Designing AI Systems With Human-Like Learning And Reasoning Abilities

Heta Desai

Abstract

Recent advancements in artificial intelligence (AI) have sparked renewed interest in creating systems that learn and reason like humans. Many of these breakthroughs have stemmed from deep neural networks trained end-to-end on tasks like object recognition, playing video games, and board games—often achieving or surpassing human-level performance. However, despite being inspired by biological systems and demonstrating impressive results, these AI models still differ significantly from human intelligence.

Insights from cognitive science suggest that to build machines that truly think and learn like people, we must move beyond current engineering trends in both the **content** and **methods** of learning. We argue that such systems should:
(a) construct causal models of the world to enable explanation and deep understanding, not just pattern recognition;
(b) base learning on intuitive theories of physics and psychology to provide a richer foundation for acquiring knowledge; and
(c) leverage compositionality and meta-learning ("learning-to-learn") to enable rapid knowledge acquisition and flexible generalization across new tasks and environments.

We outline specific challenges and offer potential directions for integrating the strengths of modern neural networks with more structured, cognitively inspired models to advance toward these goals.

1. Introduction:

Artificial intelligence (AI) has experienced cycles of rapid growth and decline, but recent years have seen unprecedented advancements by most conventional standards. A major driver of this progress is **deep learning**, a method that trains large, multi-layered neural network models. These models have made significant strides across various fields, including object recognition, speech processing, and control systems (LeCun, Bengio, & Hinton, 2015; Schmidhuber, 2015).

For example, in object recognition, Krizhevsky, Sutskever, and Hinton (2012) introduced a deep convolutional neural network (convnet) that drastically reduced the error rate compared to previous state-of-the-art models. Since then, convnets have become the dominant approach, achieving near-human performance on several benchmarks (He et al., 2015; Russakovsky et al., 2015; Szegedy et al., 2014).

Similarly, in speech recognition, traditional Hidden Markov Models (HMMs)—a standard since the late 1980s—have been increasingly replaced by deep learning techniques. Fully neural network-based systems now lead the field (Graves et al., 2013; Weng et al., 2014), outperforming earlier hybrid approaches (Hinton et al., 2012).

Deep learning has also been successfully applied to complex control tasks. For instance, Mnih et al. (2015) merged deep learning with reinforcement learning to create an algorithm that learns to play Atari games directly from raw pixel data and score signals, achieving performance levels on par with or surpassing human players (see also Guo et al., 2014; Schaul et al., 2016; Stadie et al., 2016).

These breakthroughs have helped re-establish neural networks as a dominant framework in machine learning, reminiscent of their popularity in the late '80s and early '90s. The success has also extended into the tech industry, with major players like Google and Facebook heavily investing in deep learning research. These methods now power core features in mobile apps and online platforms. Media coverage often portrays these advances as evidence of neural networks' ability to mimic human thought processes and learning, thanks to their brain-inspired design.

In this article, we take the current excitement around artificial intelligence as an opportunity to explore what it truly means for a machine to learn and think like a human. We begin by examining criteria that cognitive scientists, developmental psychologists, and AI researchers have previously proposed. Next, we outline what we believe are the fundamental components needed to build machines that think and learn like people, drawing on both theoretical frameworks and experimental evidence from cognitive science.

We then analyze how modern AI—especially deep learning—measures up against these components. Our assessment reveals that while deep learning has made great strides, it still lacks many essential aspects of human-like intelligence. This suggests that such systems may be solving problems in fundamentally different ways than humans do. We conclude by outlining promising directions for creating machines that more closely resemble human thinkers. These include integrating deep learning with key cognitive elements—such as attention, working memory, and data structures like stacks and queues—drawn from classic psychology and computer science, which have traditionally seemed at odds with neural network approaches.

Beyond listing specific components, we highlight a deeper divide between two major approaches to intelligence. The first, **statistical pattern recognition**, focuses on prediction within well-defined tasks like classification, regression, or control. In this view, learning is about identifying patterns or features that consistently correlate with specific outcomes across large, diverse datasets.

The second approach prioritizes **world modeling**, where learning involves constructing internal models to make sense of the world. Here, cognition centers on using these models for explanation, imagination, and planning—understanding what we observe, contemplating alternative possibilities, and determining how to influence outcomes. This contrast—between pattern recognition and model-building, or between prediction and explanation—is at the heart of our understanding of human intelligence.

Just as scientists aim to **explain** natural phenomena rather than merely **predict** them, we argue that human cognition is primarily a model-building endeavor. Although pattern recognition isn’t the full story of intelligence, it can play a supporting role by enabling efficient model-building through experience-based, “model-free” learning that makes key inferences easier to compute.

