1 Introduction

Artificial neural networks support distributed computations in which concepts are represented as patterns of activity in the units that comprise the network layers, and inference is carried out by propagating activation levels between layers weighted by learned connection weights. Artificial neural networks provide a type of fast, flexible computing well suited to handling ambiguity of the sort we routinely encounter in real-world environments, and, by doing so, they complement traditional symbolic computing technologies.

Engineers frequently borrow ideas from nature and generally find it more practical to translate these ideas into current technology rather than attempt to reproduce nature's solutions in detail. Indeed, the basic idea of artificial neural networks has been implemented multiple times using different technologies in order to approximate the connectivity patterns and signal transmission characteristics of real neural circuits while largely ignoring the physiology of real neurons in their implementation.

The human brain supports a wide array of learning and memory systems. Some we have begun to understand functionally and behaviorally, others we can only hypothesize must exist, and still others about which we haven't a clue. Just knowing *that* the brain supports a particular capability can serve as an important clue in engineering complex AI systems. Knowing *how* can lead to an innovative design, enhanced performance and extended competence. In particular, knowing something about how specific biological circuits relate to behavior helps in designing novel network architectures.

We are interested in designing neural network architectures that leverage what is known about biological information processing to solve complex real-world problems. To focus our efforts, we have set out to design end-to-end systems that assist human programmers in writing, debugging and modifying software. We benefit considerably from working closely with scientists from diverse subdisciplines of neuroscience to seek solutions to specific problems and identify additional problems we may have overlooked. The following section explains why this commingling of people, ideas and technologies is so valuable to us in pursuit of our objectives.

2 Neuroscience

From the brain of an Etruscan shrew weighing in at less than a tenth of a gram to a sperm whale brain weighing more than eight kilograms, it is clear that natural selection has stumbled on a basic brain plan and set of developmental strategies that enables it to construct a diverse set of special-purpose brain architectures for efficiently expressing a wide range of sophisticated behavior [**?**, **?**]. The human brain with its approximately 100 billion neurons and the shrew brain with approximately 1 million neurons share the same basic architecture.

The mouse brain has homologues of most human subcortical nuclei and has contributed significantly to our understanding of the human brain and human neurodegenerative disease in particular. The differences between between human and chimpanzee brains are subtle [**?**] and yet humans display a much wider range of behavior and express a much larger repertoire of genes than any other species [**?**]. So what makes the difference?

It's the connections between neurons that matter or, more generally, it's the different types of communication between neurons that biologists refer to as *pathways*. There are electrical, chemical and genetic pathways and each of them obey different constraints and are used for different purposes. They include point-to-point and broadcast methods of communication [**?**]. They transfer information at different speeds and using different coding strategies. Layered architectures are

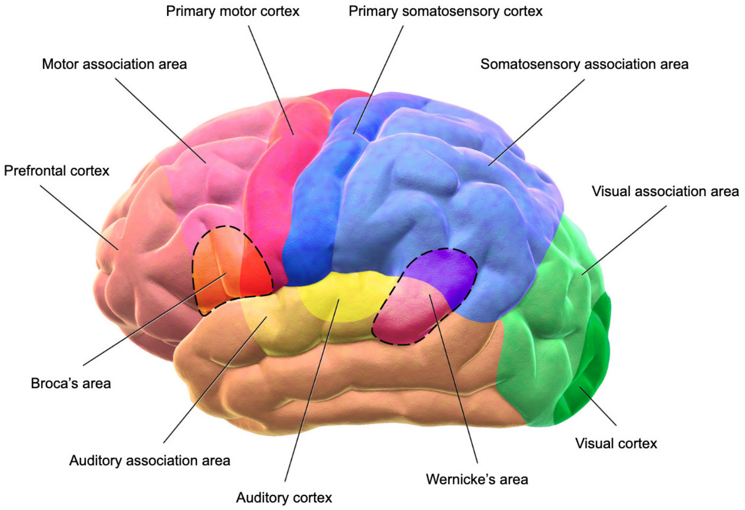

Figure 1: A highly stylized rendering of the major functional areas of the human cortex shown from the side with the head facing to left. Highlighted regions include the occipital lobe shown in shades of green including the primary visual cortex; the parietal lobe shown in shades of blue, including the primary somatosensory cortex; the temporal lobe shown in shades of yellow including the primary auditory cortex; and the frontal lobe shown in shades of pink, including the primary motor and prefrontal cortex. The region outlined by a dashed line on the left is Broca's area and it is historically associated with the production of speech and hence its position relative to the motor cortex. The region outlined by a dashed line on the right is Wernicke's area and it is historically associated with the understanding of speech and hence its position relative to the sensory cortex. Broca's and Wernicke's areas are found only in the dominant hemisphere which is usually the left as shown here.

common not just in the cortex but throughout the brain. It's the wiring that sets humans apart.

