Cohen, J. D. & O'Reilly, R. C. (2005) Prefrontal cortex and flexible cognitive control: Rules without symbols. *Proceedings of the National Academy of Sciences of the United States of America* 102(20):7338–43. [aBML]

Rozenblit, L. & Keil, F. (2002) The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science* 26(5):521–62. [EJL, NC]

Ruciński, M. (2014) Modelling learning to count in humanoid robots. Ph.D. thesis, University of Plymouth, UK. [SW]

Rumelhart, D. E., Hinton, G. & Williams, R. (1986a) Learning representations by back-propagating errors. *Nature* 323(9):533–36. [aBML]

Rumelhart, D. E. & McClelland, J. L. (1986) On learning the past tenses of English verbs. In: *Parallel distributed processing: Explorations in the microstructure of cognition, Vol. 1*, ed. Rumelhart, D. F., McClelland, J. L. & PDP Research Group, pp. 216–71. MIT Press. [aBML]

Rumelhart, D. E., McClelland, J. L. & PDP Research Group. (1986b) *Parallel distributed processing: Explorations in the microstructure of cognition, Vol. 1.* MIT Press. [aBML]

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C. & Fei-Fei, L. (2015) ImageNet large scale visual recognition. *International Journal of Computer Vision* 115 (3):211–52. [aBML]

Russell, S. & Norvig, P. (2003) *Artificial intelligence: A modern approach*. Prentice–Hall. [aBML]

Rusu, A. A., Rabinowitz, N. C., Desjardins, G., Soyer, H., Kirkpatrick, J., Kavukcuoglu, K., Pascanu, R. & Hadsell, R. (2016) Progressive neural networks. *arXiv preprint 1606.04671*. Available at: http://arxiv.org/abs/1606.04671. [aBML]

Ryan, R. M. & Deci, E. L. (2007) Intrinsic and extrinsic motivations: classic definitions and new directions. *Contemporary Educational Psychology* 25:54–67. [aBML]

Salakhutdinov, R., Tenenbaum, J. & Torralba, A. (2012) One-shot learning with a hierarchical nonparametric Bayesian model. *JMLR Workshop on Unsupervised and Transfer Learning* 27:195–207. [aBML]

Salakhutdinov, R., Tenenbaum, J. B. & Torralba, A. (2013) Learning with hierarchical-deep models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35(8):1958–71. [aBML]

Salakhutdinov, R., Torralba, A. & Tenenbaum, J. (2011) Learning to share visual appearance for multiclass object detection. In: *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Colorado Springs, CO, June 20–25, 2011*, pp. 1481–88. IEEE. [aBML]

Sanborn, A. N., Mansingkha, V. K. & Griffiths, T. L. (2013) Reconciling intuitive physics and Newtonian mechanics for colliding objects. *Psychological Review* 120(2):411–37. [aBML, ED]

Santoro, A., Bartunov, S., Botvinick, M., Wierstra, D. & Lillicrap, T. (2016). Meta-learning with memory-augmented neural networks. Presented at the 33rd International Conference on Machine Learning, New York, NY, June 19–24, 2016. *Proceedings of Machine Learning Research* 48:1842–50. [MB, rBML]

Santoro, A., Bartunov, S., Botvinick, M., Wierstra, D., & Lillicrap, T. (2016). One-shot learning with memory-augmented neural networks. *arXiv preprint 1605.06065*. Available at: https://arxiv.org/abs/1605.06065. [SSH]

Santucci, V. G., Baldassarre, G. & Mirolli, M. (2016), GRAIL: A goal-discovering robotic architecture for intrinsically-motivated learning, *IEEE Transactions on Cognitive and Developmental Systems* 8(3):214–31. [GB]

Saxe, A. M., McClelland, J. L. & Ganguli, S. (2013) Dynamics of learning in deep linear neural networks. Presented at the NIPS 2013 Deep Learning Workshop, Lake Tahoe, NV, December 9, 2013. [LRC]

Saxe, A. M., McClelland, J. L. & Ganguli, S. (2014) Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. Presented at the International Conference on Learning Representations, Banff, Canada, April 14–16, 2014. *arXiv preprint 1312.6120*. Available at:https://arxiv.org/abs/1312.6120. [LRC]