2. Cognitive and Neural Inspiration in Artificial Intelligence

The relationship between AI and human cognitive psychology has deep roots—predating even the terms "artificial intelligence" and "cognitive psychology" themselves. Alan Turing once proposed that instead of replicating adult human intelligence directly, it might be more feasible to build and educate a "child-machine." He imagined such a machine as starting with a mostly blank slate—like a notebook with minimal built-in mechanisms—and learning through experiences of reward and punishment, a concept reminiscent of reinforcement learning. Turing's perspective aligned with the behaviorist psychology dominant in his era and also shares common ground with the modern connectionist view that much of our knowledge can be learned from sensory patterns in the environment.

Cognitive science later moved beyond the simplicity of behaviorism and became foundational to early AI research. For instance, Newell and Simon (1961) developed the "General Problem Solver" as both an AI system and a model of how humans solve problems, which they validated through experiments. Other early AI researchers often referenced human cognition in their work, publishing in cognitive psychology journals and aiming to emulate how children learn rather than hard-coding intelligence. Schank (1972), for example, expressed a desire to build systems that learn as children do, rather than being programmed with vast amounts of pre-set knowledge. Minsky (1974) shared a similar view, suggesting that theories of human thinking and intelligent machines are so closely related that they should be developed together.

During this time, much AI research assumed that human thought could be understood through **symbolic representations**—discrete, structured units used in reasoning, planning, language, and vision. Alongside this symbolic tradition, another approach was emerging: **subsymbolic computation**. This model was based on neuron-like units inspired more by neuroscience than psychology, with early work from researchers like Rosenblatt (1958), Fukushima (1980), and Grossberg (1976). These ideas later evolved into the influential **parallel distributed processing (PDP)** framework developed by McClelland, Rumelhart, and colleagues in the 1980s.

PDP emphasized that intelligent computation could emerge from many simple units operating in parallel, with knowledge represented in a distributed fashion across these units—unlike the localized representations in symbolic systems. The current wave of enthusiasm for **deep learning** is a modern extension of this idea. While benefiting from more advanced hardware, vast datasets, and deeper architectures, deep learning still retains many of the principles introduced in PDP, building powerful models through stacked layers of learned representations (see LeCun et al., 2015; Schmidhuber, 2015).

It's important to note that the **Parallel Distributed Processing (PDP)** approach isn't limited to just pattern recognition—it can also support **model-building**. In fact, some of the early PDP work (Rumelhart, McClelland, & the PDP Research Group, 1986) leaned more towards building internal models than simply identifying patterns. In contrast, many of today's large-scale deep learning systems are more narrowly focused on discriminative pattern recognition (as Bottou, 2014, also discusses). Still, key questions remain about the nature of the learned representations—specifically their **form**, **compositional structure**, and **ability to generalize** or **transfer**—as well as the initial learning setup or "startup software" that helped these models get started. This paper zeroes in on those aspects.

Neural network models and the PDP perspective propose a view of intelligence that is **sub-symbolic**, where learning occurs with minimal pre-defined structure or inductive biases. Supporters of this view argue that traditional notions of structured knowledge—like rules, graphs, grammars, or object hierarchies—might not reflect how thinking actually works. Instead, these structures could be **emergent byproducts** of more fundamental sub-symbolic processes (McClelland et al., 2010). In this view, learning starts from a nearly **blank slate**, similar to Turing's idea of a child-like machine mind with little built-in knowledge.

A typical research strategy within this framework is to start by training a **simple, general-purpose neural network** on a task, and only add complexity if necessary. This approach has produced impressive results: networks have successfully mimicked structured behaviors, like learning past-tense rules in language (Rumelhart & McClelland, 1986), solving basic physics problems (McClelland, 1988), or categorizing living things in a tree-like hierarchy (Rogers & McClelland, 2004).

Modern deep networks trained on object recognition tasks (e.g., He et al., 2015; Krizhevsky et al., 2012) also generate **high-level features** that align with human neural responses in the brain (Khaligh-Razavi & Kriegeskorte, 2014), and can even predict human judgments about image similarity and typicality (Lake et al., 2015; Peterson et al., 2016). Generic neural networks have also been trained to take on more complex, structured behaviors—such as learning to play video games via **Deep Q-learning Networks (DQNs)** (Mnih et al., 2015).

Given these wide-ranging successes—in vision, language, and control—and the ability of neural networks to reproduce behaviors that *appear* rule-based or structured, the key question becomes: **Do we need more than this to build truly human-like learning and thinking machines?** Or can relatively generic neural networks alone take us all the way to that goal?

