### Neuroscience-Inspired Artificial Intelligence

> “We begin with the premise that building human-level general AI is a daunting task, because the search space of possible solutions is vast and likely only very sparsely populated. We argue that this therefore underscores the utility of scrutinizing the inner workings of the human brain—the only existing proof that such an intelligence is even possible.” (245)

Main upshot:

1. Neuroscience provides *inspiration* for AI.
2. Neuroscience can provide *validation* for AI.

# The Past

Neural nets, deep learning, backpropogation, ‘dropout’, reinforcement…

# The Present

“Reading the contemporary AI literature, one gains the impression that the earlier engagement with neuroscience has diminished. However, if one scratches the surface, one can uncover many cases in which recent developments have been inspired and guided by neuroscientific considerations.” (247)

### Attention

Beneficial to focus on different elements of some incoming data stream:

> “…attentional mechanisms have been a source of inspiration for AI architectures that take”glimpses" of the input image at each step, update internal state representations, and then select the next location to sample." (247)

Also helps in producing outputs:

> “Deep generative models…have recently shown striking successes in producing synthetic outputs…via the incorporation of attention-like mechanisms. For example, in one state-of-the-art generative model known as DRAW, attention allows the system to build up an image incrementally, attending to one portion of a”mental canvas" at a time" (247)

### Episodic Memory

Long-term storage of particular performances can help with more efficient learning:

> “One key ingredient in DQN is”experience replay," whereby the network stores a subset of the training data in an instance-based way, and then “replays” it offline, learning anew from successes or failures that occurred in the past. Experience replay is critical to maximizing data efficiency, avoids the destabilizing effects of learning from consecutive correlated experiences, and allows the network to learn a viable value function even in complex, highly structured sequential environments such as video games." (247-8)

Such ‘memories’ can also be directly used for more efficient action selection:

> “These networks store specific experiences (e.g., actions and reward outcomes associated with particular Atari game screens) and select new actions based on the similarity between the current situation input and the previous events stored in memory, taking the reward associated with those previous events into account.” (249)

**Working Memory**

Long-short-term memory (LSTM) networks, which perform really well on a number of tasks, use a kind of working memory architecture:

> “LTSMs allow information to be gated into a fixed activity state and maintained until an appropriate output is required” (249)

But more complex alternatives more closely mirror the biological structures:

> “For example, the differential neural computer (DNC) involves a neural network controller that attends to and reads/writes from an external memory matrix. This externalization allows the network controller to learn from scratch (i.e., via end-to-end optimization) to perform a wide range of complex memory and reasoning tasks that currently elude LSTMs, such as finding the shortest path through a graph-like structure, such as a subway map. . . ” (249)

**Continual Learning**

Animal cognition doesn’t suffer from the same kind of “catastrophic forgetting” (249) that artificial neural networks suffer from. But there’s been progress:

> “Together, these findings from neuroscience have inspired the development of AI algorithms that address the challenge of continual learning in deep networks by implementing of a form of”elastic" weight consolidation (EWC), which acts by slowing down learning in a subset of network weights identified as important to previous tasks, thereby anchoring these parameters to previously found solutions." (250)

# The Future

### Intuitive Understanding of the Physical World

> “Among these capabilities are knowledge of core concepts relating to the physical world, such as space, number, and objectness, which allow people to construct compositional mental models that can guide inference and prediction.” (250)

> “Importantly, the latent representations learned by such generative models exhibit compositional properties, supporting flexible transfer to novel tasks” (250)

### Efficient Learning

Humans can learn new categories from just a handful of examples.

> “Encouragingly, recent AI algorithms have begun to make progress on tasks like the characters challenge, through both structured probabilistic models (Lake et al., 2015) and deep generative models based on the abovementioned DRAW model (Rezende et al., 2016b). Both classes of system can make inferences about a new concept despite a poverty of data and generate new samples from a single example concept.” (250)

### Transfer Learning

> “In the neuroscience litera ture, one hallmark of transfer learning has been the ability to reason relationally, and AI researchers have also begun to make progress in building deep networks that address problems of this nature, for example by solving visual analogies. More generally however, how humans or other animals achieve this sort of high-level transfer learning is unknown, and remains a relatively unexplored topic in neuroscience.” (251-2)

### Imagination and Planning

Humans plan in an efficient way, using crude (but effective) simulations.

> “This”model-free" RL is computationally inexpensive but suffers from two major drawbacks: it is relatively data inefficient, requiring large amounts of experience to derive accurate estimates, and it is inflexible, being insensitive to changes in the value of outcomes. By contrast, humans can more flexibly select actions based on forecasts of long-term future outcomes through simulation-based planning, which uses predictions generated from an internal model of the environment learned through experience." (252)

> “Research into human imagination emphasizes its constructive nature, with humans able to construct fictitious mental scenarios by

> recombining familiar elements in novel ways, necessitating compositional/disentangled representations of the form present in certain generative models" (253)

**Virtual Brain Analytics**

Tools developed to represent what the brain is doing might be used to help us understand what 'black box' algorithms are doing:

> "For example, visualizing brain states through dimensionality reduction is commonplace in neuroscience, and has recently been applied to neural networks."

# From AI to Neuroscience

Two main strands: querying external memory, meta-reinforcement learning.

> "neural networks with external memory typically allow the controller to iteratively query or"hop through" the contents of memory. This mechanism is critical for reasoning over multiple supporting input statements that relate to a particular query. Previous proposals in neuroscience have argued for a similar mechanism in human cognition, but any potential neural substrates, potentially in the hippocampus, remain to be described" (254)

> " meta-reinforcement learning, where RL is used to optimize the weights of a recurrent network such that the latter is able to implement a second, emergent RL algorithm that is able to learn faster than the original" (254)