Scellier, B. & Bengio, Y. (2016) Towards a biologically plausible backprop. *arXiv preprint 1602.05179*. Available at: https://arxiv.org/abs/1602.05179v2. [aBML]

Schank, R. C. (1972) Conceptual dependency: A theory of natural language understanding. *Cognitive Psychology* 3:552–631. [aBML]

Schaul, T., Quan, J., Antonoglou, I. & Silver, D. (2016) Prioritized experience replay. Presented at International Conference on Learning Representations (ICLR), San Diego, CA, May 7–9, 2015. *arXiv preprint 1511.05952*. Available at: https://arxiv.org/abs/1511.05952. [aBML, MB]

Schlegel, A., Alexander, P. & Peter, U. T. (2015) Information processing in the mental workspace is fundamentally distributed. *Journal of Cognitive Neuroscience* 28(2):295–307. [DG]

Schlottmann, A., Cole, K., Watts, R. & White, M. (2013) Domain-specific perceptual causality in children depends on the spatio-temporal configuration, not motion onset. *Frontiers in Psychology* 4:365. [aBML]

Schlottmann, A., Ray, E. D., Mitchell, A. & Demetriou, N. (2006) Perceived physical and social causality in animated motions: Spontaneous reports and ratings. *Acta Psychologica* 123:112–43. [aBML]

Schmidhuber, J. (1991) Curious model-building control systems. *Proceedings of the IEEE International Joint Conference on Neural Networks* 2:1458–63. [P-YO]

Schmidhuber, J. (2015) Deep learning in neural networks: An overview. *Neural Networks* 61:85–117. [aBML]

Schmidt, R.C. & Richardson, M.J. (2008). Dynamics of interpersonal coordination. In: *Coordination: Neural, behavioural and social dynamics,* ed. A. Fuchs & V. Jirsa, pp. 281–307. Springer-Verlag. [LM]

Scholl, B. J. & Gao, T. (2013) Perceiving animacy and intentionality: Visual processing or higher-level judgment? In: *Social perception: detection and interpretation of animacy, agency, and intention*, ed. M. D. Rutherford & V. A. Kuhlmeier. MIT Press Scholarship Online. [aBML]

Schuller, I. K., Stevens, R. & Committee Chairs (2015) *Neuromorphic computing: From materials to architectures. Report of a roundtable convened to consider neuromorphic computing basic research needs*. Office of Science, U.S. Department of Energy. [KBC]

Schultz, W., Dayan, P. & Montague, P. R. (1997) A neural substrate of prediction and reward. *Science* 275:1593–99. [aBML]

Schulz, L. (2012a) Finding new facts; thinking new thoughts. Rational constructivism in cognitive development. *Advances in Child Development and Behavior* 43:269–94. [rBML]

Schulz, L. (2012b) The origins of inquiry: Inductive inference and exploration in early childhood. *Trends in Cognitive Sciences* 16(7):382–89. [arBML]

Schulz, L. E., Gopnik, A. & Glymour, C. (2007) Preschool children learn about causal structure from conditional interventions. *Developmental Science* 10:322–32. [aBML]

Scott, R. (Director). (2007) Blade Runner: The Final Cut: Warner Brothers (original release, 1982). [DEM]

Scott, S. H. (2004) Optimal feedback control and the neural basis of volitional motor control. *Nature Reviews Neuroscience* 5(7):532–46. [GB]

Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R. & LeCun, Y. (2014) OverFeat: Integrated recognition, localization and detection using convolutional networks. Presented at the International Conference on Learning Representations (ICLR), Banff, Canada, April 14–16, 2014. *arXiv preprint 1312.6229v4*. Available at: https://arxiv.org/abs/1312.6229. [aBML]

Shadmehr, R. & Krakauer, J. W. (2008) A computational neuroanatomy for motor control. *Experimental Brain Research* 185(3):359–81. [GB]

Shafto, P., Goodman, N. D. & Frank, M. C. (2012) Learning from others: The consequences of psychological reasoning for human learning. *Perspectives on Psychological Science* 7(4):341–51. [MHT]