3. Challenges for building more human-like machines:

Although cognitive science hasn't yet reached a unified theory of the mind or intelligence, the idea that the mind is made up solely of general-purpose neural networks with minimal built-in structure is seen as quite **extreme** by most experts today. Instead, a more widely accepted view emphasizes the role of **innate inductive biases**—such as early-developing concepts of numbers, space, agents, and physical objects. These built-in foundations, along with **powerful learning algorithms** that use prior knowledge, allow humans to learn from **very limited data**. The knowledge we acquire tends to be deeply **structured and theory-like**, supporting the flexible reasoning and creative thinking that are hallmarks of human cognition.

To illustrate this, the authors introduce **two key challenge problems** for AI and machine learning:

1. **Learning simple visual concepts** (Lake, Salakhutdinov, & Tenenbaum, 2015)

2. **Learning to play the Atari game *Frostbite*** (Mnih et al., 2015)

These two examples are used throughout the paper to highlight the importance of incorporating **core cognitive components** into AI systems.

3.1 The Characters Challenge:

The first challenge revolves around recognizing handwritten characters, a long-standing task used to evaluate different machine learning techniques. Hofstadter (1985) once suggested that understanding characters the way humans do—whether handwritten or printed—captures many of the core difficulties in artificial intelligence. Whether or not that's entirely accurate, it does underscore how even seemingly basic concepts like letters involve deep complexity. On a more practical note, people—both kids and adults—need to learn this skill, and it has real-world uses, such as reading addresses on mail or processing handwritten checks at

ATMs. Compared to broader object recognition tasks, recognizing characters is relatively straightforward: characters are flat, visually isolated from their background, and rarely blocked by other elements. Given this, it seems more achievable in the short term to design algorithms that can perceive the same meaningful patterns in characters that humans do.

The MNIST dataset is the widely used standard for evaluating digit recognition systems, where the goal is to classify images of handwritten digits from ‘0’ to ‘9’ (LeCun, Bottou, Bengio, & Haffner, 1998). It includes 60,000 training images, with 6,000 examples for each digit. Because of the large volume of training data, a variety of machine learning algorithms have achieved strong performance. For instance, K-nearest neighbors reports around a 5% test error, support vector machines bring it down to about 1%, and convolutional neural networks (CNNs) perform even better, with error rates under 1%. Some of the most advanced deep CNN models have pushed this error down to just 0.2%, which is comparable to how well humans do (Ciresan, Meier, & Schmidhuber, 2012). Similarly, CNNs have also made significant progress on the more difficult ImageNet benchmark, approaching human-level accuracy in object recognition tasks (Russakovsky et al., 2015).

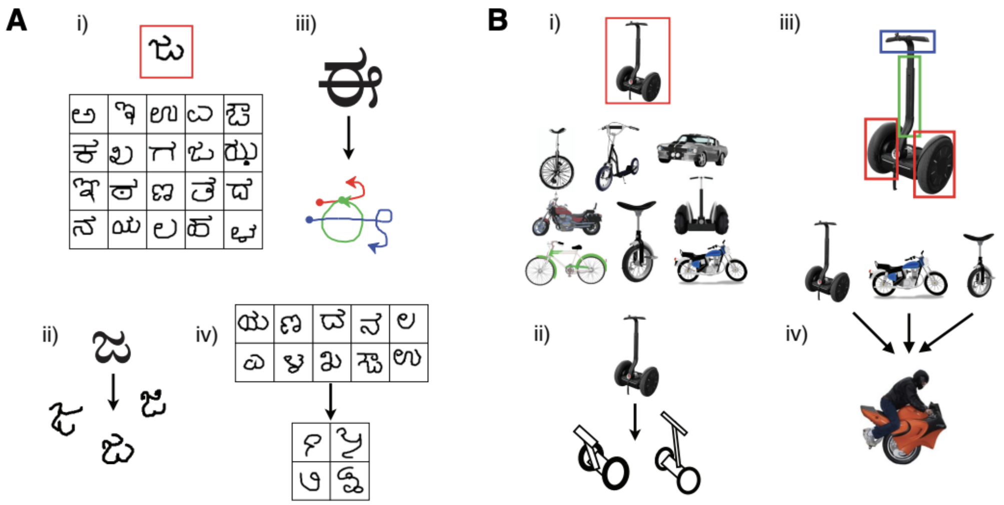


Figure 1: The characters challenge: human-level learning of a novel handwritten characters (A), with the same abilities also illustrated for a novel two-wheeled vehicle (B). A single example of a new visual concept (red box) can be enough information to support the (i) classification of new examples,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

3.2 The Frostbite Challenge:

The second challenge involves the Atari game *Frostbite* (see Figure 2), which was among the games tackled by the Deep Q-Network (DQN) developed by V. Mnih et al. (2015). DQN marked a major milestone in reinforcement learning by demonstrating that a single algorithm could learn to handle a broad range of complex games. It was trained on 49 classic Atari games (as proposed by Bellemare et al., 2013) and achieved human-level or better performance in 29 of them. However, the model particularly struggled with *Frostbite* and other games that demand planning over extended time frames.

