Paper one

Studying animal cognition and its neural implementation also has a vital role to play, as it can provide a window into various important aspects of higher-level general intelligence.

The benefits to developing AI of closely examining biological intelligence are two-fold. First, neuroscience provides a rich source of inspiration for new types of algorithms and architec-tures, independent of and complementary to the mathematical and logic-based methods and ideas that have largely dominated traditional approaches to AI.

The precise mechanisms by which this is physically realized in a biological substrate are less relevant here (the implementation level). Note this is where our approach to neuroscience-inspired AI differs from other initiatives, such as the Blue Brain Project ( Markram, 2006) or the field of neuromorphic computing systems (Esser et al., 2016), which attempt to closely mimic or directly reverse engineer the specifics of neural circuits (albeit with different goals in mind). By focusing on the computational and algorithmic levels, we gain transferrable insights into gen-eral mechanisms of brain function, while leaving room to accommodate the distinctive opportunities and challenges that arise when building intelligent machines in silico

.

we employ the terms ‘‘neuroscience’’ and ‘‘AI.’’ We use these terms in the widest possible sense. When we say neuroscience, we mean to include all fields that are involved with the study of the brain, the behaviors that it generates, and the mechanisms by which it does so, including cognitive neuroscience, systems neuroscience and psychology. When we say AI, we mean work in machine learning, statistics, and AI research that aims to build intelligent machines.

The brain does not learn by implementing a single, global optimi-zation principle within a uniform and undifferentiated neural network . Rather, biological brains are modular, with distinct but interacting subsystems underpin-ning key functions such as memory, language, and cognitive control .This insight from neuroscience has been imported, often in an unspoken way, into many areas of current AI.

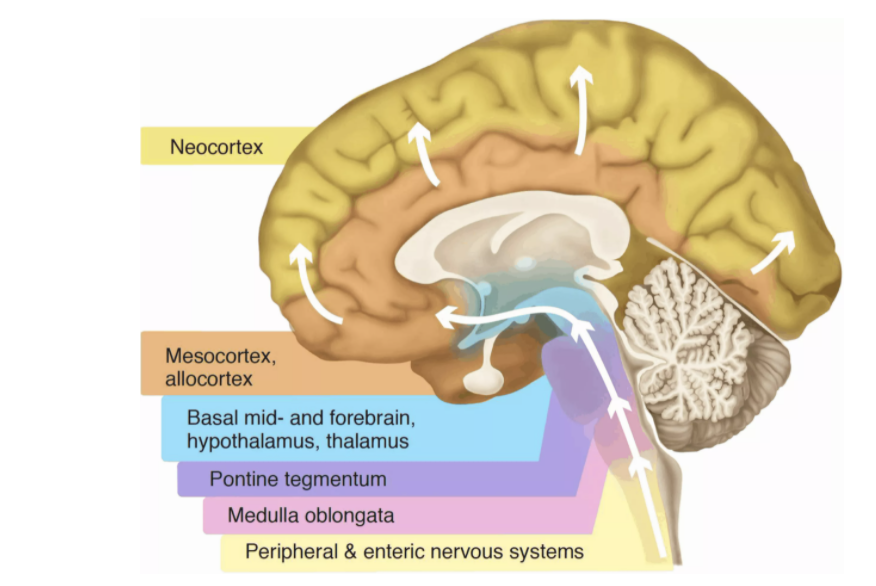
**Differences between AI and the brain when it comes to visual cortex(visual information processing) :**

Up until quite lately, most CNN models worked directly on entire images or video frames, with equal priority given to all image pixels at the earliest stage of processing. The primate visual system works differently. Rather than processing all input in parallel, visual attention shifts strategically among locations and objects, centering processing resources and representational coordinates on a series of regions in turn.

**Hippocampus**:



Hippocampus is a complex brain structure embedded deep into the temporal lobe. It has a major role in **learning and memory**. It is a plastic and vulnerable structure that gets damaged by a variety of stimuli. Studies have shown that it also gets affected in a variety of neurological and psychiatric disorders.The hippocampus is thought to be principally involved in **storing long-term memories** and in making those memories resistant to forgetting, though this is a matter of debate. It is also thought to play an important role in spatial processing and navigation.