2.1 Connectivity

Figure 1 shows the major functional areas of the human neocortex including the primary and secondary sensory and motor areas. The proximal locations of the areas provide a very rough idea of how different functions might relate to another. The graphic shown belies the fact that the brain is three dimensional and much of its functional circuitry hidden under the cortical sheet. The human cerebral cortex is complexly folded to fit within the skull with much of it hidden within the folds. This folded sheet of tissue accounts for more than 75% of the human brain by volume [?] and is largely responsible for the rich behavioral repertoire that humans exhibit. It is worth pointing out in this context that the cortical sheet enshrouds a collection of evolutionarily preserved and highly specialized circuits homologues of which are found in all mammals and without which the cortex would be useless.

The graphic shown in Figure 1 is a simplification of the standard medical textbook diagram. In particular, several of the association areas are not shown and those that are shown are labeled somewhat differently than is common practice. The organizing biological principle is that, the further away from the primary sensory areas, associative functions become more general by integrating information from multiple modalities to construct abstract representations tailored to serve

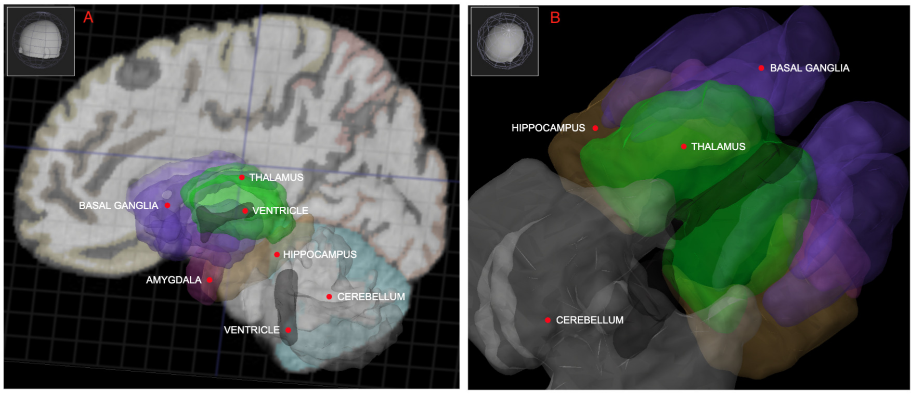

Figure 2: Two 3-D renderings of the human brain generated by the Allen Institute Brain Explorer from the Allen Human Brain Reference Atlas [?]. The inset shown in the left upper corner of each panel indicates the orientation of the head. The left panel (A) shows 3-D reconstructions of several subcortical nuclei featured in this paper. A cross-sectional view of the whole brain is projected on the mid-sagittal plane dividing the right and left sides of the brain illustrating how the cortex envelopes the subcortical regions. The right panel (B) shows the same subcortical nuclei as seen from above (horizontal plane) and to the rear of the brain illustrating how the thalamus is located between the cortical sheet and the subcortical nuclei serving in its role as a relay between the two regions.

ecologically relevant objectives [?]. It is worth contemplating the arrangement of areas to note that they converge on locations in the cortex that will play an important role in decision making and higher-order cognition more generally.

Figure 2 highlights the 3-D structure of several subcortical nuclei emphasized in this paper showing how they relate anatomically to one another and to the cortex. The reconstructions were generated from data generated by *functional magnetic resonance imaging* (fMRI) of adult human subjects [?] and offer additional functional insight complementing conventional histological studies [?]. They don't however provide detailed information concerning either local or long-range connectivity.

Traditionally, tracing connections between individual neurons has been accomplished using a variety of techniques including conventional histological and staining techniques, electrophysiology, neurotropic retroviruses and transgenic organisms expressing fluorescent proteins. However, these methods yield relatively sparse reconstructions and don't scale well to large tissue samples [?, ?].

Small samples of neural tissue can be fixed, stained and sliced into thin sections. Each of the sections is then scanned with an electron microscope and the resulting digital images analyzed with computer vision software to reconstruct neurons and identify synapses [?]. The process is time consuming but can be fully automated and scaled to handle larger samples [?, ?].

It is also possible to reconstruct the major *white matter tracts* formed by bundles of myelinated fibers using diffusion-weighted fMRI and averaging over multiple subjects after registering the individual brain scans with a reference atlas [?, ?]. Unlike the previous technologies, this method is not destructive so it can be applied to human subjects and accuracy is improved by averaging over multiple subjects after registering the individual brain scans with a reference atlas

These major tracts increase the speed of signal transmission between regions allowing for more distant communication in larger brains. The differences between the neocortex in humans and