In *Frostbite*, the player controls the character Frostbite Bailey, who must build an igloo before the timer runs out. This is done by jumping across moving ice floes—each jump on a white (active) floe contributes a piece to the igloo’s construction (Figure 2A–C). The complexity arises because the ice floes are constantly moving in different directions, and only active ones count toward building. Along the way, the player can earn extra points by collecting fish, but must also avoid various dangers like falling into the water, snow geese, and polar bears. Completing the game level requires forming a long-term plan, achieving sub-goals such as reaching particular floes, and doing so while avoiding hazards. Once the igloo is fully built, the player must reach it before time runs out to finish the level (Figure 2C).

The DQN learns to play Atari games like Frostbite by combining a deep convolutional neural network (CNN), which acts as a powerful pattern recognizer, with a simple, model-free reinforcement learning algorithm (Q-learning). This combination enables the network to translate

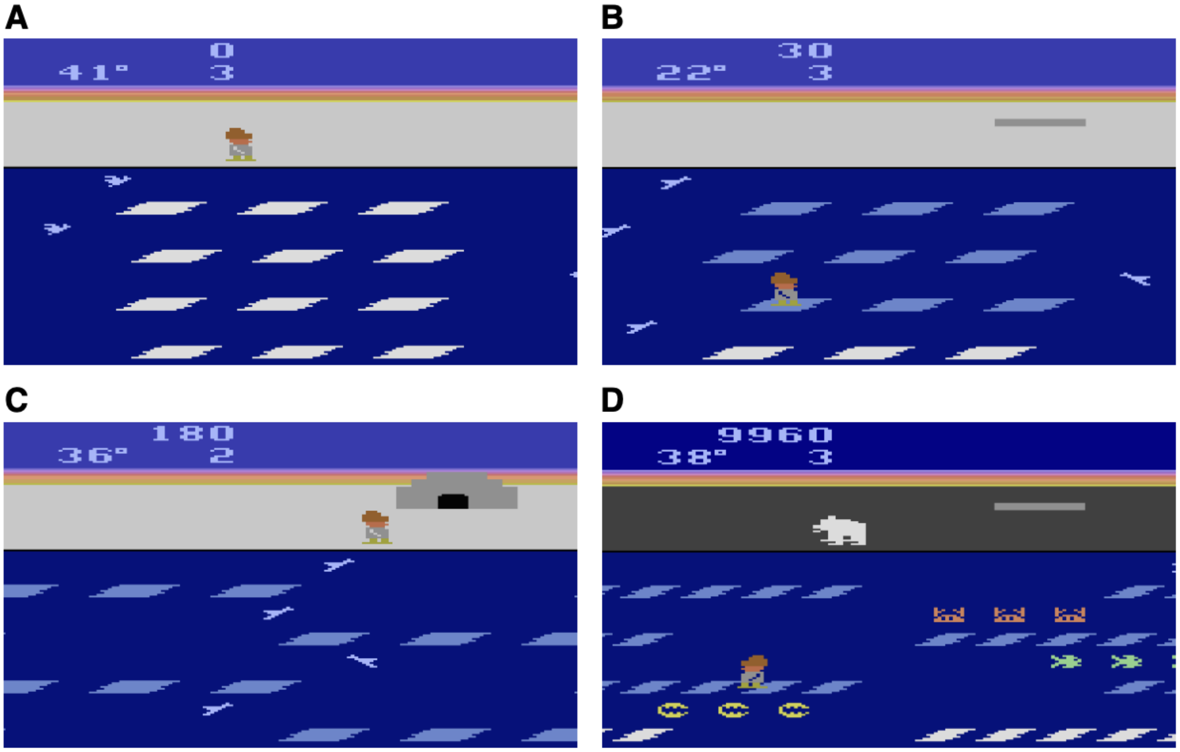


Figure 2: Screenshots of Frostbite, a 1983 video game designed for the Atari game console. A) The start of a level in Frostbite. The agent must construct an igloo by hopping between ice floes and avoiding obstacles such as birds. The floes are in constant motion (either left or right), making multi-step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