**Neocortex**

The neocortex is the center for higher brain functions, such as perception, decision-making and language. Our group focuses on the mechanisms governing neocortex development, with a strong interest on the role and regulation of the neural stem cells.

The majority of the cells in the brain are found in the cerebellum and neocortex. Cerebellum makes use of a lot of cells for coordinated movement while the neocortex uses cells for high precision in planning complex behavior and sensory discrimination.

The replay buffer in DQN might thus be thought of as a very primitive hippocampus, permitting complementary learning in silico much as is proposed for biological brains. Later work showed that the benefits of experience replay in DQN are enhanced when replay of highly rewarding events is prioritized ( Schaul et al., 2015 ), just as hippocampal replay seems to favor events that lead to high levels of reinforcement.

Working Memory

Human intelligence is characterized by a remarkable ability to maintain and manipulate information within an active store, known as working memory, which is thought to be instantiated within the prefrontal cortex and interconnected areas (Gold-man-Rakic, 1990 ). Classic cognitive theories suggest that this functionality depends on interactions between a central controller (‘‘executive’’) and separate, domain-specific memory buffers (e.g., visuo-spatial sketchpad) Baddeley, 2012 ). AI research has drawn inspiration from these models, by building architec-tures that explicitly maintain information over time.

**Prefrontal Cortex**

The prefrontal cortex is the section of the frontal cortex that lies at the very front of the brain, in front of the premotor cortex. The prefrontal cortex can be divided into several subregions. The method of anatomically subdividing the prefrontal cortex varies depending on the source, but common demarcations include the dorsolateral, dorsomedial, ventrolateral, ventromedial, and orbitofrontal regions. There is one category of cognition, however, that the prefrontal cortex is probably best known for: executive function.The term executive function is defined slightly differently depending on where you find the definition. In general, executive functions focus on controlling short-sighted, reflexive behaviors to take part in things like planning, decision-making, problem-solving, self-control, and acting with long-term goals in mind. They are higher-level cognitive processes that people tend to display greater proficiency in than other animals—thus you could argue they are some of the functions that truly help to make human cognition unique.