Shafto, P., Goodman, N. D. & Griffiths, T. L. (2014) A rational account of pedagogical reasoning: Teaching by, and learning from, examples. *Cognitive Psychology* 71:55–89. [aBML]

Shahaeian, A., Peterson, C. C., Slaughter, V. & Wellman, H. M. (2011) Culture and the sequence of steps in theory of mind development. *Developmental Psychology* 47(5):1239–47. [JMC]

Shallice, T. & Cooper, R. P. (2011) *The organisation of mind*. Oxford University Press. [RPK]

Shultz, T. R. (2003) *Computational developmental psychology*. MIT Press. [aBML]

Siegler, R. (1976) Three aspects of cognitive development. *Cognitive Psychology* 8 (4):481–520. [ED]

Siegler, R. S. & Chen, Z. (1998) Developmental differences in rule learning: A microgenetic analysis. *Cognitive Psychology* 36(3):273–310. [aBML]

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Driessche, G. V. D., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K, Graepel, T. & Hassabis, D. (2016) Mastering the game of go with deep neural networks and tree search. *Nature* 529(7585):484–89. [arBML, MB]

Silver, D., van Hasselt, H., Hessel, M., Schaul, T., Guez, A., Harley, T., Dulac-Arnold, G. Reichert, D., Rabinowitz, N., Barreto, A. & Degris, T. (2017) The predictron: End-to-end learning and planning. In: *Proceedings of the 34rd International Conference on Machine Learning, Sydney, Australia,* ed. M. F. Balcan & K. Q. Weinberger. [MB]

Silverman, R. D. & Hendrix, K. S. (2015) Point: Should childhood vaccination against measles be a mandatory requirement for attending school? Yes. *CHEST Journal* 148(4):852–54. [EJL]

Simon, H. A. (1967) Motivational and emotional controls of cognition. *Psychological Review* 74:29–39. [CDG]

Sizemore, A., Giusti, C., Betzel, R. F. & Bassett, D. S. (2016) Closures and cavities in the human connectome. *arXiv preprint 1608.03520*. Available at: https://arxiv.org/abs/1608.03520. [DG]

Smith, L. B., Jones, S. S., Landau, B., Gershkoff-Stowe, L. & Samuelson, L. (2002) Object name learning provides on-the-job training for attention. *Psychological Science* 13(1):13–19. [aBML]

Socher, R., Perelygin, A., Wu, J. Y., Chuang, J., Manning, C. D., Ng, A. Y. & Potts, C. (2013) Recursive deep models for semantic compositionality over a sentiment treebank. In: *Proceedings of the Conference on EmpiricalMethods in Natural Language Processing (EMNLP), Seattle, WA*, vol. 1631, p. 1642. Association for Computational Linguistics. [rBML]

Solomon, K., Medin, D. & Lynch, E. (1999) Concepts do more than categorize. *Trends in Cognitive Sciences* 3(3):99–105. [aBML]

Spelke, E. S. (1990) Principles of object perception. *Cognitive Science* 14(1):29–56. [aBML, JMC]

Spelke, E. S. (2003) What makes us smart? Core knowledge and natural language. Spelke ES. What makes us smart? Core knowledge and natural language. In: *Language in mind: Advances in the Investigation of language and thought*, ed. D. Gentner & S. Goldin-Meadow, pp. 277–311. MIT Press. [arBML]

Spelke, E. S., Gutheil, G. & Van de Walle, G. (1995) The development of object perception. In: *An invitation to cognitive science: vol. 2. Visual cognition,* 2nd ed. pp. 297–330. Bradford. [aBML]

Spelke, E. S. & Kinzler, K. D. (2007) Core knowledge. *Developmental Science* 10 (1):89–96. [arBML]

Spelke, E. S. & Lee, S. A. (2012). Core systems of geometry in animal minds. *Philosophical Transactions of the Royal Society, B: Biological Sciences* 367 (1603):2784–93. [ESS]

Squire, L. (1992). Memory and the hippocampus: A synthesis from findings with rats, monkeys and humans. *Psychological Review* 99(2):195–231. [ESS]

