In press at *Behavioral and Brain Sciences.*

Building Machines That Learn and Think Like People

Brenden M. Lake,[1] Tomer D. Ullman,[2,4] Joshua B. Tenenbaum,[2,4] and Samuel J. Gershman[3,4]
[1]Center for Data Science, New York University
[2]Department of Brain and Cognitive Sciences, MIT
[3]Department of Psychology and Center for Brain Science, Harvard University
[4]Center for Brains Minds and Machines

Abstract

Recent progress in artificial intelligence (AI) 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 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 (a) build causal models of the world that support explanation and understanding, rather than merely solving pattern recognition problems; (b) ground learning in intuitive theories of physics and psychology, to support and enrich the knowledge that is learned; and (c) harness compositionality and learning-to-learn to rapidly acquire and generalize knowledge to new tasks and situations. We suggest concrete challenges and promising routes towards 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. These models have achieved remarkable gains in many domains spanning object recognition, speech recognition, and control (LeCun, Bengio, & Hinton, 2015; Schmidhuber, 2015). In object recognition, Krizhevsky, Sutskever, and Hinton (2012) trained a deep convolutional neural network (convnets; LeCun et al., 1989) that nearly halved the error rate of the previous state-of-the-art 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, Zhang, Ren, & Sun, 2015; 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, Mohamed, & Hinton, 2013; Weng, Yu, Watanabe, & Juang, 2014). Ideas from deep learning have also been applied to learning complex control problems. V. 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 these games (see also Guo, Singh, Lee, Lewis, & Wang, 2014; Schaul, Quan, Antonoglou, & Silver, 2016; Stadie, Levine, & Abbeel, 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 has 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 thus their ability to emulate human learning and human cognition.

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 AI researchers. Second, we articulate what we view as the essential ingredients for building such 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 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 towards 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). While 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, Rumelhart, & the PDP Research Group, 1986; Rumelhart, McClelland, & the PDP Research Group, 1986) – a 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. 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, etc. 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, while 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

[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).

Table 1: **Glossary**

Neural network: A network of simple neuron-like processing units that collectively perform 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 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 instance, in a classification task with examples X and class labels y, a generative model specifies the distribution of data given labels $P(X\|y)$, as well as a prior on labels $P(y)$, which can be used for sampling new examples or for classification by using Bayes' rule to compute $P(y\|X)$. A discriminative model specifies $P(y\|X)$ 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.

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, Duvenaud, Grosse, Tenenbaum, & Ghahramani, 2014), automated techniques for model building and selection (Grosse, Salakhutdinov, Freeman, & Tenenbaum, 2012), and probabilistic programming languages (e.g., Gelman, Lee, & Guo, 2015; Goodman, Mansinghka, Roy, Bonawitz, & Tenenbaum, 2008; Mansinghka, Selsam, & Perov, 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, but 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 paper is to propose a set of core ingredients for building more human-like learning and thinking machines. We will 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** (Section 4.1.1): Infants have primitive object concepts that allow them to track objects over time and allow them 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. While a task may be new, physics still works the same way. A second type of software present in early development is **intuitive psychology** (Section 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. While 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 (Section 4.2.2). Under 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 to 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 (Sections 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 (Section 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, and 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 will discuss ways in which these "model-free" methods can accelerate slow model-based inferences in perception and cognition (Section 4.3.1). 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, so quickly.

We will also discuss the integration of model-based and model-free methods in reinforcement learning (Section 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 will 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 is 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 over-simplified 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 & 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.*

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.*

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, McClelland, & the PDP Research Group, 1986). 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," share 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; 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 datasets, 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, McClelland, & the PDP Research Group, 1986) 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 paper.

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, programs, etc. 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 to 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., 2015; 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, Zaremba, Fergus, & Gureckis, 2015) and similarity ratings (Peterson, Abbott, & Griffiths, 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 (V. Mnih et al., 2015). 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 towards this goal?

3 Challenges for building more human-like machines

While 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, Salakhutdinov, & Tenenbaum, 2015) and learning to play the Atari game Frostbite (V. 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 machine learning algorithms. Hofstadter (1985) argued that the problem of recognizing characters in all 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 right, 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 ATM machine. 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 to 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 MNIST data set for digit recognition, which involves classifying images of digits into the categories '0'-'9' (LeCun, Bottou, Bengio, & Haffner, 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, Meier, & Schmidhuber, 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).

While 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 as well as learning more general classes of objects (Figure 1). People can learn to recognize a new handwritten character from a single example (Figure 1A-i), allowing them to discriminate between novel instances drawn by other people and similar looking non-instances (Lake, Salakhutdinov, & Tenenbaum, 2015; E. G. Miller, Matsakis, & Viola, 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 (Figure 1A-ii), parse a character into its most important parts and relations (Figure 1A-iii; Lake, Salakhutdinov, and Tenenbaum (2012)), and generate new characters given a small set of related characters (Figure 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, Salakhutdinov, & Tenenbaum, 2015), yet aspects of the full human cognitive ability remain out of reach. While both people and model represent characters as a sequence of pen strokes and relations, people have

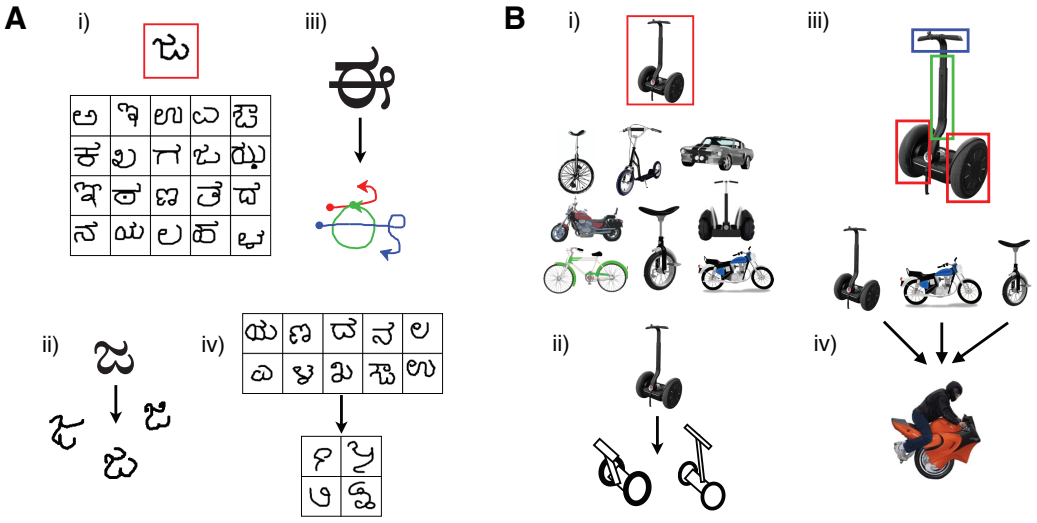

Figure 1: The characters challenge: human-level learning of a novel handwritten characters (A), with the same abilities also illustrated for a novel two-wheeled vehicle (B). A single example of a new visual concept (red box) can be enough information to support the (i) classification of new examples, (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, Salakhutdinov, and Tenenbaum (2015).

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. 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 (Figure 2), which was one of the control problems tackled by the DQN of V. 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, Naddaf, Veness, & Bowling, 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 (Figure 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

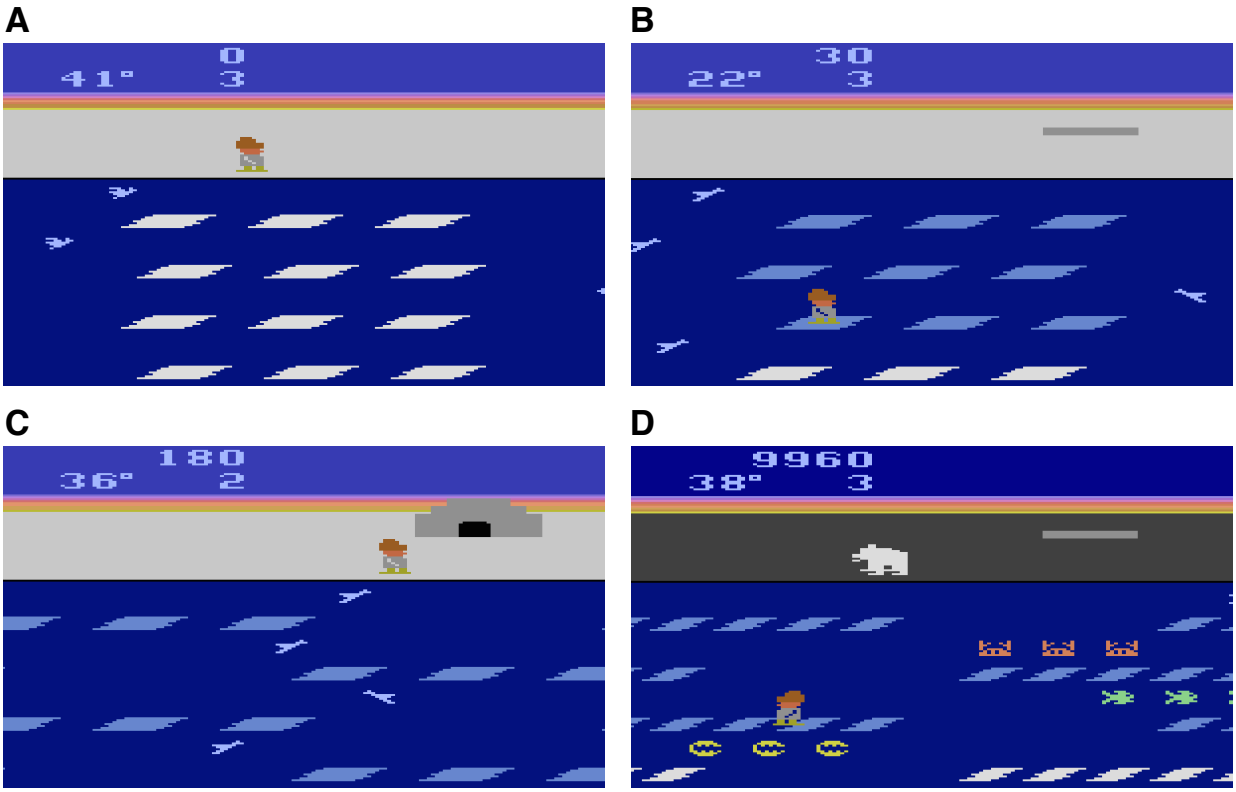

Figure 2: Screenshots of Frostbite, a 1983 video game designed for the Atari game console. A) The start of a level in Frostbite. The agent must construct an igloo by hopping between ice floes and avoiding obstacles such as birds. The floes are in constant motion (either left or right), making multi-step 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).