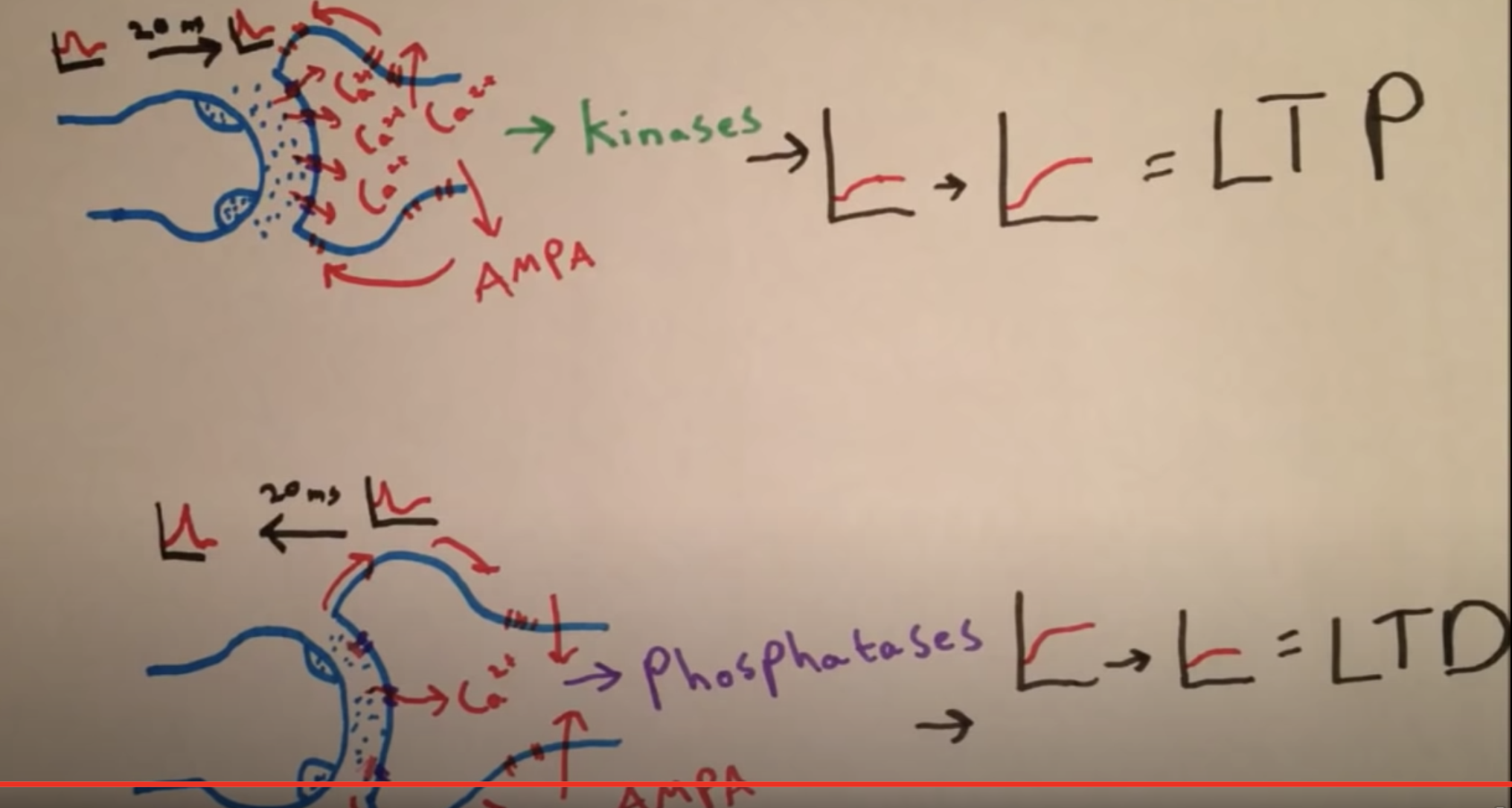
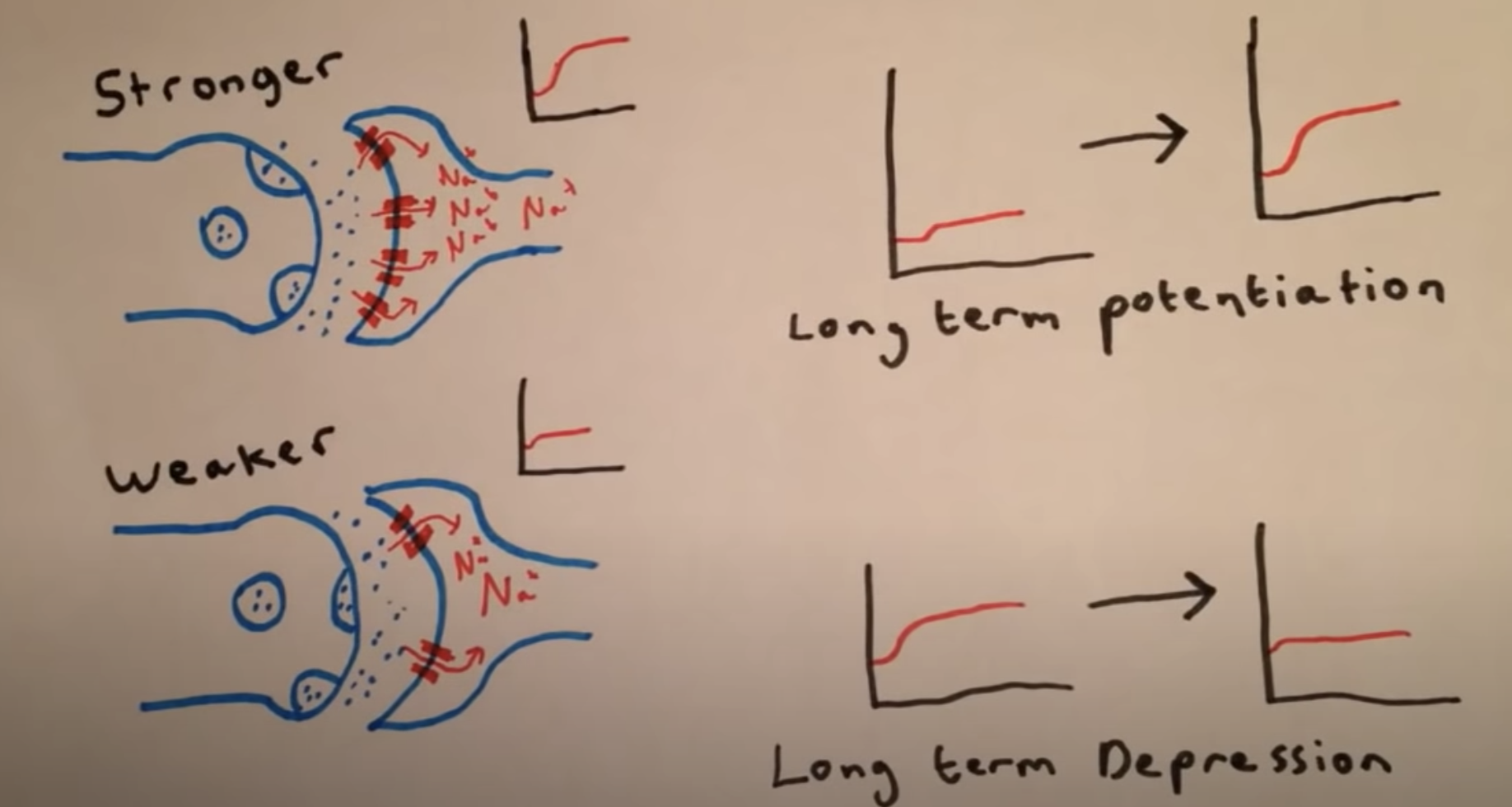