visual inputs (pixel frames) into a policy for a limited set of actions, optimizing for long-term rewards, such as the game score. The network follows a largely empirical approach, common in connectionist models, where only basic assumptions about image structure are encoded into the convolutional layers. As a result, the network must learn both a visual and conceptual system from scratch for each new game. In the study by V. Mnih et al. (2015), the network architecture and hyper-parameters were fixed, but it was trained individually for each game, making the visual system and policy highly specific to the game at hand. Later research has demonstrated how game-specific networks can share visual features (Rusu et al., 2016) or be used in multi-task networks (Parisotto, Ba, & Salakhutdinov, 2016), leading to small improvements in transfer learning when playing new games.

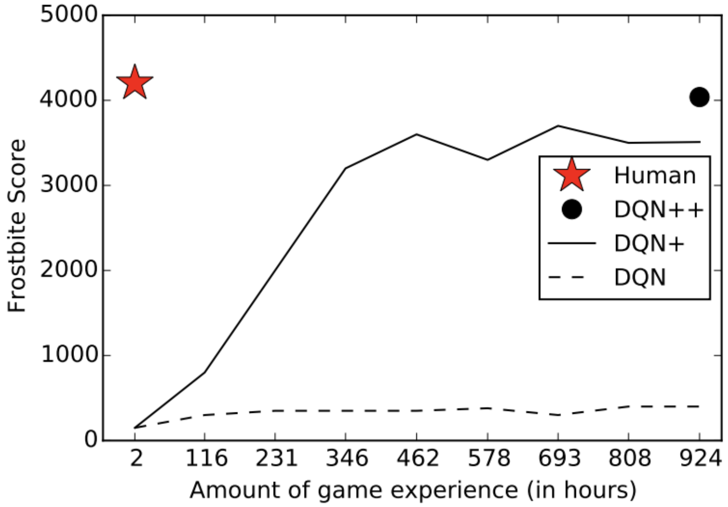


Figure 3: Comparing learning speed for people versus Deep Q-Networks (DQNs). Test performance on the Atari 2600 game "Frostbite" is plotted as a function of game experience (in hours at a frame rate of 60 fps), which does not include additional experience replay. Learning curves (if available) and scores 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).

There are other behavioral differences that highlight the contrasting ways humans and the DQN represent and learn tasks. For example, in the game Frostbite, the DQN receives incremental rewards for reaching each active ice floe, which helps it identify the sub-goals needed to complete the larger task of building an igloo. Without these sub-goals, the DQN would have to rely on random actions until it accidentally builds an igloo and gets rewarded for finishing the level. On the other hand, humans likely don't depend on incremental rewards in the same way when learning a new game. In Frostbite, a person can figure out the overarching goal of building an igloo without needing the incremental feedback. Similarly, sparse feedback poses challenges in other Atari 2600 games like Montezuma's Revenge, where humans significantly outperform current DQN methods.

4. Core ingredients of human intelligence:

In the Introduction, we outlined what we consider to be the fundamental components of intelligence. In this section, we examine these components in more detail and compare them to the current state of neural network modeling. While these are not the only necessary ingredients for human-like learning and thinking (as discussed in Section 5 regarding language), they are critical building blocks that are typically absent in most current learning-based AI systems, especially when not all of them are integrated together. We believe that combining these components could significantly enhance AI systems, making them more powerful and capable of human-like learning and reasoning.

Before diving deeper into each component, it's important to clarify that when we refer to "core ingredients," we don't mean elements that are necessarily hardwired by genetics or must be "built in" to any learning algorithm. Our discussion remains neutral regarding the origins of these key ingredients. By the time a child or adult is learning a new character or figuring out how to play Frostbite, they bring a wealth of real-world experience that deep learning systems lack—experience that would be difficult to replicate in a general way. While the core ingredients are shaped by this experience, some may even be a direct result of it. Whether they are learned, innate, or enriched by experience, the key point is that these ingredients are vital for enabling human-like learning and thought, in ways that current machine learning has not yet achieved.