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 thus complete the level before time expires (Figure 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 V. 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 multi-task network (Parisotto, Ba, & Salakhutdinov, 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 V. Mnih et al. (2015), the DQN was compared with a professional gamer who received approximately two 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 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 8 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 have demonstrated superior performance (Schaul et al., 2016; Stadie et al., 2016; van Hasselt, Guez, & Silver, 2016; Wang et al., 2016), reaching 83% of the professional gamer's score by incorporating smarter experience replay (Schaul et al., 2016) and 96% by using smarter replay and more efficient parameter sharing (Wang et al., 2016) (see DQN+ and DQN++ in Fig. 3).[3] But they requires a lot of experience to reach this level: the learning curve provided in Schaul et al. (2016) shows performance is around 46% after 231 hours, 19% after 116 hours, and below 3.5% after just 2 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. While 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. While novice players may make some mistakes, such as inferring that fish are harmful rather than helpful, they can learn to play better than chance within a few minutes. If humans are able to first watch an expert playing for a few minutes, 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 two minutes, we found that we were able to reach scores comparable to or

[2]The time required to train the DQN (compute time) is not the same as the game (experience) time. Compute time can be longer.

[3]The reported scores use the "human starts" measure of test performance, designed to prevent networks from just memorizing long sequences of successful actions from a single starting point. Both faster learning (Blundell et al., 2016) and higher scores (Wang et al., 2016) have been reported using other metrics, but it is unclear how well the networks are generalizing with these alternative metrics.

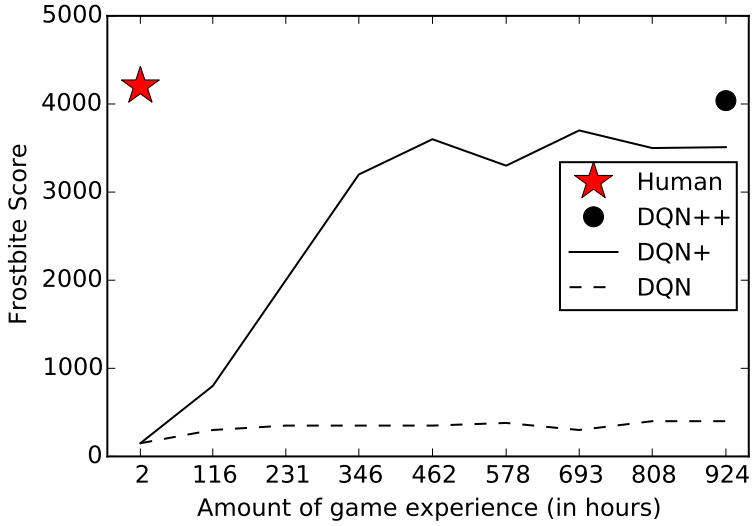

Figure 3: Comparing learning speed for people versus Deep Q-Networks (DQNs). Test performance on the Atari 2600 game "Frostbite" is plotted as a function of game experience (in hours at a frame rate of 60 fps), which does not include additional experience replay. Learning curves (if available) and scores are shown from different networks: DQN (V. Mnih et al., 2015), DQN+ (Schaul et al., 2016), and DQN++ (Wang et al., 2016). Random play achieves a score of 66.4. The "human starts" performance measure is used (van Hasselt et al., 2016).

better than the human expert reported in V. Mnih et al. (2015) after at most 15-20 minutes of total practice.[4]

There are other behavioral signatures that suggest fundamental differences in representation and learning between people and the DQN. For instance, 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 where 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. While any specific model is necessarily

[4]More precisely, the human expert in V. Mnih et al. (2015) scored an average of 4335 points across 30 game sessions of up to five minutes of play. In individual sessions lasting no longer than five 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 10-15 minutes, and 6710 after no more than 20 minutes. TDU and JBT each watched approximately two 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).

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:

- Get the lowest possible score.
- Get closest to 100, or 300, or 1000, or 3000, or any level, without going over.
- Beat your friend, who's playing next to you, but just barely, not by too much, so as not to embarrass them.
- Go as long as you can without dying.
- Die as quickly as you can.
- 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).
- Get to the furthest unexplored level without regard for your score.
- See if you can discover secret Easter eggs.
- Get as many fish as you can.
- Touch all the individual ice floes on screen once and only once.
- 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. While 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, Noelle, Braver, Cohen, & O'Reilly, 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 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. While these are hardly the only ingredients needed for human-like learning and thought (see our discussion of language in Section 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, they are 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. 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 & Meltzoff, 1999; Gweon, Tenenbaum, & Schulz, 2010; L. Schulz, 2012; 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, Goodman, & Schulz, 2011; Gweon et al., 2010; L. E. Schulz, Gopnik, & Glymour, 2007; Stahl & Feigenson, 2015; Tsividis, Gershman, Tenenbaum, & Schulz, 2013). We will 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 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, Gutheil, & Van de Walle, 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 through barriers, while solid objects cannot (Hespos, Ferry, & Rips, 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, Li, Ng, & Yuan, 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, Yildirim, Tenenbaum, & Battaglia, 2015; Battaglia, Hamrick, & Tenenbaum, 2013; Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2015; Sanborn, Mansinghka, & Griffiths, 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

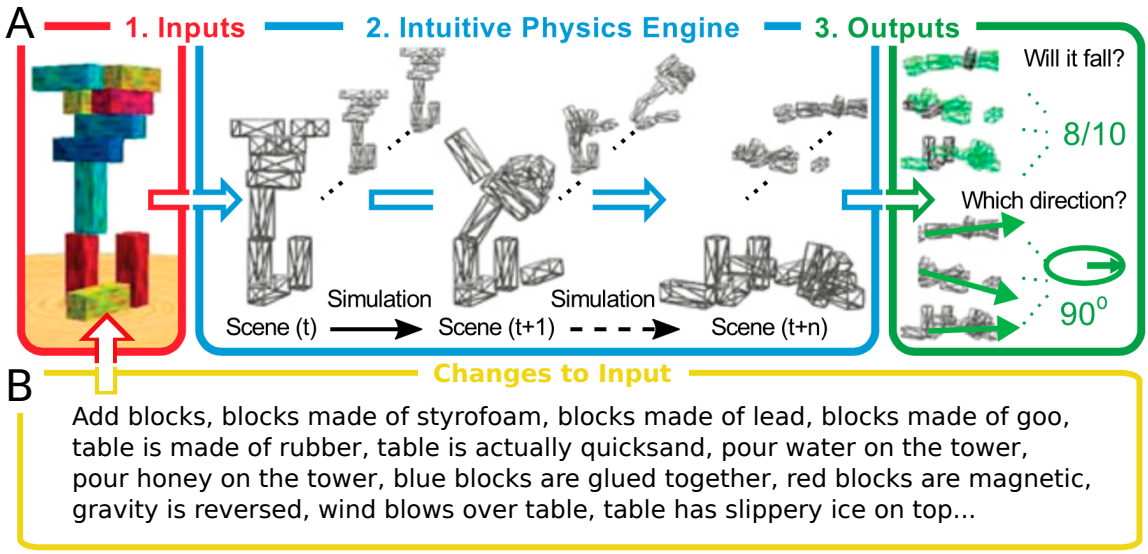

Figure 4: The intuitive physics-engine approach to scene understanding, illustrated through tower stability. (A) The engine takes in inputs through perception, language, memory and other faculties. It then constructs a physical scene with objects, physical properties and forces, simulates the scene's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

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 (Figure 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 are taken away, more blocks are added, or the table supporting the tower is 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 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 distance, velocity, and time in motion (Buckingham & Shultz, 2000), but 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, Gross, & Fergus, 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, Clune, Bengio, & Lipson, 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 (Figure 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 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 (Figure 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 towards 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, Slaughter, & Carey, 1998; Premack & Premack, 1997; Schlottmann, Ray, Mitchell, & Demetriou, 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 towards those goals subject to constraints (Csibra, 2008; Csibra, Biro, Koos, & Gergely, 2003; Spelke & Kinzler, 2007). These goals can be socially directed; at around three months of age, infants begin to discriminate anti-social agents that hurt or hinder others from neutral agents (Hamlin, 2013; Hamlin, Wynn, & Bloom, 2010), and they later distinguish between anti-social, neutral, and pro-social agents (Hamlin, Ullman, Tenenbaum, Goodman, & Baker, 2013; Hamlin, Wynn, & Bloom, 2007).

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, Cole, Watts, & White, 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 towards 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.'