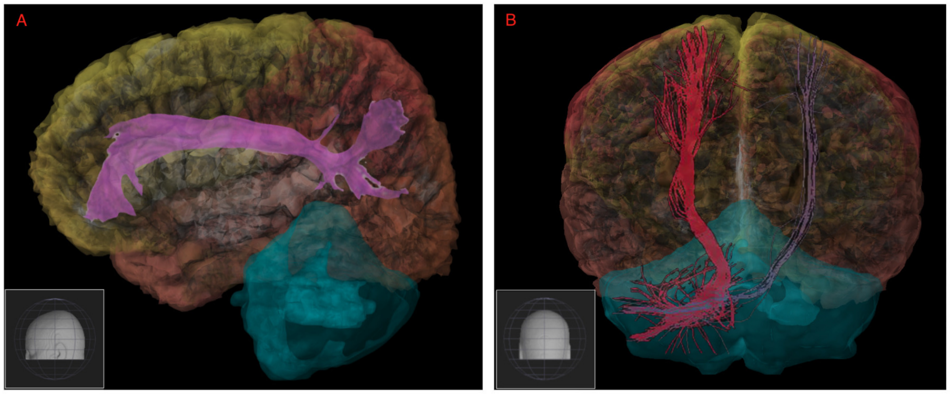

Figure 3: White matter tracts corresponding to bundles of myelinated neurons speed the transmission of information between distant regions of the brain. The left panel (A) shows the connections between the prefrontal cortex and circuits in the parietal and temporal cortex that shape conscious awareness, guide attention and support short-term memory maintenance [?, ?]. The parietal and temporal cortices are known for being home to *association areas* that integrate information from multiple sensory systems thereby creating rich representations necessary for abstract thinking. In humans, white matter tracts between the frontal cortex and the cerebellum — shown in the right panel (B) — facilitate higher-order cognitive function in addition to their role in supporting motor function in all mammals. For example, these connections are particularly important in the development of reading skills in children and adolescents [?, ?].

chimpanzees are subtle [?]; however, white matter connections observed in humans but not in chimpanzees particularly link multimodal areas of the temporal, lateral parietal, and inferior frontal cortices, including tracts important for language processing [?, ?].

The cerebellum in mammals serves to shape motor activities selected in the basal ganglia by ensuring they are precisely timed and well-coordinated. Such activities are initiated by the basal ganglia with executive oversight from the prefrontal cortex. In humans, the cerebellum also supports cognitive functions such as those involved in reading [?]. Figure 3 (B) shows the white matter tracts connecting the cerebellum and prefrontal cortex where such abstract cognitive functions originate.

A white matter bundle called the *arcuate fasciculus* — Figure 3 (A) — provides reciprocal connections between the frontal cortex and association areas in the parietal and temporal lobes plays a key role in attention and the active maintenance of short-term working memory, including support for language processing in the left hemisphere and visuospatial processing in the right hemisphere [?].

The human brain exhibits structure at many scales, the white matter tracts being but one example. A common pattern involves paths that connect multiple circuits that have their own internal components and connections. At a global scale, processing begins in primary sensory areas, propagates forward through dorsal regions integrating additional sources of information to produce composite representations that are processed in the frontal cortex before returning through ventral regions responsible for motivation and action selection.

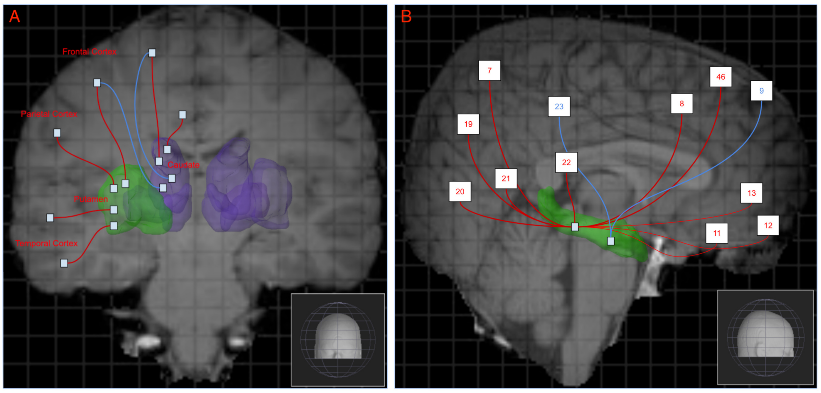

Figure 4: The left panel (A) illustrates the reciprocal connections between two subnuclei of the basal ganglia, the *putamen* and *caudate nucleus*, and locations in prefrontal cortex responsible for influencing action selection. The distinctions between frontal, parietal and temporal cortical areas provide only a very general indication of how their function relates to that of the basal ganglia. The right panel (B) highlights reciprocal connections between cortical regions — identified by the Brodmann areas 7, 8, 9, 11, 12, 13, 19, 20, 21, 22, 23 and 46 — and the hippocampal complex via the adjacent perirhinal (blue) and the parahippocampal (red) areas. The indicated Brodmann areas generally provide a more nuanced understanding of their possible function than does simply stipulating the cortical lobe they reside in.