4.1 Developmental start-up software

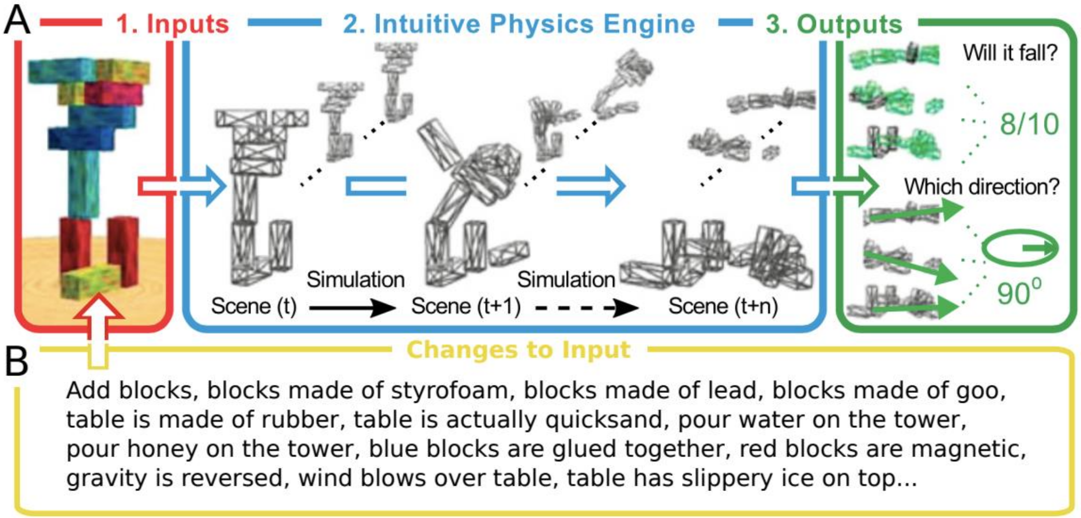
In early development, humans possess a basic understanding of several core domains, including numbers (numerical and set operations), space (geometry and navigation), physics (inanimate objects and mechanics), and psychology (agents and groups). These domains serve as fundamental frameworks for cognition, each organized around specific entities and abstract principles that relate them. The cognitive representations within these domains can be seen as "intuitive theories," with causal structures resembling scientific theories. The idea of the "child as scientist" suggests that learning itself is a scientific process. Recent studies show that children actively seek out new data to test hypotheses, isolate variables, assess causal relationships, draw conclusions from data, and selectively learn from others. We will explore the learning mechanisms in more detail in Section 4.2.

Each of these core domains has been extensively studied, and they are believed to be universally shared across cultures and, to some extent, with non-human animals. While all these domains may enhance current machine learning, we will particularly focus on the early understanding of objects and agents in this section.

4.1.1 Intuitive physics:

Young children possess a rich understanding of intuitive physics, with key physical concepts emerging at a much earlier age than when they learn to play games like Frostbite. Whether these concepts are innate or learned, they are available at a very early stage and can be applied to solve everyday physics-related problems. By as early as 2 months, and possibly even earlier, infants expect inanimate objects to follow principles such as persistence, continuity, cohesion, and solidity. For instance, they believe objects should move along smooth paths, not disappear and reappear, not pass through each other, and not exert influence across distances (Spelke, 1990; Spelke, Gutheil, & Van de Walle, 1995). These expectations help infants segment objects early on, before they rely on appearance-based cues like color or texture (Spelke, 1990).

As infants grow, these early expectations guide their learning. By around 6 months, they begin to distinguish between rigid bodies, soft bodies, and liquids. For example, they expect liquids to pass through barriers, while solid objects cannot (Hespos, Ferry, & Rips, 2009). By their first birthday, infants have already grasped basic physical concepts such as inertia, support, containment, and collisions (Baillargeon, 2004; Baillargeon, Li, Ng, & Yuan, 2009; Hespos & Baillargeon, 2008).



Add blocks, blocks made of styrofoam, blocks made of lead, blocks made of goo, table is made of rubber, table is actually quicksand, pour water on the tower, pour honey on the tower, blue blocks are glued together, red blocks are magnetic, gravity is reversed, wind blows over table, table has slippery ice on top...

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).

4.1.2 Intuitive psychology

Intuitive psychology is another early-developing ability that significantly influences human learning and thinking. Even before they can speak, infants are able to distinguish between animate agents and inanimate objects. This ability is partially based on innate or early-present detectors for low-level cues, such as the presence of eyes, motion from rest, and biological motion (Johnson, Slaughter, & Carey, 1998; Premack & Premack, 1997; Schlottmann, Ray, Mitchell, & Demetriou, 2006; Tremoulet & Feldman, 2000). These cues are often helpful, but not always necessary, for identifying agency. In addition to these low-level signals, infants also expect agents to act in a contingent and reciprocal manner, have goals, and pursue those goals efficiently within constraints (Csibra, 2008; Csibra, Biro, Koos, & Gergely, 2003; Spelke & Kinzler, 2007). These goals can be social in nature; by around three months of age, infants start to distinguish between anti-social agents, who harm or hinder others, and neutral agents (Hamlin, 2013; Hamlin, Wynn, & Bloom, 2010). Later, they further differentiate between anti-social, neutral, and pro-social agents (Hamlin, Ullman, Tenenbaum, Goodman, & Baker, 2013; Hamlin, Wynn, & Bloom, 2007).