While 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 towards A could then be seen as good (Hamlin, 2013). Or suppose something harmful was in the box which A didn't 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") models of Baker, Saxe, and Tenenbaum (2009) or the "naive utility calculus" models of Jara-Ettinger, Gweon, Tenenbaum, and Schulz (2015) (See also Jern and Kemp (2015) and Tauber and Steyvers (2011), and a related alternative based on predictive coding from Kilner, Friston, and Frith (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 (or 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 social-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 ex-

[5]While 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, Shultz, & Onishi, 2013), we are referring here to scenarios which require inferring goals, utilities, and relations.

tremely 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, Goodman, & Griffiths, 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 instance, 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 the Frostbite Bailey die because of the bird – in order 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 their goals. This agent can be useful in different ways under 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, while 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. While 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 (S. Ullman, Harari, & Dorfman, 2012).[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. While the origins of intuitive psychology is 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, Cooper, & Munro, 1982), back-propagation (Rumelhart, Hinton, & Williams, 1986), the wake-sleep algorithm (Hinton, Dayan, Frey, & Neal, 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 towards gradually matching the statistics of the model's internal patterns with the statistics of the input data.

[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.

In recent years, machine learning has found particular success using backpropagation and large data sets to solve difficult pattern recognition problems. While 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 in order 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, Smith, & Jones, 1988; E. M. Markman, 1989; Smith, Jones, Landau, Gershkoff-Stowe, & Samuelson, 2002; F. Xu & Tenenbaum, 2007, although see Horst and Samuelson 2008 regarding memory limitations). Children may only need to see a few examples of the concepts *hairbrush*, *pineapple* or *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 nine or ten 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, Bienenstock, & Doursat, 1992). Benchmark tasks such as the ImageNet data set for object recognition provides hundreds or thousands of examples per class (Krizhevsky et al., 2012; Russakovsky et al., 2015) – 1000 hairbrushes, 1000 pineapples, etc. In the context of learning new handwritten characters or learning to play Frostbite, the MNIST benchmark includes 6000 examples of each handwritten digit (LeCun et al., 1998), and the DQN of V. Mnih et al. (2015) played each Atari video game for approximately 924 hours of unique training experience (Figure 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, evolution, etc. 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 (A. B. Markman & Ross, 2003; Solomon, Medin, & Lynch, 1999). Beyond classification, concepts support prediction (Murphy & Ross, 1994; Rips, 1975), action (Barsalou, 1983), communication (A. B. 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 (Figure 1B-ii), parse the concept into its most important components (Figure 1B-iii), or even create a new complex concept through the combination of familiar concepts (Figure 1B-iv). Likewise, as discussed in the context of Frostbite, a learner who has acquired the basics of the game could flexibly apply their knowledge to an infinite set of Frostbite variants (Section 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 suggests 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, Joulin, & Baroni, 2016; Mitchell, Keller, & Kedar-cabelli, 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 utilize large data sets and little domain knowledge.

There has been recent work on other types of tasks including learning generative models of images (Denton, Chintala, Szlam, & Fergus, 2015; Gregor, Danihelka, Graves, Rezende, & Wierstra, 2015), caption generation (Karpathy & Fei-Fei, 2015; Vinyals, Toshev, Bengio, & Erhan, 2014; K. Xu et al., 2015), question answering (Sukhbaatar, Szlam, Weston, & Fergus, 2015; Weston, Chopra, & Bordes, 2015), and learning simple algorithms (Graves, Wayne, & Danihelka, 2014; Grefenstette, Hermann, Suleyman, & Blunsom, 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 for generalizing from small data sets (although see Rezende, Mohamed, Danihelka, Gregor, & Wierstra, 2016, for a deep learning approach to the Character Challenge). And it has been difficult to learn neural-network-style representations that effortlessly generalize to new tasks that they were not trained on (see Davis & Marcus, 2015; Marcus, 1998, 2001). What additional ingredients may be needed in order to rapidly learn more powerful and more general-purpose representations?

A relevant case study is from our own work on the Characters Challenge (Section 3.1; Lake, 2014; Lake, Salakhutdinov, & Tenenbaum, 2015). 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 – that is, structured procedures that generate new examples of a concept when executed (Figure 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 reuse of stochastic primitives that can combine in new ways to create new concepts.

Note that we are overloading the word "model" to refer to both 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

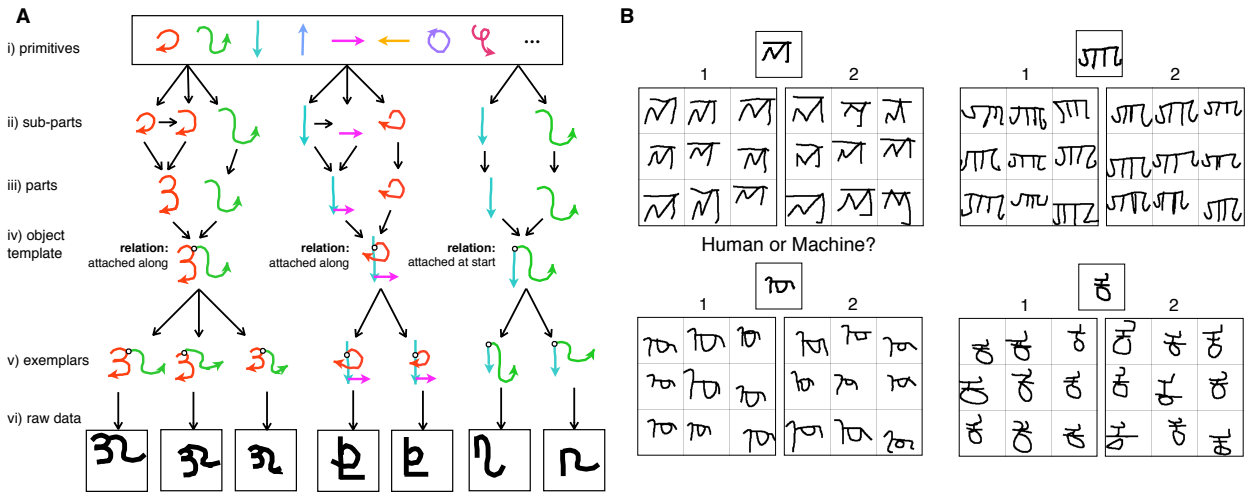

Figure 5: A causal, compositional model of handwritten characters. A) New types are generated compositionally by choosing primitive actions (color coded) from a library (i), combining these sub-parts (ii) to make parts (iii), and combining parts with relations to define simple programs (iv). These programs 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 9 people and the other is 9 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, Salakhutdinov, and Tenenbaum (2015).

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

Learning models of this form allows BPL to perform a challenging one-shot classification task at human level performance (Figure 1A-i) and to outperform current deep learning models such as convolutional networks (Koch, Zemel, & Salakhutdinov, 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., Figure 5B). These tasks include generating new examples (Figure 1A-ii and Figure 5B), parsing objects into their essential components (Figure 1A-iii), and generating new concepts in the style of a particular alphabet (Figure 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. While 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.

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

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 together 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, like a part hierarchy for describing complex objects or scenes (Bienenstock, Geman, & Potter, 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, etc. Parts can themselves be composed of sub-parts, forming a "partonomy" of part-whole relationships (G. A. Miller & Johnson-Laird, 1976; Tversky & Hemenway, 1984). In the novel vehicle example, the parts and relations can be shared and reused from existing related concepts, such as cars, scooters, motorcycles, and unicycles. Since 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. While 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, Salakhutdinov, and Tenenbaum (2015) 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 (Figure 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, Lee, Glass, & Tenenbaum, 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, etc. (Figure 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, Cohen, & Littman, 2008). Many repetitions of the same objects are present at different locations in the scene, and thus representing each as an identical instance of the same object with the

Figure 6: Perceiving scenes without intuitive physics, intuitive psychology, compositionality, and causality. Image captions are generated by a deep neural network (Karpathy & Fei-Fei, 2015) 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`.

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 (Figure 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 3D 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, Narasimhan, Saeedi, & Tenenbaum, 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 utilize causality are usually generative (as opposed to discriminative; see Glossary in Table 1), but not every generative model is also causal. While 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. While generative neural networks such as Deep Belief Networks (Hinton, Osindero, & Teh, 2006) or variational auto-encoders (Gregor, Besse, Rezende, Danihelka, & Wierstra, 2016; Kingma, Rezende, Mohamed, & Welling, 2014) may generate compelling handwritten digits, they mark one end of the "causality spectrum," since 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 Bayesian Program Learning (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 its 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 instance, Liberman, Cooper, Shankweiler, and Studdert-Kennedy (1967) argued that the richness of speech perception is best explained by inverting the production plan, at the level of vocal tract movements, in order 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 (Figure 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, Salakhutdinov, & Tenenbaum, 2015) (Figure 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 they learn and generalize. 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 while 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 due to 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 while 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 together, can lead to revealing errors. Consider image captions generated by a deep neural network (Figure 6; Karpathy & Fei-Fei, 2015). 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 towards deep neural networks and related approaches that learn causal models. Lopez-Paz, Muandet, Scholköpf, and Tolstikhin (2015) introduced a discriminative, data-driven framework for distinguishing the direction of causality from examples. While 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. While 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). While the model demonstrates an impressive ability to make plausible generalizations that go beyond the training examples, 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, Salakhutdinov, and Tenenbaum (2015) (Figure 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 to 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 utilize learning-to-learn, described next), potentially leading to a more computationally efficient and neurally grounded variant of the BPL model of handwritten characters (Figure 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 (Figure 2). Deep neural networks could play a role in two ways: serving as a bottom-up proposer to make probabilistic inference more tractable in a structured generative model (Section 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, Chater, Kemp, Perfors, & Tenenbaum, 2010; Tenenbaum, Kemp, Griffiths, & Goodman, 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", "multi-task learning" or "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 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, Carlin, Stern, & Rubin, 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, Tenenbaum, & Torralba, 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, Bottou, Scholköpf, & Vapnik, 2016; Rusu et al., 2016; Salakhutdinov, Torralba, & Tenenbaum, 2011; Srivastava & Salakhutdinov, 2013; Torralba, Murphy, & Freeman, 2007; Zeiler & Fergus, 2014). Neural networks can also learn-to-learn by optimizing hyperparameters, including the form of their weight update rule (Andrychowicz et al., 2016), over a set of related tasks.