2.2 Reciprocity

Many of the connections within such paths are reciprocal allowing feedback to adjust behavior and improve prediction. Similar reflective and self-corrective patterns arise within subcortical regions including the hippocampal complex and basal ganglia, e.g., between the dentate gyrus and CA1 in the hippocampus and as layered networks inside individual subcomponents such as the mossy fiber network within the dentate gyrus. Each level solves different problems, offers general insights and provides hints about how one might realize such solutions in artificial systems.

Figure 4 describes how subcortical nuclei such as the hippocampal complex and basal ganglia interact with cortical regions. Such attributions provide insight on how to construct complex artificial neural architectures composed of simpler subnetworks ostensibly responsible for component functions including perception, action selection and episodic memory.

Here we consider two levels of granularity: the first is coarse grained relying on major anatomical divisions illustrated in Figure 1. The second is somewhat finer grained and relies on areal divisions based on cytoarchitectural distinctions involving cell types, neural processes including dendrites and axons, and other histological characteristics.

The former generally employs Korbinian Brodmann's decomposition of the cortex into 52 areas published in 1909 [**?**] and revised several times since then to take advantage of more modern staining and imaging technologies as well as improved methods for functional localization. In many cases, identifying the Brodmann area associated with the endpoint of a neural connection can tell us a good deal about the functional relationship between two brain regions.

The left side of Figure 4 (A) highlights the reciprocal connections between two subnuclei of the basal ganglia, the *putamen* and *caudate nucleus*, and locations in prefrontal cortex responsible for influencing action selection by adjusting input to the basal ganglia and, by way of the thalamus, locations in the parietal and temporal cortex that provide information about the current state relevant to decision making.

We can often improve functional descriptions if we localize to specific Brodmann areas. For example, the *orbitofrontal cortex* (OFC) is located in the prefrontal cortex is a region of the frontal lobes involved in the cognitive process of decision-making. In humans it consists of *Brodmann area 10, 11 and 47.* It is defined as the part of the prefrontal cortex that receives projections from the medial dorsal nucleus of the thalamus, and is thought to represent emotion and reward in decision making [**?**]. The prefrontal cortex, consisting of Brodmann areas 8, 9, 10, 11, 12, 13, 44, 45, 46 and 47, includes the OFC but covers a wider range of functionality.

The right side of Figure 4 (B) highlights reciprocal connections between cortical areas — Brodmann areas 7, 8, 9, 11, 12, 13, 19, 20, 21, 22, 23 and 46 — and the hippocampal complex via the adjacent *perirhinal* cortex (shown as blue connections) and the *parahippocampal* cortex (shown as red connections) that are involved in representing and recognizing objects and environmental scenes.

The anatomy of the brain and patterns of connectivity linking its major functional areas provide a structural account that derives from and informs function. However, functional analyses relating to human cognition require technologies that are able to record neural activity or its correlates aligned with relevant behavioral features. Non-human model systems often employed when invasive technology is required.

On the one hand, optogenetics, two-photon microscopy and conventional electrophysiology are able to record from and modify the electrical activity of tens to thousands of neurons at the resolution of a few microns. While this represents progress, the coverage is inadequate for many studies, and the methods are, for the most part, limited to non-human subjects due to the invasive nature of their practical application [**?**, **?**, **?**, **?**].

Conversely, fMRI can used to study awake, behaving humans performing a wide range of cognitive tasks, but relies on signals that are at best indirectly related to neural activity as in the case of blood oxygen levels, and that are currently limited to spatial resolutions on the order of tens of millimeters and temporal resolutions on the order of hundreds of milliseconds [**?**, **?**, **?**].

Moreover, the electrical activity of individual neurons is but one marker for neural function. Other pathways including diffuse signaling by way of chemical neuromodulation and genetic activity and protein transport at the cellular level are becoming increasingly important as markers for behavior at multiple time scales [**?**]. Despite these limitations, neuroscientists have made considerable progress by combining information from different model systems using multiple recording technologies.

2.3 Sensorimotor Hierarchy

Much of the cortex is in the business of learning representations of concepts relevant to survival. Perception is the means by which we apprehend and act on the physical realization of the concepts we have learned. It seems obvious that perception serves action. It may not seem so obvious that action serves perception, but the fact is we are almost always moving our head, hands and torso in order to resolve ambiguities in what we see, feeling the shape of unfamiliar objects in order to grasp them firmly and twisting about to see who is behind us calling our name or to get a better idea of where we've come from in order to ensure we can retrace our steps. These are complex sensorimotor activities we depend on every day.