It is generally accepted that infants expect agents to act in a goal-directed, efficient, and socially sensitive way (Spelke & Kinzler, 2007). However, there is less agreement on the computational structure that supports this reasoning and whether it involves referencing mental states and explicit goals. One possibility is that intuitive psychology operates through cues alone, without deeper mental constructs (Schlottmann, Cole, Watts, & White, 2013; Scholl & Gao, 2013), though this would require an increasing number of cues as scenarios grow more complex. For example, in a scenario where Agent A is moving toward a box and Agent B blocks A from reaching it, both infants and adults are likely to interpret B's behavior as "hindering" (Hamlin, 2013). This inference could be captured by a cue such as "if an agent's expected path is blocked, the blocking agent is given a negative association."

4.2 Learning as rapid model building:

Since their inception, neural network models have emphasized the importance of learning. There are various algorithms used for neural networks, such as the perceptron algorithm (Rosenblatt, 1958), Hebbian learning (Hebb, 1949), backpropagation (Rumelhart, Hinton, & Williams, 1986), and others. These algorithms generally adjust connection strengths gradually, whether for supervised learning, where the aim is to improve pattern recognition, or for unsupervised learning, which focuses on matching the internal patterns of the model to the statistics of the input data.

Recently, machine learning has made significant progress using backpropagation and large datasets to solve complex pattern recognition tasks. Although these methods have achieved human-level performance on some benchmarks, they still fall short of human learning abilities in other areas. Deep neural networks, for instance, often require much more data than humans to solve similar problems, such as learning to recognize a new object or playing a new game. When learning the meanings of words, children can make meaningful generalizations from very limited data (Carey & Bartlett, 1978; Landau, Smith, & Jones, 1988; E. M. Markman, 1989). Children may only need a few examples of concepts like "hairbrush," "pineapple," or "lightsaber" to grasp the boundaries of these concepts, differentiating them from an infinite set of potential objects. Children learn new concepts rapidly, acquiring around nine or ten new words a day until high school (Bloom, 2000; Carey, 1978). Even in adulthood, this capacity for "one-shot" learning remains—an adult might need to see only a single image or video of a new two-wheeled vehicle to understand the concept and distinguish it from similar objects (Fig. 1B-i).

In contrast to human learning, neural networks are notoriously data-hungry, as they are general function approximators (Geman, Bienenstock, & Doursat, 1992). For example, the ImageNet dataset for object recognition contains hundreds or thousands of examples per class (Krizhevsky et al., 2012; Russakovsky et al., 2015)—such as 1000 images of hairbrushes or pineapples. For tasks like learning new handwritten characters or playing Frostbite, the MNIST benchmark provides 6000 examples per digit (LeCun et al., 1998), and the DQN used by V. Mnih et al. (2015) required about 924 hours of unique training to play each Atari game (Figure 3). Clearly, these algorithms are less efficient with information compared to humans performing the same tasks.

It's also worth noting that some types of concepts are harder for humans to learn. For instance, concepts learned in school, such as mathematical functions, logarithms, derivatives, and scientific concepts like atoms or evolution, are much more challenging. In certain areas, machine learners even outperform humans, such as analyzing financial or weather data. However, for the majority of cognitively natural concepts—those learned by children as part of acquiring language—humans still far exceed machines in learning ability. This section focuses on this type of learning, which is central to reverse engineering and understanding the principles behind human learning. It also presents opportunities for incorporating these principles into the next generation of machine learning and AI algorithms, potentially improving progress in learning both easy and difficult concepts for humans.

4.3 Thinking Fast:

The previous section focused on learning complex models from limited data and suggested key ingredients for achieving human-like learning abilities. These cognitive abilities become even more impressive when considering the speed at which humans perceive, think, and make decisions. Typically, richer and more structured models require more complex and slower inference algorithms, much like how complex models demand more data. This makes the speed of human perception and thought all the more remarkable.

The combination of detailed models with efficient inference presents another area where psychology and neuroscience could offer valuable insights for AI. It also suggests ways to build on the strengths of deep learning, particularly its efficient inference and scalable learning capabilities. This section explores potential solutions to the challenge of balancing fast inference with structured representations, such as using Helmholtz-machine-style approximate inference in generative models (Dayan, Hinton, Neal, & Zemel, 1995; Hinton et al., 1995) and fostering cooperation between model-free and model-based reinforcement learning systems.