While transfer learning and multi-task 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 Section 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, Salakhutdinov, and Tenenbaum (2015), 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, and 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 (Figure 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, etc.) 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 might 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 this 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 (Figure 2D), an attentive learner is sure to avoid it. Depending on the level, ice floes can be spaced far apart (Figure 2A-C) or close together (Figure 2D), suggesting the agent may be able to cross some gaps but not others. In this way,

[8]Humans typically have direct experience with 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. The best neural network classifiers (deep convolutional networks) have error rates approximately five times higher than 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, Salakhutdinov, & Tenenbaum, 2015). 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.

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) presents 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 four 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 four or five million frames of learning to reach might now be reached after one or two 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 do. A deep learning system trained on many video games may not, by itself, be enough to learn new games as quickly as people do. Yet if such a system aims to learn compositionally structured causal models of a 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 towards resolving the conflict between fast inference and structured representations, including Helmholtz-machine-style approximate inference in generative models (Dayan, Hinton, Neal, & Zemel, 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, Salakhutdinov, & Tenenbaum, 2015; 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, while 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, Horvitz, & Tenenbaum, 2015; Vul, Goodman, Griffiths, & Tenenbaum, 2014).

Popular algorithms for approximate inference in probabilistic machine learning have been proposed as psychological models (see Griffiths, Vul, & Sanborn, 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, Denison, Griffiths, & Gopnik, 2014; Gershman, Vul, & Tenenbaum, 2012; T. D. Ullman, Goodman, & Tenenbaum, 2012; 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, Reali, & Griffiths, 2009) and perceptual multistability (Gershman et al., 2012; Moreno-Bote, Knill, & Pouget, 2011). Moreover, we are beginning to understand how such methods could be implemented in neural circuits (Buesing, Bill, Nessler, & Maass, 2011; Huang & Rao, 2014; Pecevski, Buesing, & Maass, 2011).[9]

While 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 (T. D. Ullman et al., 2012). Discovering new theories can be slow and arduous, as testified by the long timescale 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 (L. Schulz, 2012), problem-solving (Sternberg & Davidson, 1995), and epoch-making discoveries in scientific research (Langley, Bradshaw, Simon, & Zytkow, 1987). Discovering new theories can also happen much more quickly – A person learning the rules of Frostbite will probably undergo a loosely ordered sequence of "Aha!" moments: they will learn that jumping on ice floes causes them to change color, changing the color of ice floes causes an igloo to be constructed piece-by-piece, that birds make you lose points, that fish make you gain points, that you can change the direction of ice floe 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 happens quickly, it is possible that people employ inductive biases not only to evaluate hypotheses, but also to guide hypothesis selection. L. Schulz (2012) 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

[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.

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, Tenenbaum, and Schulz (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, Sheskin, & Schulz, 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, Tarlow, Kohli, & Winn, 2014; Heess, Tarlow, & Winn, 2013; A. Mnih & Gregor, 2014; Stuhlmüller, Taylor, & Goodman, 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; A. Mnih & Gregor, 2014; Rezende, Mohamed, & Wierstra, 2014) or nearest-neighbor density estimation (Kulkarni, Kohli, Tenenbaum, & Mansinghka, 2015; Stuhlmüller et al., 2013). One implication of amortization is that solutions to different problems will become correlated due to 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, Whitney, Kohli, & Tenenbaum, 2015; Yildirim, Kulkarni, Freiwald, & Te, 2015). Another avenue for potential integration is through differentiable programming (Dalrmple, 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 V. 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, Dayan, & Montague, 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, Niv, & Dayan, 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, Dezfouli, & Piray, 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, Markman, & Otto, 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, Kurth-Nelson, Lübbert, Guitart-Masip, & Dolan, 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; Deci & Ryan, 1975; Harlow, 1950). While 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 multi-task 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 discussing the arguments in this paper with colleagues, three lines of questioning or critiques have come up frequently. We think it is helpful to address these points directly, to maximize the potential for moving forward together.

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 multi-task 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., 2013; Luong, Le, Sutskever, Vinyals, & Kaiser, 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 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 paper. 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 the intuitive physics and psychology.

Finally, we recognize that some researchers still hold out hope that if only they can just get big enough training datasets, 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 provides humans with. 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.

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" (Section 4.3) are prime exemplars. Neuroscience can also in principle impose constraints on cognitive accounts, both at the cellular and systems level. 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 thus 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 backwards along the axon, which does not fit with realistic models of neuronal function (although recent models circumvent this problem in various ways Liao, Leibo, & Poggio, 2015; Lillicrap, Cownden, Tweed, & Akerman, 2014; Scellier & Bengio, 2016). This has not prevented backpropagation 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 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.

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

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, Thompson, Moyer, & Disterhoft, 1997). Learned behavior is rapidly reacquired after extinction (Bouton, 2004), whereas no such facilitation is observed for LTP (de 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 level, 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.

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 lag strikingly. Certainly one could argue that language should be included on any short list of key ingredients in human intelligence: for instance, Mikolov et al. (2016) featured language prominently in their recent paper sketching challenge problems and a road map for AI. Moreover, while natural language processing is an active area of research in deep learning (e.g., Bahdanau, Cho, & Bengio, 2015; Mikolov, Sutskever, & Chen, 2013; K. 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 ourselves 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 do 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; F. 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 gives 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, Chomsky, & Fitch, 2002)? Is it the ability to reuse symbols by name (Deacon, 1998)? Is it the ability to understand others intentionally and build shared intentionality (Bloom, 2000; Frank, Goodman, & Tenenbaum, 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 the ways that acquiring language extends and enriches the ingredients of cognition we focus on in this article. The intuitive physics and psychology of infants is 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 achieved and what it has not. While 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 and the current best algorithms in AI and machine learning, people learn from less 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 (Figure 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 paper. 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; V. Mnih, Heess, Graves, & Kavukcuoglu, 2014; K. Xu et al., 2015), augmented working memory (Graves et al., 2014, 2016; Grefenstette et al., 2015; Sukhbaatar et al., 2015; Weston et al., 2015), and experience replay (McClelland, McNaughton, & O'Reilly, 1995; V. Mnih et al., 2015). These ingredients are lower-level than the key cognitive ingredients discussed in this paper, 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 (V. Mnih et al., 2014), and image caption generation (K. 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 instance, 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" (K. 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 towards 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 (Section 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 & de Freitas, 2016; Sukhbaatar et al., 2015; Weston et al., 2015). These developments are also part of a broader trend towards "differentiable programming," the incorporation of classic data structures such a random access memory, stacks, and queues, into gradient-based learning systems (Dalrmple, 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 maintains 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 & de 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.

While this new generation of neural networks has yet to tackle the types of challenge problems introduced in this paper, 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 paper – objects, forces, agents, causality, and compositionality – help to explain important facets of human learning and thinking, yet they also bring challenges for performing efficient inference (Section 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 for 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 Figure 7). Just as important, however, are the new questions and directions it opens up for the long-term project of building genuinely human-like AI.

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 before 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 (Silver, 2016). 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 at all. 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.

AI 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 (Sections 3.1 and 3.2), humans can learn the basics of the game quickly 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

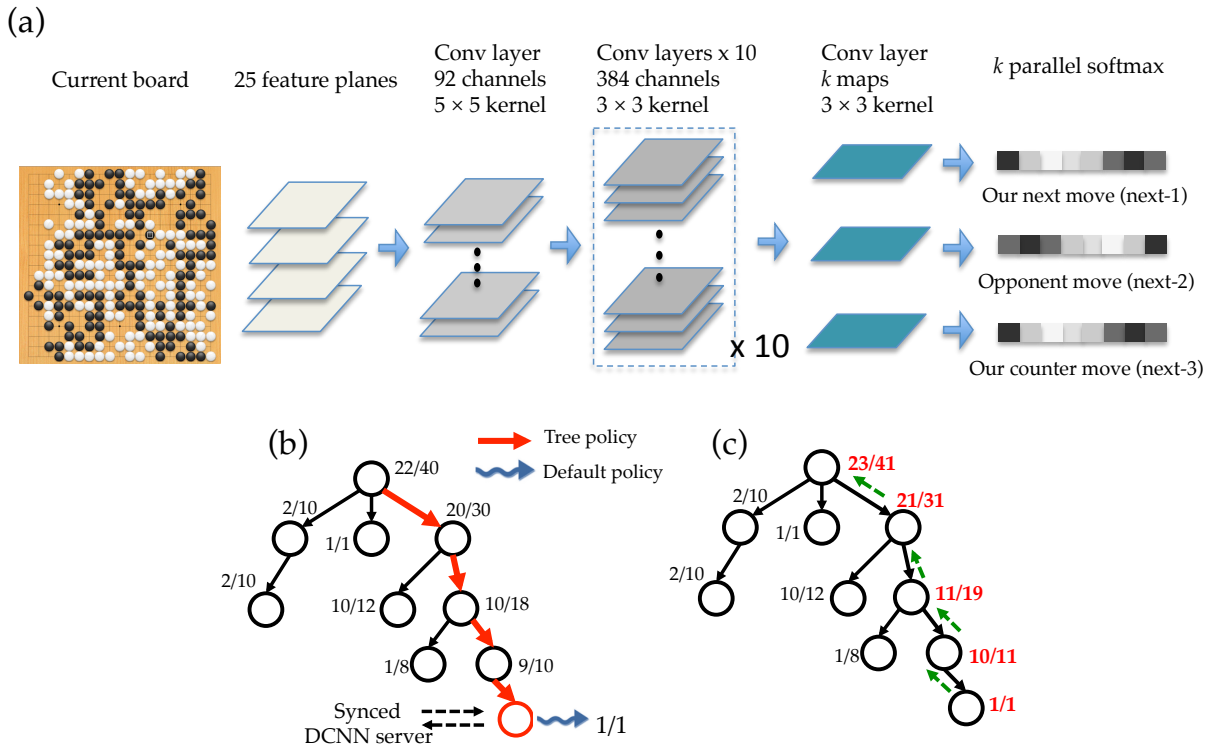

Figure 7: An AI system for playing Go combining a deep convolutional network (convnet) and model-based search through Monte-Carlo Tree Search (MCTS). (A) The convnet on its own can be used to predict the next k moves given the current board. (B) A search tree with the current board state as its root 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.