Srivastava, N. & Salakhutdinov, R. (2013) Discriminative transfer learning with tree-based priors. Presented at the 2013 Neural Information Processing Systems conference, Lake Tahoe, NV, December 5–10, 2013. In: *Advances in Neural Information Processing Systems 26 (NIPS 2013)*, ed. C J. C. Burges, L. Bottou, M. Welling, Z. Ghagramani & K. Q. Weinberger [poster]. Neural Information Processing Systems Foundation. [aBML]

Stadie, B. C., Levine, S. & Abbeel, P. (2016) Incentivizing exploration in reinforcement learning with deep predictive models. *arXiv preprint 1507.00814*. Available at: http://arxiv.org/abs/1507.00814. [aBML]

Stahl, A. E. & Feigenson, L. (2015) Observing the unexpected enhances infants' learning and exploration. *Science* 348(6230):91–94. [aBML]

Stanley, K. O. & Miikkulainen, R. (2002) Evolving neural networks through augmenting topologies. *Evolutionary Computation* 10(2):99–127. [rBML]

Stanovich, K. E. (2009) *What intelligence tests miss: The psychology of rational thought*. Yale University Press. [RJS]

Sterelny, K. (2012) *The evolved apprentice*. MIT Press. [DCD]

Sterelny, K. (2013) The informational commonwealth. In: *Arguing about human nature: Contemporary debates*, ed. L S. M. Downes & E. Machery, pp. 274–88. Routledge, Taylor & Francis. [DCD]

Sternberg, R. J. (1997) What does it mean to be smart? *Educational Leadership* 54 (6):20–24. [RJS]

Sternberg, R. J., ed. (2002) *Why smart people can be so stupid*. Yale University Press. [RJS]

Sternberg, R. J. & Davidson, J. E. (1995) *The nature of insight*. MIT Press. [aBML]

Sternberg, R. J. & Jordan, J., eds. (2005) *Handbook of wisdom: Psychological perspectives.* Cambridge University Press. [RJS]

Stuhlmüller, A., Taylor, J. & Goodman, N. D. (2013) Learning stochastic inverses. Presented at the 2013 Neural Information Processing Systems conference, Lake Tahoe, NV, December 5–10, 2013. In: *Advances in Neural Information Processing Systems 26 (NIPS 2013)*, ed. C J. C. Burges, L. Bottou, M. Welling, Z. Ghagramani & K. Q. Weinberger, pp. 3048–56. Neural Information Processing Systems Foundation. [aBML]

Sukhbaatar, S., Szlam, A., Weston, J. & Fergus, R. (2015) End-to-end memory networks. Presented at the 2015 Neural Information Processing Systems conference, Montreal, QC, Canada, December 7–12, 2015. In: *Advances in neural information processing systems 28 (NIPS 2015)*, ed. C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama & R. Garnett [oral presentation]. Neural Information Processing Systems Foundation. [arBML]

Sun, R. (2016) *Anatomy of the mind*. Oxford University Press. [CDG]

Super, C. M. & Harkness, S. (2002) Culture structures the environment for development. *Human Development* 45(4):270–74. [JMC]

Sutton, R. S. (1990) Integrated architectures for learning, planning, and reacting based on approximating dynamic programming. In: *Proceedings of the 7th International Workshop on Machine Learning (ICML), Austin, TX*, pp. 216–24. International Machine Learning Society. [aBML]

Svedholm, A. M. & Lindeman, M. (2013) Healing, mental energy in the physics classroom: Energy conceptions and trust in complementary and alternative medicine in grade 10–12 students. *Science & Education* 22(3):677–94. [EJL]

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. & Rabinovich, A. (2014) Going deeper with convolutions. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, June 7–12, 2015*, pp. 1–9. IEEE. [aBML]

Tan, L. H., Spinks, J. A., Eden, G. F., Perfetti, C. A. & Siok, W. T. (2005) Reading depends on writing, in Chinese. *Proceedings of the National Academy of Sciences of the United States of America* 102(24):8781–85. [ED]