5. Responses to Common Questions

Throughout discussions of this paper, three recurring critiques or questions have emerged. We believe it's useful to address these points directly to further the discussion and move forward collectively.

1. **Comparing human and neural network learning speeds is unfair due to humans' prior experience.** It may seem unjust to compare the learning speeds of neural networks and humans for tasks like learning to play Atari games or recognizing handwritten characters, given that humans have extensive prior experience. People have spent many hours playing different games, reading, and writing various characters, among other related activities. The argument is that if neural networks were "pre-trained" with similar experiences, they might generalize in a manner similar to humans when presented with novel tasks.

2. **The biological plausibility of neural networks suggests intelligence theories should begin there.** Our focus has been on how cognitive science can guide the creation of human-like AI, rather than beginning with neuroscience, as some deep learning proponents do. We take a practical view that the best way to formally understand human intelligence is by examining the "software" before the "hardware." In this paper, we outlined key ingredients of this "software." However, we acknowledge that neuroscience can provide valuable insights for both cognitive models and AI. For example, our focus on neural networks and model-free reinforcement learning in the "Thinking Fast" section reflects this connection. That said, what we know about the brain is not always clear or certain, and many widely accepted ideas in neural computation are biologically questionable. Therefore, challenges to cognitive theories that arise from brain-inspired models do not necessarily invalidate these theories.

3. **Language is a crucial aspect of human intelligence. Why is it not emphasized more here?** We have said relatively little about language in this paper, despite its importance to human cognition. While natural language processing (NLP) is an active field of research in deep learning (e.g., Bahdanau, Cho, & Bengio, 2015; Mikolov et al., 2013), it is well recognized that current neural networks fall far short of achieving human-like language abilities. The question arises: how can we develop machines with richer language capabilities?

4. We believe that understanding language and its role in intelligence is closely linked with understanding the foundational abilities discussed in this paper. Intuitive physics, psychology, and rapid learning with compositional, causal models are core abilities that precede language acquisition in children and serve as building blocks for linguistic meaning. We hope that by gaining a deeper understanding of these earlier ingredients and their computational implementation, we can better grasp linguistic meaning and language acquisition, ultimately contributing to the development of more sophisticated language-based AI systems.

6. Looking Forward

In recent decades, AI and machine learning have achieved significant milestones: AI systems have defeated chess masters, triumphed over Jeopardy champions, apps can recognize friends' photos, and machines have reached or surpassed human performance in large-scale object recognition. Furthermore, AI has enabled speech recognition on smartphones and is expected to make great strides in fields like self-driving cars, medicine, genetics, drug design, and robotics. These accomplishments are noteworthy, as they have moved AI research beyond academic circles and into practical applications that enhance our daily lives.

However, it's important to acknowledge both the achievements and the limitations of AI. Despite impressive progress, natural intelligence remains the gold standard. While AI algorithms may perform at or even exceed human levels in certain tasks, they don’t learn or think like humans. Achieving a deeper understanding of human-like intelligence could lead to even more powerful algorithms and may also unlock insights into the workings of the human mind.

When comparing human learning to current AI capabilities, humans stand out for their ability to learn from fewer data points and generalize more flexibly and richly. For instance, humans can recognize and generate new examples or concepts from just a few instances—something deep neural networks still struggle with, even for tasks like handwritten character recognition, where AI models require many examples to train and do not easily generalize to new tasks. We believe that the flexibility and power of human inferences come from the causal and compositional nature of our cognitive representations.

We propose that deep learning and other AI methods can come closer to human-like learning by integrating psychological principles such as those discussed in this paper. Looking ahead, we highlight several promising trends in deep learning that could lead to significant breakthroughs in AI development.

Acknowledgments

We would like to express our gratitude 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 the anonymous reviewers for their valuable feedback on earlier drafts of this manuscript. We also appreciate Tom Schaul for his assistance in clarifying questions about the DQN learning curves and Frostbite scoring. This research was funded by the Center for Minds, Brains, and Machines (CBMM), supported by NSF STC award CCF-1231216, and the Moore-Sloan Data Science Environment at NYU.

References

Bahdanau, D., Cho, K., & Bengio, Y. (2015). "Neural Machine Translation by Jointly Learning to Align and Translate." In *International Conference on Learning Representations (ICLR)*.

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).

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(1998), 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, NY, 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.

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.