"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 their 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? While techniques for handling variable sized inputs in convnets may help for 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 do. 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 they are, 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 paper 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 capturing better 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 paper argues for – as well as building on all of the progress to date that AlphaGo represents.

6.2 Future applications to practical AI problems

In this paper, 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 towards scene understanding, as evidenced by a flurry of recent work focused on generating natural language captions for images (Karpathy & Fei-Fei, 2015; Vinyals et al., 2014; K. 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 (Figure 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. In order 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 towards the other objects and agents they believe are present.

2. *Autonomous agents and intelligent devices.* Robots and personal assistants (such as cellphones) 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" or "Kofi Annan"), new gestures (a secret handshake or a "fist bump"), and new activities, and a human-like system would be able to learn to both recognize and produce new instances from a small number of examples. Like 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., Do they want to change lanes? Pass another car? Are they swerving to avoid a hidden hazard? Are they 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 is little relevant training data (e.g. navigating unusual construction zones, natural disasters, etc.)

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. While 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: While compositionality and learning-to-learn can provide the parts for new ideas, causality provides the glue that gives them coherence and purpose.

6.3 Towards 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 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 little training data as people need, and also to evaluate models on a range of human-like generalizations beyond the one task the model was trained on. We hope that the ingredients outlined in this article will prove useful for working towards 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 manuscript. Tom Schaul was 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 NSF STC award CCF-1231216, and the Moore-Sloan Data Science Environment at NYU.

References

Andrychowicz, M., Denil, M., Gomez, S., Hoffman, M. W., Pfau, D., Schaul, T., & de Freitas, N. (2016). Learning to learn by gradient descent by gradient descent. *arXiv preprint*.

Anselmi, F., Leibo, J. Z., Rosasco, L., Mutch, J., Tacchetti, A., & Poggio, T. (2016). Unsupervised learning of invariant representations. *Theoretical Computer Science*.

Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural Machine Translation by Jointly Learning to Align and Translate. In *International Conference on Learning Representations (ICLR)*. Retrieved from `http://arxiv.org/abs/1409.0473v3`

Baillargeon, R. (2004). Infants' physical world. *Current Directions in Psychological Science*, *13*, 89–94. doi: 10.1111/j.0963-7214.2004.00281.x

Baillargeon, R., Li, J., Ng, W., & Yuan, S. (2009). An account of infants physical reasoning. *Learning and the infant mind*, 66–116.

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, *113*(3), 329–349.

Barsalou, L. W. (1983). Ad hoc categories. *Memory & Cognition*, *11*(3), 211–227.

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.

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.*

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*, *110*(45), 18327–18332.

Baudiš, P., & Gailly, J.-l. (2012). Pachi: State of the art open source go program. In *Advances in computer games* (pp. 24–38). Springer.

Baxter, J. (2000). A model of inductive bias learning. *Journal of Artificial Intelligence Research*, *12*, 149–198.

Bayer, H. M., & Glimcher, P. W. (2005). Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron*, *47*, 129–141.

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*, 253–279.

Berlyne, D. E. (1966). Curiosity and exploration. *Science*, *153*, 25–33.

Berthiaume, V. G., Shultz, T. R., & Onishi, K. H. (2013). A constructivist connectionist model of transitions on false-belief tasks. *Cognition*, *126*(3), 441–458.

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

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

Bi, G.-q., & Poo, M.-m. (2001). Synaptic modification by correlated activity: Hebb's postulate revisited. *Annual Review of Neuroscience*, *24*, 139–166.

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review*, *94*(2), 115–147.

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.

Bienenstock, E., Geman, S., & Potter, D. (1997). Compositionality, MDL Priors, and Object Recognition. In *Advances in Neural Information Processing Systems.*

Bloom, P. (2000). *How Children Learn the Meanings of Words.* Cambridge, MA: MIT Press.

Blundell, C., Uria, B., Pritzel, A., Li, Y., Ruderman, A., Leibo, J. Z., ... Hassabis, D. (2016).

Model-Free Episodic Control. *arXiv preprint*.

Bobrow, D. G., & Winograd, T. (1977). An overview of KRL, a knowledge representation language. *Cognitive Science*, *1*, 3–46.

Boden, M. A. (1998). Creativity and artificial intelligence. *Artificial Intelligence*, *103*(I 998), 347–356.

Boden, M. A. (2006). *Mind as machine: A history of cognitive science*. Oxford University Press.

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.

Bottou, L. (2014). From machine learning to machine reasoning. *Machine learning*, *94*(2), 133–149.

Bouton, M. E. (2004). Context and behavioral processes in extinction. *Learning & Memory*, *11*, 485–494.

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), 305–345.

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.

Carey, S. (1978). The Child as Word Learner. In J. Bresnan, G. Miller, & M. Halle (Eds.), *Linguistic theory and psychological reality* (pp. 264–293).

Carey, S. (2004). Bootstrapping and the origin of concepts. *Daedalus*, *133*(1), 59–68.

Carey, S. (2009). *The Origin of Concepts*. New York, New York, USA: Oxford University Press.

Carey, S., & Bartlett, E. (1978). Acquiring a single new word. *Papers and Reports on Child Language Development*, *15*, 17–29.

Chouard, T. (2016, March). *The go files: AI computer wraps up 4-1 victory against human champion.* ([Online; posted 15-March-2016])

Ciresan, D., Meier, U., & Schmidhuber, J. (2012). Multi-column Deep Neural Networks for Image Classification. In *Computer Vision and Pattern Recognition (CVPR)* (pp. 3642–3649).

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.

Cook, C., Goodman, N. D., & Schulz, L. E. (2011). Where science starts: spontaneous experiments in preschoolers' exploratory play. *Cognition*, *120*(3), 341–9.

Crick, F. (1989). The recent excitement about neural networks. *Nature*, *337*, 129–132.

Csibra, G. (2008). Goal attribution to inanimate agents by 6.5-month-old infants. *Cognition*, *107*, 705–717.

Csibra, G., Biro, S., Koos, O., & Gergely, G. (2003). One-year-old infants use teleological representations of actions productively. *Cognitive Science*, *27*, 111–133.

Dalrmple, D. (2016). *Differentiable Programming.* Retrieved from `https://www.edge.org/response-detail/26794`

Davis, E., & Marcus, G. (2015). Commonsense Reasoning and Commonsense Knowledge in Artificial Intelligence. *Communications of the ACM*, *58*(9), 92–103.

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, *8*, 1704–1711.

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

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

Deci, E. L., & Ryan, R. M. (1975). *Intrinsic motivation.* Wiley Online Library.

de Jonge, M., & Racine, R. J. (1985). The effects of repeated induction of long-term potentiation in the dentate gyrus. *Brain Research*, *328*, 181–185.

Denton, E., Chintala, S., Szlam, A., & Fergus, R. (2015). Deep Generative Image Models using a Laplacian Pyramid of Adversarial Networks. In *Advances in Neural Information Processing Systems 29.* Retrieved from `http://arxiv.org/abs/1506.05751`

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)* (pp. 240–247).

Dolan, R. J., & Dayan, P. (2013). Goals and habits in the brain. *Neuron*, *80*, 312–325.

Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., & Darrell, T. (2013). Decaf: A deep convolutional activation feature for generic visual recognition. *arXiv preprint arXiv:1310.1531*.

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.

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), 751–776.

Eden, M. (1962). Handwriting and Pattern Recognition. *IRE Transactions on Information Theory*, 160–166.

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), 1202–1205.

Elman, J. L. (2005). Connectionist models of cognitive development: Where next? *Trends in Cognitive Sciences*, *9*(3), 111–117.

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

Eslami, S. M. A., Heess, N., Weber, T., Tassa, Y., Kavukcuoglu, K., & Hinton, G. E. (2016). Attend, infer, repeat: Fast scene understanding with generative models. *arXiv preprint arXiv:1603.08575*.

Eslami, S. M. A., Tarlow, D., Kohli, P., & Winn, J. (2014). Just-in-time learning for fast and flexible inference. In *Advances in Neural Information Processing Systems* (pp. 154–162).

Fodor, J. A. (1975). *The Language of Thought.* Harvard University Press.

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *28*, 3–71.

Frank, M. C., Goodman, N. D., & Tenenbaum, J. B. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science*, *20*, 578–585.

Freyd, J. (1983). Representing the dynamics of a static form. *Memory and Cognition*, *11*(4), 342–346.