Tauber, S. & Steyvers, M. (2011) Using inverse planning and theory of mind for social goal inference. In: *Proceedings of the 33rd Annual Conference of the Cognitive Science Society, Boston, MA, July 20–23, 2011,* pp. 2480–85. Cognitive Science Society. [aBML]

Taylor, E. G. & Ahn, W.-K. (2012) Causal imprinting in causal structure learning. *Cognitive Psychology* 65:381–413. [EJL]

Téglás, E., Vul, E., Girotto, V., Gonzalez, M., Tenenbaum, J. B. & Bonatti, L. L. (2011) Pure reasoning in 12-month-old infants as probabilistic inference. *Science* 332(6033):1054–59. [aBML]

Tenenbaum, J. B., Kemp, C., Griffiths, T. L. & Goodman, N. D. (2011) How to grow a mind: Statistics, structure, and abstraction. *Science* 331(6022):1279–85. [aBML]

Thomaz, A. L. & Cakmak, M. (2013) Active social learning in humans and robots. In: *Social learning theory: Phylogenetic considerations across animal, plant, and microbial taxa*, ed. K. B. Clark, pp. 113–28. Nova Science. ISBN 978-1-62618-268-4. [KBC]

Thorwart, A. & Livesey E. J. (2016) Three ways that non-associative knowledge may affect associative learning processes. *Frontiers in Psychology* 7:2024. doi: 10.3389/fpsyg.2016.02024. [EJL]

Thurstone, L. L. (1919) The learning curve equation. *Psychological Monographs* 26 (3):2–33. [LRC]

Tian, Y. & Zhu, Y. (2016) Better computer Go player with neural network and long-term prediction. Presented at the International Conference on Learning Representations (ICLR), San Juan, Puerto Rico, May 2–4, 2016. *arXiv preprint 1511.06410*. Available at: https://arxiv.org/abs/1511.06410. [aBML]

Todd, P. M. & Gigerenzer, G. (2007) Environments that make us smart: Ecological rationality. *Current Directions in Psychological Science* 16(3):167–71. doi:10.1111/j.1467-8721.2007.00497.x. [PMP]

Tomai, E. & Forbus, K. (2008) Using qualitative reasoning for the attribution of moral responsibility. In: *Proceedings of the 30th Annual Conference of the Cognitive Science Society, Washington, DC, July 23–26, 2008*. Cognitive Science Society. [KDF]

Tomasello, M. (1999) *The cultural origins of human cognition*. Harvard University Press. [MHT]

Tomasello, M. (2010) *Origins of human communication*. MIT Press. [aBML]

Tompson, J. J., Jain, A., LeCun, Y. & Bregler, C. (2014). Joint training of a convolutional network and a graphical model for human pose estimation. Presented at the 28th Annual Conference on Neural Information Processing Systems, Montreal, Canada. In: *Advances in Neural Information Processing Systems* 27 (NIPS 2014), ed. Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence & K. Q. Weinberger, pp. 1799–807. Neural Information Processing Systems Foundation. [rBML]

Torralba, A., Murphy, K. P. & Freeman, W. T. (2007) Sharing visual features for multiclass and multiview object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29(5):854–69. [aBML]

Toshev, A. & Szegedy, C. (2014). Deeppose: Human pose estimation via deep neural networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH,* pp. 1653–60. IEEE. [rBML]

Tremoulet, P. D. & Feldman, J. (2000) Perception of animacy from the motion of a single object. *Perception* 29:943–51. [aBML]

Trettenbrein, P. C. (2016) The demise of the synapse as the locus of memory: A looming paradigm shift? *Frontiers in Systems Neuroscience* 10:88. [DG]

Tsividis, P., Gershman, S. J., Tenenbaum, J. B. & Schulz, L. (2013) Information selection in noisy environments with large action spaces. In: *Proceedings of the 36th Annual Conference of the Cognitive Science Society, Austin, TX*, pp. 1622–27. Cognitive Science Society. [aBML]

Tsividis, P., Tenenbaum, J. B. & Schulz, L. E. (2015) Constraints on hypothesis selection in causal learning. *Proceedings of the 37th Annual Conference of the Cognitive Sciences, Pasadena, CA, July 23–25, 2015*, pp. 2434–439. Cognitive Science Society. [aBML]