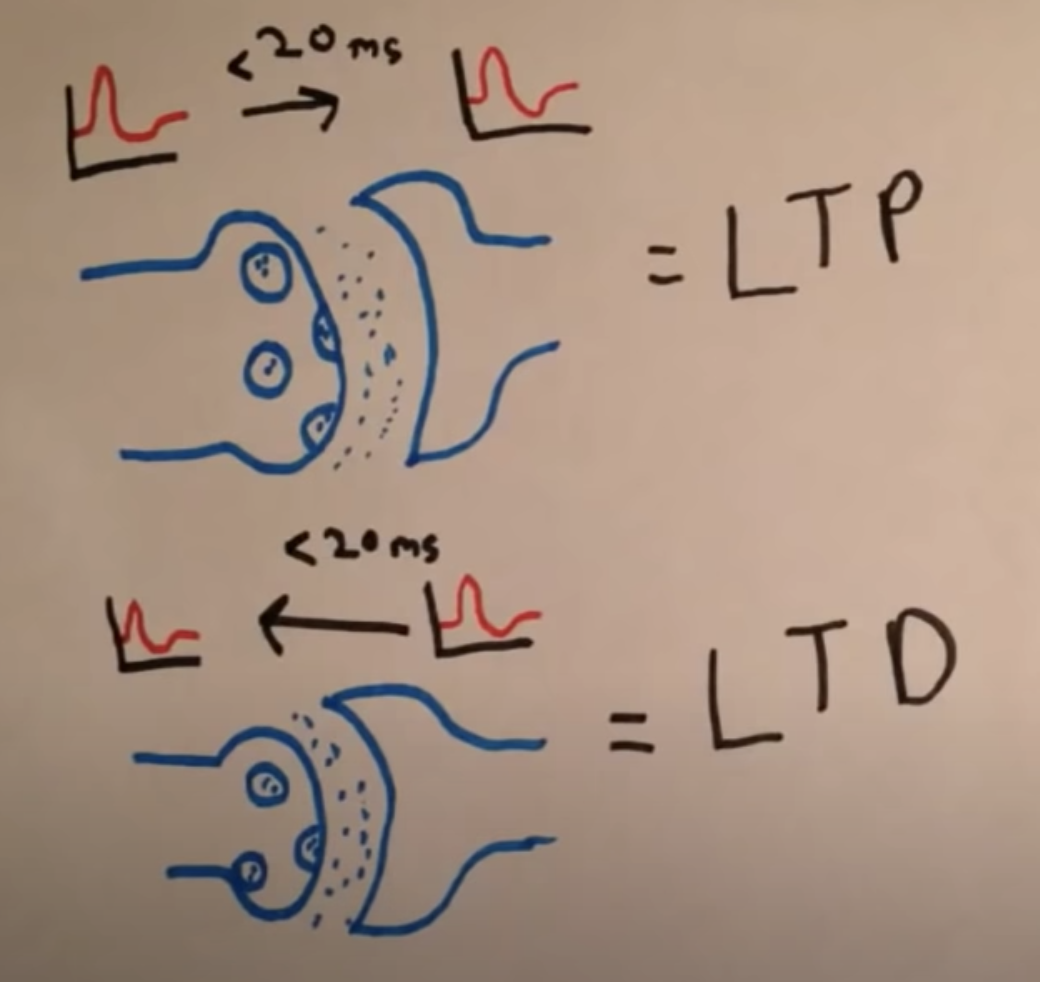
**Continual Learning**