Freyd, J. (1987). Dynamic Mental Representations. *Psychological Review*, *94*(4), 427–438.

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.

Gallistel, C., & Matzel, L. D. (2013). The neuroscience of learning: beyond the Hebbian synapse.

Annual Review of Psychology, *64*, 169–200.

Gelly, S., & Silver, D. (2008). Achieving master level play in 9 x 9 computer go..

Gelly, S., & Silver, D. (2011). Monte-carlo tree search and rapid action value estimation in computer go. *Artificial Intelligence*, *175*(11), 1856–1875.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian Data Analysis*. Chapman and Hall/CRC.

Gelman, A., Lee, D., & Guo, J. (2015). Stan a probabilistic programming language for Bayesian inference and optimization. *Journal of Educational and Behavioral Statistics*, *40*, 530–543.

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

Gershman, S. J., & Goodman, N. D. (2014). Amortized inference in probabilistic reasoning. In *Proceedings of the 36th Annual Conference of the Cognitive Science Society*.

Gershman, S. J., Horvitz, E. J., & Tenenbaum, J. B. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, *349*, 273–278.

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*, 182–194.

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

Gerstenberg, T., Goodman, N. D., Lagnado, D. a., & Tenenbaum, J. B. (2015). How, whether, why: Causal judgments as counterfactual contrasts. *Proceedings of the 37th Annual Conference of the Cognitive Science Society*.

Ghahramani, Z. (2015). Probabilistic machine learning and artificial intelligence. *Nature*, *521*, 452–459.

Goodman, N. D., Mansinghka, V. K., Roy, D. M., Bonawitz, K., & Tenenbaum, J. B. (2008). Church: A language for generative models. *Uncertainty in Artificial Intelligence*.

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.

Gopnik, A., & Meltzoff, A. N. (1999). Words, Thoughts, and Theories. *Mind: A Quarterly Review of Philosophy*, *108*, 0.

Graves, A. (2014). Generating sequences with recurrent neural networks. *arXiv preprint*. Retrieved from `http://arxiv.org/abs/1308.0850`

Graves, A., Mohamed, A.-r., & Hinton, G. (2013). Speech recognition with deep recurrent neural networks. In *Acoustics, speech and signal processing (icassp), 2013 ieee international conference on* (pp. 6645–6649).

Graves, A., Wayne, G., & Danihelka, I. (2014). Neural Turing Machines. *arXiv preprint*. Retrieved from `http://arxiv.org/abs/1410.5401v1`

Graves, A., Wayne, G., Reynolds, M., Harley, T., Danihelka, I., Grabska-Barwińska, A., ... Hassabis, D. (2016). Hybrid computing using a neural network with dynamic external memory. *Nature*.

Grefenstette, E., Hermann, K. M., Suleyman, M., & Blunsom, P. (2015). Learning to Transduce with Unbounded Memory. In *Advances in Neural Information Processing Systems*.

Gregor, K., Besse, F., Rezende, D. J., Danihelka, I., & Wierstra, D. (2016). Towards Conceptual Compression. *arXiv preprint*. Retrieved from `http://arxiv.org/abs/1604.08772`

Gregor, K., Danihelka, I., Graves, A., Rezende, D. J., & Wierstra, D. (2015). DRAW: A Recurrent Neural Network For Image Generation. In *International Conference on Machine Learning (ICML)*.

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.

Griffiths, T. L., Vul, E., & Sanborn, A. N. (2012). Bridging levels of analysis for probabilistic models of cognition. *Current Directions in Psychological Science*, *21*, 263–268.

Grossberg, S. (1976). Adaptive pattern classification and universal recoding: I. parallel development and coding of neural feature detectors. *Biological Cybernetics*, *23*, 121–134.

Grosse, R., Salakhutdinov, R., Freeman, W. T., & Tenenbaum, J. B. (2012). Exploiting compositionality to explore a large space of model structures. In *Uncertainty in Artificial Intelligence*.

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* (pp. 3338–3346).

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*, *107*, 9066–9071. doi: 10.1073/pnas.1003095107

Halle, M., & Stevens, K. (1962). Speech Recognition: A Model and a Program for Research. *IRE Transactions on Information Theory*, *8*(2), 155–159.

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*, 186–193. doi: 10.1177/0963721412470687

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*, 209–226. doi: 10.1111/desc.12017

Hamlin, K. J., Wynn, K., & Bloom, P. (2007). Social evaluation by preverbal infants. *Nature*, *450*, 557–560.

Hamlin, K. J., Wynn, K., & Bloom, P. (2010). Three-month-olds show a negativity bias in their social evaluations. *Developmental Science*, *13*, 923–929. doi: 10.1111/j.1467-7687.2010.00951.x

Harlow, H. F. (1949). The formation of learning sets. *Psychological Review*, *56*(1), 51–65.

Harlow, H. F. (1950). Learning and satiation of response in intrinsically motivated complex puzzle performance by monkeys. *Journal of Comparative and Physiological Psychology*, *43*, 289–294.

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*, 1569–1579.

Hayes-Roth, B., & Hayes-Roth, F. (1979). A cognitive model of planning. *Cognitive Science*, *3*, 275–310.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep Residual Learning for Image Recognition. *arXiv preprint*. Retrieved from `http://arxiv.org/abs/1512.03385`

Hebb, D. O. (1949). *The organization of behavior*. Wiley.

Heess, N., Tarlow, D., & Winn, J. (2013). Learning to pass expectation propagation messages. In *Advances in Neural Information Processing Systems* (pp. 3219–3227).

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, 304–316.

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), 603–611.

Hinton, G. E. (2002). Training products of experts by minimizing contrastive divergence. *Neural Computation*, *14*(8), 1771–800.

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.

Hinton, G. E., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-r., Jaitly, N., ... Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal Processing Magazine*, *29*, 82–97.

Hinton, G. E., Osindero, S., & Teh, Y. W. (2006). A fast learning algorithm for deep belief nets. *Neural Computation*, *18*, 1527–1554.

Hoffman, D. D., & Richards, W. A. (1984). Parts of recognition. *Cognition*, *18*, 65–96.

Hofstadter, D. R. (1985). *Metamagical themas: Questing for the essence of mind and pattern.* New York: Basic Books.

Horst, J. S., & Samuelson, L. K. (2008). Fast Mapping but Poor Retention by 24-Month-Old Infants. *Infancy*, *13*(2), 128–157.

Huang, Y., & Rao, R. P. (2014). Neurons as Monte Carlo samplers: Bayesian? inference and learning in spiking networks. In *Advances in neural information processing systems* (pp. 1943–1951).

Hummel, J. E., & Biederman, I. (1992). Dynamic binding in a neural network for shape recognition. *Psychological Review*, *99*(3), 480–517.

Jackendoff, R. (2003). *Foundations of Language.* Oxford University Press.

Jara-Ettinger, J., Gweon, H., Tenenbaum, J. B., & Schulz, L. E. (2015). Childrens understanding of the costs and rewards underlying rational action. *Cognition*, *140*, 14–23.

Jern, A., & Kemp, C. (2013). A probabilistic account of exemplar and category generation. *Cognitive Psychology*, *66*(1), 85–125.

Jern, A., & Kemp, C. (2015). A decision network account of reasoning about other peoples choices. *Cognition*, *142*, 12–38.

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*, 233–238. doi: 10.1111/1467 -7687.00036

Juang, B. H., & Rabiner, L. R. (1990). Hidden Markov models for speech recognition. *Technometric*, *33*(3), 251–272.

Karpathy, A., & Fei-Fei, L. (2015). Deep Visual-Semantic Alignments for Generating Image Desscriptions. In *Computer Vision and Pattern Recognition (CVPR).*

Kemp, C. (2007). *The acquisition of inductive constraints.* Unpublished doctoral dissertation, MIT.

Keramati, M., Dezfouli, A., & Piray, P. (2011). Speed/accuracy trade-off between the habitual and the goal-directed processes. *PLoS Computational Biology*, *7*, e1002055.

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. *PLoS Computational Biology*, *10*(11), e1003915.

Kilner, J. M., Friston, K. J., & Frith, C. D. (2007). Predictive coding: An account of the mirror neuron system. *Cognitive Processing*, *8*(3), 159–166.

Kingma, D. P., Rezende, D. J., Mohamed, S., & Welling, M. (2014). Semi-supervised Learning

with Deep Generative Models. In *Neural Information Processing Systems (NIPS).*

Koch, G., Zemel, R. S., & Salakhutdinov, R. (2015). Siamese neural networks for one-shot image recognition. In *ICML Deep Learning Workshop.*

Kodratoff, Y., & Michalski, R. S. (2014). *Machine learning: An artificial intelligence approach* (Vol. 3). Morgan Kaufmann.

Koza, J. R. (1992). *Genetic programming: on the programming of computers by means of natural selection* (Vol. 1). MIT press.

Kriegeskorte, N. (2015). Deep Neural Networks: A New Framework for Modeling Biological Vision and Brain Information Processing. *Annural Review of Vision Science*, *1*, 417–446.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25* (pp. 1097–1105).

Kulkarni, T. D., Kohli, P., Tenenbaum, J. B., & Mansinghka, V. (2015). Picture: A probabilistic programming language for scene perception. In *Computer Vision and Pattern Recognition (CVPR).*

Kulkarni, T. D., Narasimhan, K. R., Saeedi, A., & Tenenbaum, J. B. (2016). Hierarchical Deep Reinforcement Learning: Integrating Temporal Abstraction and Intrinsic Motivation. *arXiv preprint*.