Tsividis, P. A., Pouncy, T., Xu, J. L., Tenenbaum, J. B. & Gershman, S. J. (2017) Human learning in Atari. In: *Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI) Spring Symposium on Science of Intelligence: Computational Principles of Natural and Artificial Intelligence, Stanford University, Palo Alto, CA, March 25–27, 2017.* AAAI Press. [MHT, rBML]

Turing, A. M. (1950) Computing machine and intelligence. *Mind* 59:433–60. Available at: http://mind.oxfordjournals.org/content/LIX/236/433. [aBML]

Turovsky, B. (2016) Found in translation: More accurate, fluent sentences in Google Translate. Available at: https://blog.google/products/translate/found-translation-more-accurate-fluent-sentences-google-translate/. [DEM]

Tversky, B. & Hemenway, K. (1984) Objects, parts, and categories. *Journal of Experimental Psychology: General* 113(2):169–91. [aBML]

Ullman, S., Harari, D. & Dorfman, N. (2012a) From simple innate biases to complex visual concepts. *Proceedings of the National Academy of Sciences of the United States of America* 109(44):18215–20. [aBML]

Ullman, T. D., Baker, C. L., Macindoe, O., Evans, O., Goodman, N. D. & Tenenbaum, J. B. (2009). Help or hinder: Bayesian models of social goal inference. Presented at the 2009 Annual Conference on Neural Information Systems Processing, Vancouver, BC, Canada, December 7–10, 2009. In: *Advances in Neural Information Processing Systems 22 (NIPS 2009)*, ed. Y. Bengio, D. Schuumans, J. D. Lafferty, C. K. I. Williams & A. Culotta. Neural Information Processing Systems Foundation. [rBML]

Ullman, T. D., Goodman, N. D. & Tenenbaum, J. B. (2012b) Theory learning as stochastic search in the language of thought. *Cognitive Development* 27(4):455–80. [aBML]

U.S. Postal Service Historian (2016) Pieces of mail handled, number of post offices, income, and expenses since 1789. Available at: https://about.usps.com/who-we-are/postal-history/pieces-of-mail-since-1789.htm. [DG]

van den Hengel, A., Russell, C., Dick, A., Bastian, J., Pooley, D., Fleming, L. & Agapitol, L. (2015) Part-based modelling of compound scenes from images. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, June 7–12, 2015,* pp. 878–86. IEEE. [aBML]

van den Oord, A., Kalchbrenner, N. & Kavukcuoglu, K. (2016). Pixel recurrent neural networks. Presented at the 33rd International Conference on Machine Learning, New York, NY. *Proceedings of Machine Learning Research* 48:1747–56. [MB]

van Hasselt, H., Guez, A. & Silver, D. (2016) Deep learning with double Q-learning. In: *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence and the Twenty-Eighth Innovative Applications of Artificial Intelligence Conference on Artificial Intelligence, Phoenix, AZ.* AAAI Press. [aBML]

Varlet, M., Marin, L., Capdevielle, D., Del-Monte, J., Schmidt, R.C., Salesse, R.N., Boulenger, J.-P., Bardy, B. & Raffard, S. (2014). Difficulty leading interpersonal coordination: Towards an embodied signature of social anxiety disorder. *Frontiers in Behavioral Neuroscience* 8:1–9. [LM]

Vinyals, O., Blundell, C., Lillicrap, T. & Wierstra, D. (2016) Matching networks for one shot learning. Vinyals, O., Blundell, C., Lillicrap, T. Kavukcuoglu, K. & Wierstra, D. (2016). Matching networks for one shot learning. Presented at the 2016 Neural Information Processing Systems conference, Barcelona, Spain, December 5–10, 2016. In: *Advances in Neural Information Processing Systems 29 (NIPS 2016),* ed. D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon & R. Garnett, pp. 3630–38. Neural Information Processing Systems Foundation. [arBML, MB, SSH]