Intelligent agents must be able to learn and remember many different tasks that are encountered over multiple timescales. Both biological and artificial agents must thus have a capacity for continual learning, that is, an ability to master new tasks without forgetting how to perform prior tasks. While animals appear relatively adept at continual learning, neural networks suffer from the problem of **catastrophic forgetting**. This occurs as the network parameters shift toward the optimal state for performing the second of two successive tasks, overwriting the configuration that allowed them to perform the first. Given the importance of continual learning, this liability of neural networks remains a significant challenge for the development of AI.

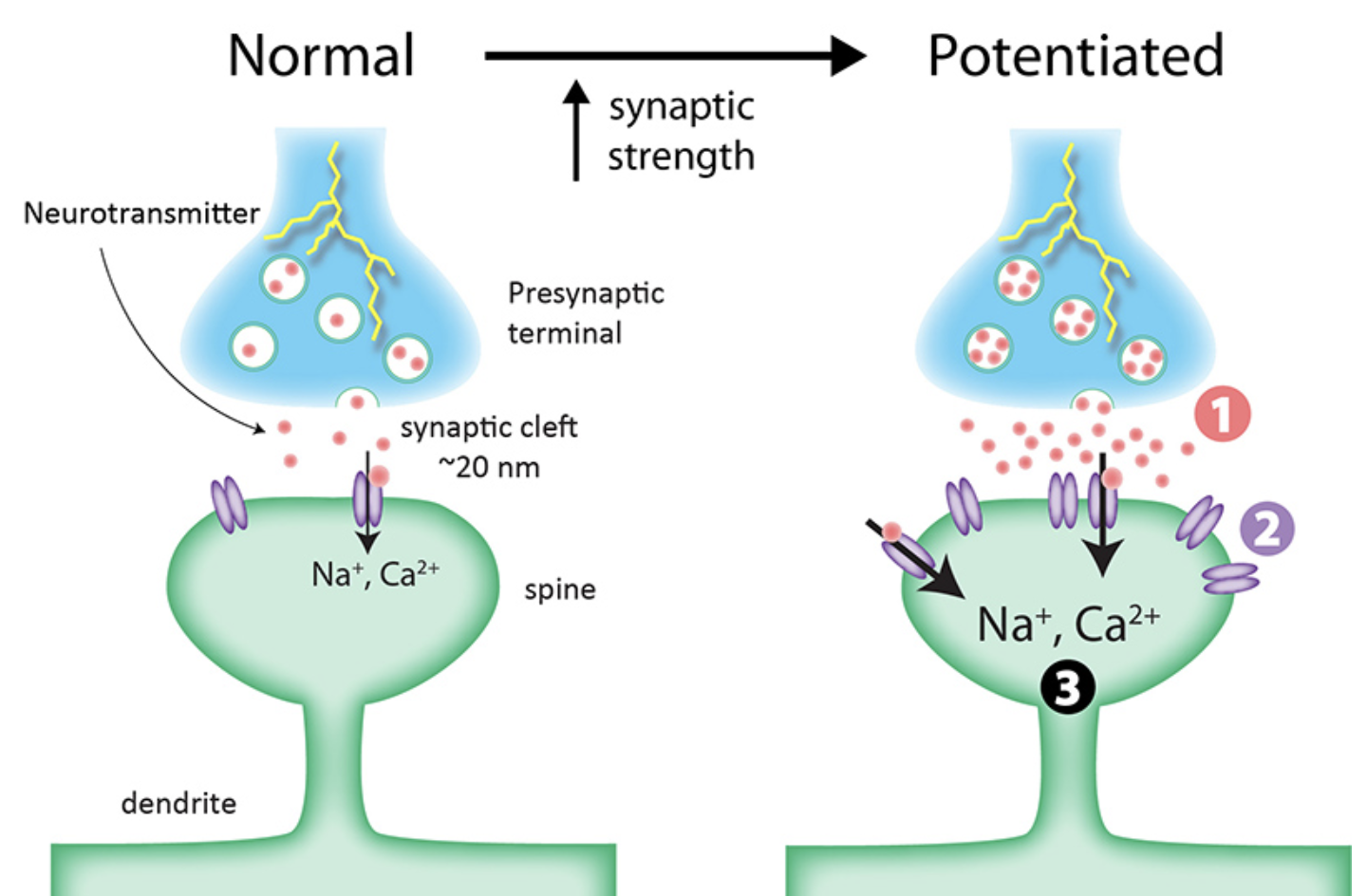
**Strengthening the synaptic plasticity:**