Kulkarni, T. D., Whitney, W., Kohli, P., & Tenenbaum, J. B. (2015). Deep Convolutional Inverse Graphics Network. In *Computer Vision and Pattern Recognition (CVPR).*

Lake, B. M. (2014). *Towards more human-like concept learning in machines: Compositionality, causality, and learning-to-learn*. Unpublished doctoral dissertation, MIT.

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* (pp. 803–808).

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.*

Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science*, *350*(6266), 1332–1338.

Lake, B. M., Zaremba, W., Fergus, R., & Gureckis, T. M. (2015). Deep Neural Networks Predict Category Typicality Ratings for Images. In *Proceedings of the 37th Annual Conference of the Cognitive Science Society.*

Landau, B., Smith, L. B., & Jones, S. S. (1988). The importance of shape in early lexical learning. *Cognitive Development*, *3*(3), 299–321.

Langley, P., Bradshaw, G., Simon, H. A., & Zytkow, J. M. (1987). *Scientific discovery: Computational explorations of the creative processes*. MIT press.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*, 436–444.

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*, 541–551.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2323.

Lerer, A., Gross, S., & Fergus, R. (2016). Learning Physical Intuition of Block Towers by Example. *arXiv preprint*. Retrieved from `http://arxiv.org/abs/1603.01312`

Levy, R. P., Reali, F., & Griffiths, T. L. (2009). Modeling the effects of memory on human online sentence processing with particle filters. In *Advances in Neural Information Processing Systems* (pp. 937–944).

Liao, Q., Leibo, J. Z., & Poggio, T. (2015). How important is weight symmetry in backpropagation? *arXiv preprint arXiv:1510.05067*.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, *74*(6), 431–461.

Lillicrap, T. P., Cownden, D., Tweed, D. B., & Akerman, C. J. (2014). Random feedback weights support learning in deep neural networks. *arXiv preprint arXiv:1411.0247*.

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* (Vol. 2, pp. 1242–1250).

Lombrozo, T. (2009). Explanation and categorization: How "why?" informs "what?". *Cognition*, *110*(2), 248–53.

Lopez-Paz, D., Bottou, L., Scholköpf, B., & Vapnik, V. (2016). Unifying distillation and privileged information. In *International Conference on Learning Representations (ICLR)*.

Lopez-Paz, D., Muandet, K., Scholköpf, B., & Tolstikhin, I. (2015). Towards a Learning Theory of Cause-Effect Inference. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*.

Luong, M.-T., Le, Q. V., Sutskever, I., Vinyals, O., & Kaiser, L. (2015). Multi-task sequence to sequence learning. *arXiv preprint arXiv:1511.06114*.

Lupyan, G., & Bergen, B. (2016). How Language Programs the Mind. *Topics in Cognitive Science*, *8*(2), 408–424. Retrieved from `http://doi.wiley.com/10.1111/tops.12155`

Lupyan, G., & Clark, A. (2015). Words and the world: Predictive coding and the language-perception-cognition interface. *Current Directions in Psychological Science*, *24*(4), 279–284.

Macindoe, O. (2013). *Sidekick agents for sequential planning problems*. Unpublished doctoral dissertation, Massachusetts Institute of Technology.

Magid, R. W., Sheskin, M., & Schulz, L. E. (2015). Imagination and the generation of new ideas. *Cognitive Development*, *34*, 99–110.

Mansinghka, V., Selsam, D., & Perov, Y. (2014). Venture: A higher-order probabilistic programming platform with programmable inference. *arXiv preprint arXiv:1404.0099*.

Marcus, G. (1998). Rethinking Eliminative Connectionism. *Cognitive Psychology*, *282*(37), 243–282.

Marcus, G. (2001). *The algebraic mind: Integrating connectionism and cognitive science*. MIT press.

Markman, A. B., & Makin, V. S. (1998). Referential communication and category acquisition. *Journal of Experimental Psychology: General*, *127*(4), 331–54.

Markman, A. B., & Ross, B. H. (2003). Category use and category learning. *Psychological Bulletin*, *129*(4), 592–613.

Markman, E. M. (1989). *Categorization and Naming in Children*. Cambridge, MA: MIT Press.

Marr, D. C. (1982). *Vision*. San Francisco, CA: W.H. Freeman and Company.

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*, *200*(1140), 269–94.

McClelland, J. L. (1988). *Parallel distributed processing: Implications for cognition and development*

(Tech. Rep.). DTIC Document.

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.

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.

McClelland, J. L., Rumelhart, D. E., & the PDP Research Group. (1986). *Parallel Distributed Processing: Explorations in the microstructure of cognition. Volume II.* Cambridge, MA: MIT Press.

Mikolov, T., Joulin, A., & Baroni, M. (2016). A Roadmap towards Machine Intelligence. *arXiv preprint*. Retrieved from `http://arxiv.org/abs/1511.08130`

Mikolov, T., Sutskever, I., & Chen, K. (2013). Distributed Representations of Words and Phrases and their Compositionality. *Advances in Neural Information Processing Systems*.

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.*

Miller, G. A., & Johnson-Laird, P. N. (1976). *Language and perception.* Cambridge, MA: Belknap Press.

Minsky, M. L. (1974). A framework for representing knowledge. *MIT-AI Laboratory Memo 306*.

Minsky, M. L., & Papert, S. A. (1969). *Perceptrons: An introduction to computational geometry.* MIT Press.

Mitchell, T. M., Keller, R. R., & Kedar-cabelli, S. T. (1986). Explanation-Based Generalization: A Unifying View. *Machine Learning*, *1*, 47–80.

Mnih, A., & Gregor, K. (2014). Neural variational inference and learning in belief networks. In *Proceedings of the 31st International Conference on Machine Learning* (pp. 1791–1799).

Mnih, V., Heess, N., Graves, A., & Kavukcuoglu, K. (2014). Recurrent Models of Visual Attention. In *Advances in Neural Information Processing Systems 27* (pp. 1–9).

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., ... Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature*, *518*(7540), 529–533.

Mohamed, S., & Rezende, D. J. (2015). Variational information maximisation for intrinsically motivated reinforcement learning. In *Advances in neural information processing systems* (pp. 2125–2133).

Moreno-Bote, R., Knill, D. C., & Pouget, A. (2011). Bayesian sampling in visual perception. *Proceedings of the National Academy of Sciences*, *108*, 12491–12496.

Murphy, G. L. (1988). Comprehending complex concepts. *Cognitive Science*, *12*(4), 529–562.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review*, *92*(3), 289–316.

Murphy, G. L., & Ross, B. H. (1994). Predictions from Uncertain Categorizations. *Cognitive Psychology*, *27*, 148–193.

Neisser, U. (1966). *Cognitive Psychology.* New York: Appleton-Century-Crofts.

Newell, A., & Simon, H. A. (1961). *Gps, a program that simulates human thought.* Defense Technical Information Center.

Newell, A., & Simon, H. A. (1972). *Human problem solving.* Prentice-Hall.

Niv, Y. (2009). Reinforcement learning in the brain. *Journal of Mathematical Psychology*, *53*, 139–154.

O'Donnell, T. J. (2015). *Productivity and Reuse in Language: A Theory of Linguistic Computation and Storage*. Cambridge, MA: MIT Press.

Osherson, D. N., & Smith, E. E. (1981). On the adequacy of prototype theory as a theory of concepts. *Cognition*, *9*(1), 35–58.

Parisotto, E., Ba, J. L., & Salakhutdinov, R. (2016). Actor-Mimic: Deep Multitask and Transfer Reinforcement Learning. In *International Conference on Learning Representations (ICLR).* Retrieved from `http://arxiv.org/abs/1511.06342`

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.

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.*

Piantadosi, S. T. (2011). *Learning and the language of thought*. Unpublished doctoral dissertation, Massachusetts Institute of Technology.

Pinker, S. (2007). *The Stuff of Thought*. Penguin.

Pinker, S., & Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*, *28*, 73–193.

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*, 1184–1187.

Premack, D., & Premack, A. J. (1997). *Infants Attribute Value to the Goal-Directed Actions of Self-propelled Objects* (Vol. 9). doi: 10.1162/jocn.1997.9.6.848

Reed, S., & de Freitas, N. (2016). Neural Programmer-Interpreters. In *International Conference on Learning Representations (ICLR).* Retrieved from `http://arxiv.org/abs/1511.06279`

Rehder, B. (2003). A causal-model theory of conceptual representation and categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(6), 1141–59.

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), 323–360.

Rehling, J. A. (2001). *Letter Spirit (Part Two): Modeling Creativity in a Visual Domain*. Unpublished doctoral dissertation, Indiana University.

Rezende, D. J., Mohamed, S., Danihelka, I., Gregor, K., & Wierstra, D. (2016). One-Shot Generalization in Deep Generative Models. In *International Conference on Machine Learning (ICML).* Retrieved from `http://arxiv.org/abs/1603.05106v1`

Rezende, D. J., Mohamed, S., & Wierstra, D. (2014). Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning (ICML).*

Rips, L. J. (1975). Inductive judgments about natural categories. *Journal of Verbal Learning and Verbal Behavior*, *14*(6), 665–681.

Rips, L. J., & Hespos, S. J. (2015). Divisions of the physical world: Concepts of objects and substances. *Psychological Bulletin*, *141*, 786–811.

Rogers, T. T., & McClelland, J. L. (2004). *Semantic Cognition*. Cambridge, MA: MIT Press.

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*, 386–408.

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 (PNAS)*, *102*(20), 7338–7343.

Rumelhart, D. E., Hinton, G., & Williams, R. (1986). Learning representations by back-propagating errors. *Nature*, *323*(9), 533–536.

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* (pp. 216–271). Cambridge, MA: MIT Press.