In thinking about physically realizable concepts we think first about what they look, feel, sound and smell like. The sensory cortex is responsible for constructing a hierarchy of representations to characterize such concepts, not to capture everything we sense, but rather to account for what we

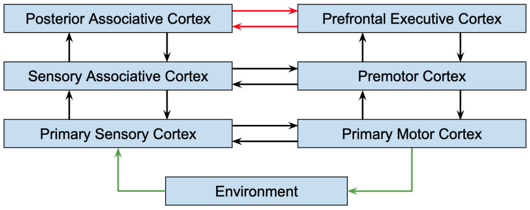
Posterior Associative Cortex
Prefrontal Executive Cortex
Sensory Associative Cortex
Premotor Cortex
Primary Sensory Cortex
Primary Motor Cortex
Environment

Figure 5: A simplified block diagram of the cortex. The column on the left represents the posterior cortex including the occipital, temporal and parietal lobes. The column on the right represents the frontal lobe of the cortex corresponding to the primary motor cortex, premotor cortex (association motor cortex) and prefrontal cortex. Green arrows represent interaction with the environment, black arrows represent sensorimotor abstractions and red arrows indicate cognitive activity relating speech, planning and abstract thinking. See the main text for more detail. Adapted from Figure 8.9 in [**?**]

need to know about concepts to survive. Reconstructing scenes with photographic realism is not what our sensory systems were designed for. Circuits of the primary sensory cortex feed into the circuits of the (unimodal) association sensory cortex that feed into (multimodal) sensory cortex. All of these representations are abstract and yet patterns of regionalization are remarkably preserved within species [**?**, **?**, **?**].

Concepts arise in patterns of neural activity that account for what we need to know about them, including how they appear to us so we can recognize them, what affordances they offer for us to make use of them and how we might predict their occurrence in decision making. Many of the concepts that are represented in our brains serve to model the dynamics of physical systems that we interact with every day, such as riding a bike, working with tools, opening doors, negotiating stairs and riding escalators in department stores. Just as important, if not more so, are the social dynamics we deal with at work and school with their constantly shifting personal relationships and status rankings.

If you are a software engineer designing robot control systems, you might give action much the same scrutiny as perception and build a parallel hierarchy of representations that describes the concepts that relate to movement including navigation, articulation and manipulation ranging from servo-motor commands to strategies for moving furniture, but designing or learning these hierarchies independently is generally a bad idea. In mammals, these two hierarchies are tightly coupled to account for how they depend on one another [**?**].

Indeed, determining what sensory representations to learn depends upon and influences what motor representations to learn and *vice versa*, where we follow the convention of using the term *motor* as a catchall term for concepts relating to muscles and movement. As pointed out in the introduction, there is evidence to suggest that circuits occurring early in the ventral visual stream code for object-selective features and exhibit large-scale organization characterized by the high-level properties of animacy and object size [**?**, **?**].

Figure 5 is a simplified block diagram of the cortex organized as two columns. The left column represents the posterior cortex consisting of the occipital, temporal and parietal lobes that are primarily concerned with processing sensory information. The relevant brain areas are summarized in three blocks roughly corresponding to primary sensory cortex, unimodal association cortex and multimodal association cortex stacked so the least abstract concepts are on the bottom and most abstract on the top. The combined area is often referred to as *semantic memory* and characterized as long-term declarative memory [?].

The right column represents the frontal lobe of the cortex corresponding to the primary motor cortex, premotor cortex (associative motor cortex) and prefrontal cortex. The primary motor cortex is responsible for creating abstract representations of motor activity throughout the body. The premotor cortex is responsible for integrating sensory and motor abstractions to construct sensorimotor representations. The prefrontal cortex orchestrates cognitive behavior including speech, planning and abstract thinking, and is reciprocally connected to the association areas just mentioned as well subcortical structures including the basal ganglia and hippocampus.

The two columns are connected with one another at multiple levels: by physical interaction with the environment (green arrows), by sensorimotor abstraction and alignment (black arrows), and by cognitive effort in directing activity mediated through subcortical structures (red arrows). This arrangement supports the formation of rich representations that serve a wide range of cognitive function. The sensorimotor connections and feedback through the environment provide an inductive bias to guide learning, ground inference and reduce sample complexity by reducing reliance on labeled data and enabling opportunities for unsupervised learning [?].

Simple as this model of cortical function may seem, it may be one of the most important architectural contributions of neuroscience to the development of artificial intelligence patterned after the human brain. Some of the lessons have already been integrated into the discipline of control theory through exposure to early work in biological cybernetics [?, ?, ?, ?, ?, ?], but some of the most important lessons impact the application of machine learning in building autonomous embodied systems including robots and digital assistants as alluded to above.

Box A: Pattern Separation, Completion and Integration

As discussed in Section **??**, *pattern separation* reduces the similarity between input patterns of activity by orthogonalizing inputs to minimize interference between patterns and increase hippocampal storage capacity [?]. Pattern separation involves primarily DG and CA3 — see Section **??** for an explanation of the acronyms. The DG maps input from EHC to a much larger and sparsely active GC population. In rats, the number of neurons in the DG exceeds that in EHC by about 5:1 [?]. This expansion coding with strong inhibitory interneurons and a competitive learning rule can greatly reduce the overlap between inputs. The DG connects to CA3 mainly through *mossy fibers* that reliably activate CA3 pyramidal neurons and sustain activation for tens of seconds [?]. Each CA3 neuron receives a small number of these connections from DG so the degree of sparsity is maintained [?].