What is synaptic plasticity:

Synaptic plasticity controls how effectively two neurons communicate with each other. The strength of communication between two synapses can be likened to the volume of a conversation. When neurons talk, they do so at different volumes – some neurons whisper to each other while others shout. The volume setting of the synapse, or the synaptic strength, is not static, but rather can change in both the short term and long term. Synaptic plasticity refers to these changes in synaptic strength.



Long-term synaptic plasticity was first reported in 1973. Studying a pathway in the rabbit hippocampus, researchers [discovered](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1350458/) that rapidly and repeatedly activating the synapses made them stronger; the volume control was turned up and stayed that way. They called this long-lasting increase in synaptic strength long-term potentiation, or LTP. The reverse phenomenon, in which synapses become weaker for extended periods, also exists, and is called long-term depression, or LTD



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) ( Kirkpatrick et al., 2017 ), 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 ( Figure 1D). This allows multiple tasks to be learned without an increase in network capacity, with weights shared efficiently between tasks with related structure. In this way, the EWC algo-rithm allows deep RL networks to support continual learning at large scale.

**Efficient Learning**

Human cognition is distinguished by its ability to rapidly learn about new concepts from only a handful of examples, leveraging prior knowledge to enable flexible inductive inferences. In order to highlight this human ability as a challenge for AI, Lake and col- leagues recently posed a ‘‘characters challenge’’ ( Lake et al., 2016 ). Here, an observer must distinguish novel instances of an unfamiliar handwritten character from other, similar items after viewing only a single exemplar. Humans can perform this task well, but it is difficult for classical AI systems.

Humans also excel at generalizing or transferring generalized knowledge gained in one context to novel, previously unseen do-mains.

Despite their strong performance on goal-directed tasks, deep RL systems such as DQN operate mostly in a reactive way, learning the mapping from perceptual inputs to actions that maximize future value. 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

AI research on planning, however, has yet to capture some of the key characteristics that give human planning abilities their power. In particular, we suggest that a general solution to this problem will require understanding how rich internal models, which in practice will have to be approximate but sufficiently accurate to support planning, can be learned through experience, without strong priors being handcrafted into the network by the experimenter. We also argue that AI research will benefit from a close reading of the related literature on how humans imagine possible scenarios, envision the future, and carry out simulation- based planning, functions that depend on a common neural substrate in the hippocampus Although imagina- tion has an intrinsically subjective, unobservable quality, we have reason to believe that it has a conserved role in simula-tion-based planning across species.or example, when paused at a choice point, ripples of neural activity in the rat hippocampus resemble those observed during subsequent navigation of the available trajectories (‘‘preplay’’), as if the animal were ‘‘imagining’’ each possible alternative.