Rumelhart, D. E., McClelland, J. L., & the PDP Research Group. (1986). *Parallel Distributed Processing: Explorations in the microstructure of cognition. Volume I.* Cambridge, MA: MIT Press.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... Fei-Fei, L. (2015). *ImageNet large scale visual recognition challenge* (Tech. Rep.).

Russell, S., & Norvig, P. (2003). *Artificial Intelligence: A Modern Approach.* Upper Saddle River, NJ: Prentice Hall.

Rusu, A. A., Rabinowitz, N. C., Desjardins, G., Soyer, H., Kirkpatrick, J., Kavukcuoglu, K., ... Hadsell, R. (2016). Progressive Neural Networks. *arXiv preprint*. Retrieved from `http://arxiv.org/abs/1606.04671`

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.

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.

Salakhutdinov, R., Torralba, A., & Tenenbaum, J. (2011). Learning to Share Visual Appearance for Multiclass Object Detection. In *Computer Vision and Pattern Recognition (CVPR).*

Sanborn, A. N., Mansinghka, V. K., & Griffiths, T. L. (2013). Reconciling intuitive physics and newtonian mechanics for colliding objects. *Psychological Review*, *120*(2), 411.

Scellier, B., & Bengio, Y. (2016). Towards a biologically plausible backprop. *arXiv preprint arXiv:1602.05179*.

Schank, R. C. (1972). Conceptual dependency: A theory of natural language understanding. *Cognitive Psychology*, *3*, 552–631.

Schaul, T., Quan, J., Antonoglou, I., & Silver, D. (2016). Prioritized Experience Replay. In *International Conference on Learning Representations (ICLR).* Retrieved from `http://arxiv.org/abs/1511.05952`

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*. doi: 10.3389/fpsyg.2013.00365

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*, 112–143. doi: 10.1016/j.actpsy.2006.05.006

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, *61*, 85–117.

Scholl, B. J., & Gao, T. (2013). Perceiving Animacy and Intentionality: Visual Processing or

Higher-Level Judgment? *Social perception: Detection and interpretation of animacy, agency, and intention*.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*, 1593–1599.

Schulz, L. (2012). The origins of inquiry: Inductive inference and exploration in early childhood. *Trends in Cognitive Sciences*, *16*(7), 382–9.

Schulz, L. E., Gopnik, A., & Glymour, C. (2007). Preschool children learn about causal structure from conditional interventions. *Developmental Science*, *10*, 322–332. doi: 10.1111/j.1467-7687.2007.00587.x

Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., & LeCun, Y. (2014). OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks. In *International Conference on Learning Representations (ICLR)*.

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.

Shultz, T. R. (2003). *Computational developmental psychology*. MIT Press.

Siegler, R. S., & Chen, Z. (1998). Developmental differences in rule learning: A microgenetic analysis. *Cognitive Psychology*, *36*(3), 273–310.

Silver, D. (2016). Personal communication.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Driessche, G. V. D., ... Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, *529*(7585), 484–489.

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.

Solomon, K., Medin, D., & Lynch, E. (1999). Concepts do more than categorize. *Trends in Cognitive Sciences*, *3*(3), 99–105.

Spelke, E. S. (1990). Principles of Object Perception. *Cognitive Science*, *14*(1), 29–56.

Spelke, E. S. (2003). Core knowledge. *Attention and performance*, *20*.

Spelke, E. S., Gutheil, G., & Van de Walle, G. (1995). The development of object perception. In *Visual cognition: An invitation to cognitive science, vol. 2 (2nd ed.). an invitation to cognitive science* (pp. 297–330).

Spelke, E. S., & Kinzler, K. D. (2007). Core knowledge. *Developmental Science*, *10*(1), 89–96.

Srivastava, N., & Salakhutdinov, R. (2013). Discriminative Transfer Learning with Tree-based Priors. In *Advances in Neural Information Processing Systems 26*.

Stadie, B. C., Levine, S., & Abbeel, P. (2016). Incentivizing Exploration In Reinforcement Learning With Deep Predictive Models. *arXiv preprint*. Retrieved from `http://arxiv.org/abs/1507.00814`

Stahl, A. E., & Feigenson, L. (2015). Observing the unexpected enhances infants' learning and exploration. *Science*, *348*(6230), 91–94.

Sternberg, R. J., & Davidson, J. E. (1995). *The nature of insight*. The MIT Press.

Stuhlmüller, A., Taylor, J., & Goodman, N. D. (2013). Learning stochastic inverses. In *Advances in Neural Information Processing Systems* (pp. 3048–3056).

Sukhbaatar, S., Szlam, A., Weston, J., & Fergus, R. (2015). End-To-End Memory Networks. In *Advances in Neural Information Processing Systems 29*. Retrieved from `http://arxiv.org/abs/1503.08895`

Sutton, R. S. (1990). Integrated architectures for learning, planning, and reacting based on ap-

proximating dynamic programming. In *Proceedings of the Seventh International Conference on Machine Learning* (pp. 216–224).

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... Rabinovich, A. (2014). Going Deeper with Convolutions. *arXiv preprint*. Retrieved from `http://arxiv.org/abs/1409.4842`

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* (pp. 2480–2485).

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), 1054–9.

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.

Tian, Y., & Zhu, Y. (2016). Better Computer Go Player with Neural Network and Long-term Prediction. In *International Conference on Learning Representations (ICLR)*. Retrieved from `http://arxiv.org/abs/1511.06410`

Tomasello, M. (2010). *Origins of human communication*. MIT press.

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), 854–869.

Tremoulet, P. D., & Feldman, J. (2000). Perception of animacy from the motion of a single object. *Perception*, *29*, 943–951.

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* (pp. 1622–1627).

Tsividis, P., Tenenbaum, J. B., & Schulz, L. E. (2015). Constraints on hypothesis selection in causal learning. *Proceedings of the 37th Annual Cognitive Science Society*.

Turing, A. M. (1950). Computing Machine and Intelligence. *MIND*, *LIX*, 433–460. Retrieved from `http://mind.oxfordjournals.org/content/LIX/236/433` doi: http://dx.doi.org/10.1093\%2Fmind\%2FLIX.236.433

Tversky, B., & Hemenway, K. (1984). Objects, Parts, and Categories. *Journal of Experimental Psychology: General*, *113*(2), 169–191.

Ullman, S., Harari, D., & Dorfman, N. (2012). From simple innate biases to complex visual concepts. *Proceedings of the National Academy of Sciences*, *109*(44), 18215–18220.

Ullman, T. D., Goodman, N. D., & Tenenbaum, J. B. (2012). Theory learning as stochastic search in the language of thought. *Cognitive Development*, *27*(4), 455–480.

van den Hengel, A., Russell, C., Dick, A., Bastian, J., Pooley, D., Fleming, L., & Agapito, L. (2015). Part-based modelling of compound scenes from images. In *Computer Vision and Pattern Recognition (CVPR)* (pp. 878–886).

van Hasselt, H., Guez, A., & Silver, D. (2016). Deep Reinforcement Learning with Double Q-learning. In *Thirtieth Conference on Artificial Intelligence (AAAI)*.

Vinyals, O., Blundell, C., Lillicrap, T., Kavukcuoglu, K., & Wierstra, D. (2016). Matching Networks for One Shot Learning. *arXiv preprint*. Retrieved from `http://arxiv.org/abs/1606.04080`

Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2014). Show and Tell: A Neural Image Caption Generator. In *International Conference on Machine Learning (ICML)*.

Vul, E., Goodman, N., Griffiths, T. L., & Tenenbaum, J. B. (2014). One and Done? Optimal

Decisions From Very Few Samples. *Cognitive Science*.

Wang, Z., Schaul, T., Hessel, M., van Hasselt, H., Lanctot, M., & de Freitas, N. (2016). Dueling network architectures for deep reinforcement learning. *arXiv preprint*. Retrieved from `http://arxiv.org/abs/1511.06581`

Ward, T. B. (1994). Structured imagination: The role of category structure in exemplar generation. *Cognitive Psychology*, *27*, 1–40.

Watkins, C. J., & Dayan, P. (1992). Q-learning. *Machine Learning*, *8*, 279–292.

Wellman, H. M., & Gelman, S. A. (1992). Cognitive development: Foundational theories of core domains. *Annual Review of Psychology*, *43*, 337–75.

Wellman, H. M., & Gelman, S. A. (1998). Knowledge acquisition in foundational domains. In *The handbook of child psychology* (pp. 523–573). Retrieved from `http://doi.apa.org/psycinfo/2005-01927-010`

Weng, C., Yu, D., Watanabe, S., & Juang, B.-H. F. (2014). Recurrent deep neural networks for robust speech recognition. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*(2), 5532–5536.

Weston, J., Chopra, S., & Bordes, A. (2015). Memory Networks. In *International Conference on Learning Representations (ICLR).*

Williams, J. J., & Lombrozo, T. (2010). The role of explanation in discovery and generalization: Evidence from category learning. *Cognitive Science*, *34*(5), 776–806.

Winograd, T. (1972). Understanding natural language. *Cognitive Psychology*, *3*, 1–191.

Winston, P. H. (1975). Learning structural descriptions from examples. In P. H. Winston (Ed.), *The psychology of computer vision.* New York: McGraw-Hill.

Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review*, *114*(2), 245–272.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., ... Bengio, Y. (2015). Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In *International Conference on Machine Learning (ICML).* Retrieved from `http://arxiv.org/abs/1502.03044`

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*, *111*(23), 8619–24.

Yildirim, I., Kulkarni, T. D., Freiwald, W. A., & Te. (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.*

Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems (NIPS).*

Zeiler, M. D., & Fergus, R. (2014). Visualizing and Understanding Convolutional Networks. In *European Conference on Computer Vision (ECCV).*