Pattern completion reconstructs the complete stored pattern given a partial input. Each pyramidal neuron in CA3 receives a large number of synapses from other pyramidal cells forming a recurrent network that serves as an autoassociative memory for pattern completion [?]. During learning, recurrent connections between active CA3 neurons are strengthened and later when neurons encoding part of an episode are reactivated, they recurrently activate other connected cells to reconstruct the original episode. Basket cells in CA3 form inhibitory synapses to pyramidal cells to dampen excitatory responses thereby emphasizing key features [?].

Pattern completion provides access to relevant experience to support decision making in novel situations, and while pattern separation helps downstream discrimination, perfectly orthogonal representations are not ideal in the case we want events that occurred close together to have similar representations. In this case, *pattern integration* represents related experiences as overlapping populations. There are a number of neural mechanisms suggested to support pattern integration in the hippocampus. We consider two here, the first of which involves *neurogenesis*.

There is evidence that hundreds of new GCs are added to an adult human hippocampus everyday [?], and stronger evidence suggests that thousands of new GCs are added to rodent's hippocampus, though not all survive [?]. Unlike mature GCs that fire sparsely, immature GCs are more active and have lower threshold for induction of long-term potentiation [?, ?, ?]. Aimone *et al* [?] posit that a population of hyperactive young GCs could collectively encode events close in time to decrease pattern separation in DG. Others hypothesize that neurogenesis may increase storage capacity by protecting old GCs from new information [?, ?] or that young active GCs could improve the resolution of memory content [?].

Alternatively, pattern integration might be enabled by recurrent connections involving the hippocampus and neocortex. Recurrent connections in the hippocampus, mainly in CA3 region, can replay an entire episode given a part of it. The replayed episode is backprojected to the neocortex through EHC, that can then recirculate the replayed episode as input to hippocampus to trigger replay of another episode that has overlapping elements with previous one. Kumaran *et al* [?] propose that this kind replay between hippocampus and neocortical regions can combine representations of elements that seldom occur together but appear in similar contexts. In addition to integrating experiences with shared elements, backprojection to the medial prefrontal cortex (mPFC) may bias hippocampus to reactivate experiences that are more behaviorally relevant [?] — see Box B for more on behavioral relevance. The concurrent presentation of these memories in mPFC may further improve the learning of abstraction relations across episodes.

3 Architecture

The architecture of the human brain, at any scale you choose to consider, bears little or no resemblance to conventional computer architectures. There is no separate program memory, no centralized processing unit, no highly stable, random-access, non-volatile memory and nothing like the digital level of abstraction that enables software engineers to ignore instabilities in the analog circuits that implement logic gates. Since representations (data) are collocated with the transformations (computations) that operate on them and different parts of the brain perform different computations requiring different types of memory, the human brain has to support multiple memory systems.

Human memory is characterized along several dimensions depending on what sort of information is stored, how it is accessed and how long it remains accessible [?]. Short-term, long-term and working memory are differentiated on the basis of access, persistence, volatility and the effort required to maintain. Short term is measured in seconds, long term in days, months or years and working memory is essentially short-term memory that can be maintained (with cognitive effort) indefinitely and manipulated (very roughly) analogous to a register in the ALU of a von Neumann machine [?].

Declarative memory is defined by the ability to explicitly (consciously) recollect facts, whereas non-declarative memory is accessed unconsciously or implicitly through performance rather than recollection. Episodic memory is generally considered long-term and declarative, and is further differentiated on the basis of the kinds of relationships it can encode, including spatial, temporal and social [?, ?, ?, ?]. Procedural knowledge, including motor, visuospatial and cognitive skills, is encoded in the cerebellum, the putamen and caudate nucleus of the basal ganglia, the motor cortex, and frontal cortex.

To ground the discussion, we introduce the *programmer's apprentice* as an example of the sort of digital assistants we envision as an application of the technologies presented in this paper. We consider several core components of the apprentice architecture each of which depends on or implements one or more memory systems. Drawing upon concepts covered earlier, we consider three major elements:

1. the role of the posterior cortex role in supporting declarative knowledge and semantic memory,

2. the basal ganglia and prefrontal cortex as the basis for motivation and executive function, and

3. the hippocampal formation in supporting episodic memory formation, retrieval and consolidation.

3.1 Embodied Cognition

Embodied cognition is the theory that an organism's body shapes its understanding of the environment it inhabits and grounds its perception of and interaction with that environment. Importantly, the environment completes a loop that links perception and action enabling the organism to formulate and test predictive models that guide behavior. Such models serve as the foundation for commonsense reasoning and provide a starting point for understanding a much wider range of concrete and abstract systems, giving rise to a tendency in humans to attribute self-styled agency to both animate and inanimate objects.