Some encouraging initial progress toward simulation-based planning has been made using deep generative models ( Eslami et al., 2016; Rezende et al., 2016a, 2016b) (Figure 2). In partic-ular, recent work has introduced new architectures that have the capacity to generate temporally consistent sequences of generated samples that reflect the geometric layout of newly experienced realistic environments ( Gemici et al., 2017; Oh et al., 2015 ) ( Figure 2 E), providing a parallel to the function of the hippocampus in binding together multiple components to create an imagined experience that is spatially and temporally coherent (Hassabis and Maguire, 2007 ). Deep generative models thus show the potential to capture the rich dynamics of complex realistic environments, but using these models for simulation-based planning in agents remains a challenge for future work.

An emerging picture from neuroscience research suggests that the hippo- campus supports planning by instantiating an internal model of the environment, with goal-contingent valuation of simulated outcomes occurring in areas downstream of the hippocampus such the orbitofrontal cortex or striatum ( Redish, 2016 ). Notably, however, the mechanisms that guide the rolling forward of an internal model of the environment in the hippocampus remain uncertain and merit future scrutiny. One possibility is that this process is initiated by the prefrontal cortex through interactions with the hippocampus. Indeed, this notion has distinct parallels with proposals from AI research that a separate controller interacts with an internal model of the environment in a bidirectional fashion, querying the model based on task-relevant goals and receiving predicted simulated states as input.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.We think that ultimately these flexible, combinatorial aspects of planning will form a critical underpinning of what is perhaps the hardest challenge for AI research: to build an agent that can plan hierarchically, is truly creative, and can generate solutions to challenges that currently elude even the human mind

**Virtual Brain Analytics**

One rather different way in which neuroscience may serve AI is by furnishing new analytic tools for understanding computation in AI systems. Due to their complexity, the products of AI research often remain ‘‘black boxes’’; we understand only poorly the nature of the computations that occur, or representations that are formed, during learning of complex tasks. However, by applying tools from neuroscience to AI systems, synthetic equiv- alents of single-cell recording, neuroimaging, and lesion tech-niques, we can gain insights into the key drivers of successful learning in AI research and increase the interpretability of these systems. We call this ‘‘virtual brain analytics.’

We think that virtual brain analytics is likely to be an increasingly integral part of the pipeline of algorithmic development as the complexity of architectures increases.

Going further, we believe that building intelligent algorithms has the potential to offer new ideas about the underpinnings of intel-ligence in the brains of humans and other animals. In particular, psychologists and neuroscientists often have only quite vague notions of the mechanisms that underlie the concepts they study. AI research can help, by formalizing these concepts in a quanti-tative language and offering insights into their necessity and suf-ficiency (or otherwise) for intelligent behavior.

We also highlight two recent strands of AI research that may motivate new research in neuroscience. First, 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 state-ments that relate to a particular query (Sukhbaatar et al., 2015). 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(Kumaran and McClelland, 2012). Second, recent work high-lights the potential benefits of ‘‘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 ( Duan et al., 2016; Wang et al., 2016).

The successful transfer of insights gained from neuroscience to the development of AI algorithms is critically dependent on the interaction between researchers working in both these fields, with insights often developing through a continual hand- ing back and forth of ideas between fields. In the future, we hope that greater collaboration between researchers in neuro- science and AI, and the identification of a common language between the two fields ( Marblestone et al., 2016 ), will permit a virtuous circle whereby research is accelerated through shared theoretical insights and Common empirical advances. We believe that the quest to develop AI will ultimately also lead to a better understanding of our own minds and thought pro- cesses. Distilling intelligence into an algorithmic construct and comparing it to the human brain might yield insights into some of the deepest and the most enduring mysteries of the mind, such as the nature of creativity, dreams, and perhaps one day, even consciousness.