Vinyals, O., Toshev, A., Bengio, S. & Erhan, D. (2014) Show and tell: A neural image caption generator. *arXiv preprint 1411.4555*. Available at: https://arxiv.org/abs/1411.4555. [aBML]

Viviani, P. & Stucchi, N. (1992). Biological movements look uniform: evidence of motor-perceptual interactions. *Journal of Experimental Psychology: Human, Perception & Performance* 18(3):603–23. [LM]

Vollmer, A-L., Mühlig, M., Steil, J. J., Pitsch, K., Fritsch, J., Rohlfing, K. & Wrede, B. (2014) Robots show us how to teach them: Feedback from robots shapes tutoring behavior during action learning, *PLoS One* 9(3):e91349. [P-YO]

Vul, E., Goodman, N., Griffiths, T. L. & Tenenbaum, J. B. (2014) One and done? Optimal decisions from very few samples. *Cognitive Science* 38(4):599–637. [aBML]

Vygotsky, L. S. (1978) Interaction between learning and development. In: *Mind in society: The development of higher psychological processes*, ed. M. Cole, V. John-Steiner, S. Scribner & E. Souberman, pp. 79–91. Harvard University Press. [RPK]

Wallach, W., Franklin, S. & Allen C. (2010) A conceptual and computational model of moral decision making in human and artificial agents. *Topics in Cognitive Science* 2:454–85. [KBC]

Wang, J. X., Kurth-Nelson, Z., Tirumala, D., Soyer, H., Leibo, J. Z., Munos, R., Blundell, C., Kumaran, D. & Botvinick, M. (2017). Learning to reinforcement learn. In: Presented at the 39th Annual Meeting of the Cognitive Science Society, London, July 26–29, 2017. *arXiv preprint 1611.05763*. Available at: https://arxiv.org/abs/1611.05763. [MB]

Wang, Z., Schaul, T., Hessel, M., Hasselt, H. van, Lanctot, M. & de Freitas, N. (2016) Dueling network architectures for deep reinforcement learning. *arXiv preprint 1511.06581*. Available at: http://arxiv.org/abs/1511.06581. [aBML]

Ward, T. B. (1994) Structured imagination: The role of category structure in exemplar generation. *Cognitive Psychology* 27:1–40. [aBML]

Watkins, C. J. & Dayan, P. (1992) Q-learning. *Machine Learning* 8:279–92. [aBML]

Weigmann, K. (2006) Robots emulating children. *EMBO Reports* 7(5):474–76. [KBC]

Weizenbaum, J. (1966) ELIZA—A computer program for the study of natural language communication between man and machine. *Communications of the ACM* 9(1):36–45. [RJS]

Wellman, H. M. & Gelman, S. A. (1992) Cognitive development: Foundational theories of core domains. *Annual Review of Psychology* 43:337–75. [arBML]

Wellman, H. M. & Gelman, S. A. (1998). Knowledge acquisition in foundational domains. In: *Handbook of child psychology: Vol. 2. Cognition, perception, and language development*, 5th ed., series ed. W. Damon, vol. ed. D. Kuhn & R. S. Siegler, pp. 523–73. Wiley. [arBML]

Weng, J., McClelland, J., Pentland, A., Sporns, O., Stockman, I., Sur, M. & Thelen, E. (2001) Autonomous mental development by robots and animals. *Science* 291 (5504):599–600. [GB]

Wermter, S., Palm, G., Weber, C. & Elshaw, M. (2005) Towards biomimetic neural learning for intelligent robots. In: *Biomimetic neural learning for intelligent robots*, ed. S. Wermter, G. Palm & M. Elshaw, pp. 1–18. Springer. [SW]

Weston, J., Bordes, A., Chopra, S., Rush, A. M., van Merriënboer, B., Joulin, A. & Mikolov, T. (2015a) Towards AI-complete question answering: A set of prerequisite toy tasks. *arXiv preprint 1502.05698*. Available at: https://arxiv.org/pdf/1502.05698.pdf. [SSH]

Weston, J., Chopra, S. & Bordes, A. (2015b) Memory networks. Presented at the International Conference on Learning Representations, San Diego, CA, May 7–9, 2015. arXiv:1410.3916. Available at: https://arxiv.org/abs/1410.3916. [arBML]