To ground our discussion, we consider a personal assistant that works with a software engineer in the role of an apprentice learning on the job, as was common in the guilds and trade associations of medieval cities. The programmer's apprentice we imagine here is a novice programmer but has the intuitive skills of an idiot savant, given that the apprentice has a suite of powerful programming tools as an integral part of its brain. These tools constitute the assistant's body, its peripheral nervous system if you will.

The original programmer's apprentice was the name of project initiated at MIT by Chuck Rich and Dick Waters and Howie Shrobe to build an intelligent assistant that would help a programmer to write, debug and evolve software. Our version of the programmer's apprentice is implemented as an instance of a hierarchical neural network architecture. It has a variety of conventional inputs including speech, text and vision, as well as output modalities including the ability to run code and display program output and execution traces.

The programmer's apprentice relies on multiple sources of input, including dialogue in the form of text utterances, visual information from an editor buffer shared by the programmer and apprentice and information from a fully *instrumented integrated development environment* (FIDE) designed for analyzing, writing and debugging code adapted to interface seamlessly with the apprentice as we might move our limbs or direct our gaze. As in case of the legs you were born with, the apprentice has to learn how to use its prosthetic extensions.

This input is processed by a collection of neural networks modeled after the primary sensory areas in the primate brain. The outputs of these networks feed into a hierarchy of additional networks corresponding to uni-modal secondary and multi-modal association areas that produce increasingly abstract representations as one ascends the hierarchy as illustrate in Figure 6.

Architecturally, the apprentice's FIDE is an instance of a differentiable neural computer (DNC) introduced by Alex Graves and his colleagues at DeepMind [?]. The assistant combined with its FIDE corresponds to a neural network that can read from and write to an external memory matrix,

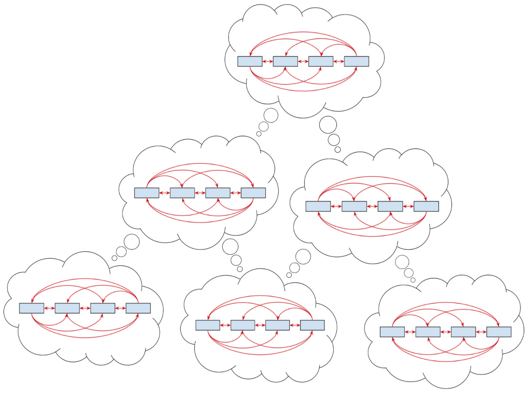

Figure 6: The architecture of the apprentice sensory cortex including the layers corresponding to abstract, multi-modal representations handled by the association areas can be realized as a multi-layer hierarchical neural network model consisting of standard neural network components. This graphic depicts these components as encapsulated in thought bubbles of the sort often employed in cartoons to indicate what some cartoon character is thinking. Analogously, the technical term ”thought vector” is used to refer to the activation state of the output layer of such a component. All of the bubbles appear to contain networks with exactly the same architecture, where one might expect sensory modality to dictate local architecture. The hierarchical architecture depicted here is modeled after the mammalian neocortex that appears to be tiled with columnar component networks called cortical columns that self-assemble into larger networks and adapt locally to accommodate their input. In practice, it may be necessary to engineer modality-specific networks for the lowest levels of the hierarchy — analogous to the primary sensory and motor areas of the neocortex, but more general-purpose networks for the higher levels in the hierarchy — analogous to the sensory and motor association areas.

combining the characteristics of a random-access memory and set of memory-mapped device drivers and programmable interrupt controllers. The interface supports a fixed number of commands and channels that provide feedback.

The integrated development environment and its associated software engineering tools constitute an extension of the apprentice's capabilities in much the same way that a piano or violin extends a musician. The extension becomes an integral part of the person possessing it and over time their brain creates a topographic map that facilitates interacting with the extension. We expect the same to occur in the case of the assistant.

3.2 Conscious Attention

Stanislas Dehaene and his colleagues have developed a computational model of consciousness that provides a practical framework for thinking about consciousness that is sufficiently detailed for much of what an engineer might care about in designing digital assistants [**?**]. Dehaene's work extends the *Global Workspace* Theory of Bernard Baars [**?**]. Dehaene's version of the theory combined with Yoshua Bengio's concept of a *consciousness prior* and deep reinforcement learning [**?**, **?**] suggest a model for constructing and maintaining the cognitive states that arise and persist during complex problem solving [**?**].

Global Workspace Theory accounts for both conscious and unconscious thought with the primary distinction for our purpose being that the former has been selected for attention and the latter has not been so selected. Sensory data arrives at the periphery of the organism. The data is initially processed in the primary sensory areas located in posterior cortex, propagates forward and is further processed in increasingly-abstract multimodal association areas. Even as information flows forward toward the front of the brain, the results of abstract computations performed in the association areas are fed back toward the primary sensory cortex. This basic pattern of activity is common in all mammals.