Williams, J. J. & Lombrozo, T. (2010) The role of explanation in discovery and generalization: Evidence from category learning. *Cognitive Science* 34(5):776–806. [aBML]

Wills, T. J., Cacucci, F., Burgess, N. & O'Keefe, J. (2010). Development of the hippocampal cognitive map in preweanling rats. *Science* 328(5985):1573–76. [ESS]

Winograd, T. (1972) Understanding natural language. *Cognitive Psychology* 3:1–191. [aBML, RJS]

Winston, P. H. (1975) Learning structural descriptions from examples. In: *The psychology of computer vision*, pp.157–210. McGraw-Hill. [aBML]

Wiskott, L. (2006). How does our visual system achieve shift and size invariance? In: *23 Problems in systems neuroscience*, ed. J. L. Van Hemmen & T. J. Sejnowski, pp. 322–40. Oxford University Press. [DG]

Wiskott, L. & von der Malsburg, C. (1996) Face recognition by dynamic link matching. In: *Lateral interactions in the cortex: structure and function*, ed. J. Sirosh, R. Miikkulainen and Y. Choe, ch 11. The UTCS Neural Networks Research Group. [DG]

Wolfram, S. (2002) *A new kind of science*. Wolfram Media. ISBN 1-57955-008-8. [KBC]

Wolpert, D. M., Miall, R. C. & Kawato, M. (1998) Internal models in the cerebellum. *Trends in Cognitive Science* 2(9):338–47. [GB]

Xu, F. & Tenenbaum, J. B. (2007) Word learning as Bayesian inference. *Psychological Review* 114(2):245–72. [aBML]

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R. & Bengio, Y. (2015) Show, attend and tell: Neural image caption generation with visual attention. Presented at the 2015 International Conference on Machine Learning. *Proceedings of Machine Learning Research* 37:2048–57. [arBML]

Yamada, Y., Mori, H. & Kuniyoshi, Y. (2010) A fetus and infant developmental scenario: Self-organization of goal-directed behaviors based on sensory constraints. In: *Proceedings of the 10th International Conference on Epigenetic Robotics, Őrenäs Slott, Sweden*, pp. 145–52. Lund University Cognitive Studies. [P-YO]

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D. & DiCarlo, J. J. (2014) Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences of the United States of America* 111(23):8619–24. [aBML, NK]

Yildirim, I., Kulkarni, T. D., Freiwald, W. A. & Tenenbaum, J. (2015) Efficient analysis-by-synthesis in vision: A computational framework, behavioral tests, and comparison with neural representations. In: *Proceedings of the 37th Annual Conference of the Cognitive Science Society, Pasadena, CA, July 22–25, 2015*. Cognitive Science Society. Available at: https://mindmodeling.org/cogsci2015/papers/0471/index.html. [aBML, NK]

Yosinski, J., Clune, J., Bengio, Y. & Lipson, H. (2014) How transferable are features in deep neural networks? Presented at the 2014 Neural Information Processing Systems conference, Montreal, QC, Canada. In: *Advances in neural information processing systems 27 (NIPS 2014)*, ed. Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence & K. Q. Weinberger [oral presentation]. Neural Information Processing Systems Foundation. [aBML]

Youyou, W., Kosinski, M. & Stillwell, D. (2015) Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences of the United States of America* 112(4):1036–40. [KBC]

Zeiler, M. D. & Fergus, R. (2014) Visualizing and understanding convolutional networks. In: *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part I*, ed. D. Fleet, T. Pajdla, B. Schiele & T. Tuytelaars, pp. 818–33. Springer. [aBML]

Zentall, T. R. (2013) Observational learning in animals. In: *Social learning theory: Phylogenetic considerations across animal, plant, and microbial taxa*, ed. K. B. Clark, pp. 3–33. Nova Science. ISBN 978-1-62618-268-4. [KBC]

Zhou, H., Friedman, H. S. & von der Heydt, R. (2000) Coding of border ownership in monkey visual cortex. *The Journal of Neuroscience* 20:6594–611. [DGe]