Humans have a large frontal cortex that includes machinery responsible for conscious awareness and that depends on an extensive network of specialized neurons called *spindle cells* that span a large portion of the posterior cortex allowing circuits in the frontal cortex to sense relevant activity throughout this area and then manage this activity by creating and maintaining the persistent state vectors that are necessary when generating extended narratives or working on complex problems that require juggling many component concepts at once. Figure 7 suggests a neural architecture combining the idea of a global workspace with that of an attentional system for identifying relevant input.

These attentional networks are connected to regions throughout the cortex and are trained via reinforcement learning to recognize events worth attending to according to the learned value function. Using extensive networks of connections — both incoming and outgoing, attentional networks are able to create a composite representation of the current situation that can serve a wide range of executive cognitive functions, including decision making and imagining possible futures. The basic idea of a neural network trained to attend to relevant parts of the input is key to a number of the systems that we'll be looking at.

In their paper [**?**] in *Nature*, The authors note that "there are interesting parallels between the memory mechanisms of a DNC and the functional capabilities of the mammalian hippocampus. DNC memory modification is fast and can be one-shot, resembling the associative long-term potentiation of hippocampal CA3 and CA1 synapses. The hippocampal dentate gyrus, a region known to support neurogenesis, has been proposed to increase representational sparsity, thereby enhancing memory capacity: usage-based memory allocation and sparse weightings may provide similar facilities." See the discussion of neurogenesis as an algorithmic technique in Box A.

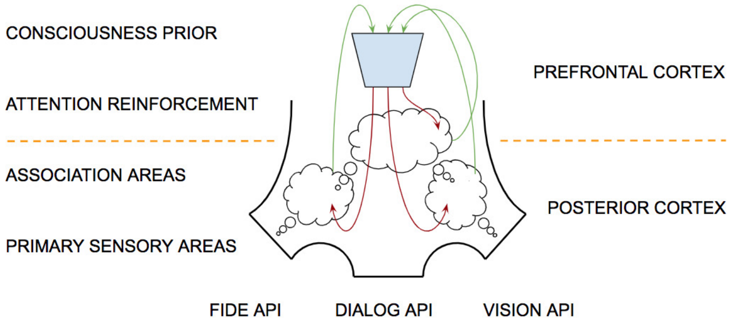

Figure 7: The basic capabilities required to support conscious awareness can be realized in a relatively simple computational architecture that represents the apprentice's global workspace and incorporates a model of attention that surveys activity throughout somatosensory and motor cortex, identifies the activity relevant to the current focus of attention and then maintains this state of activity so that it can readily be utilized in problem solving. In the case of the apprentice, new information is ingested into the model at the system interface, including dialog in the form of text, visual information in the form of editor screen images, and a collection of programming-related signals originating from a suite of software development tools. Single-modality sensory information feeds into multimodal association areas to create rich abstract representations. Attentional networks in the prefrontal cortex take as input activations occurring throughout the posterior cortex. These networks are trained by reinforcement learning to identify areas of activity worth attending to and the learned policy selects a set of these areas to attend to and sustain. This attentional process is guided by a prior that prefers low-dimensional thought vectors corresponding to statements about the world that are either true, highly probable or very useful for making decisions. Humans can sustain only a few such activations at a time. The apprentice need not be so constrained.

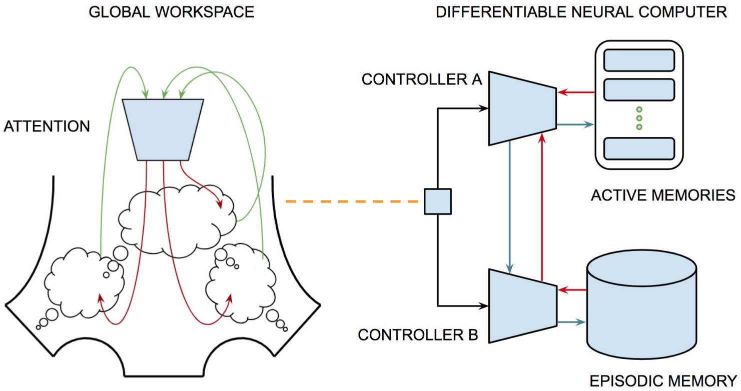

Figure 8: You can think of the episodic memory encoded in the hippocampus and entorhinal cortex as RAM and the actively maintained memories in the prefrontal cortex as the contents of registers in a conventional von Neumann architecture. Since the activated memories have different temporal characteristics and functional relationships with the contents of the global workspace, we implement them as two separate NTM memory systems each with its own special-purpose controller. Actively maintained information highlighted in the global workspace is used to generate keys for retrieving relevant memories that augment the highlighted activations. While the associative keys required to access locations only partially match locations, they can can still be used to guide attention allowing the NTM to recognize and even partially merge related locations. In general, locations in memory correspond to thought vectors that can be composed with other thought vectors to shape the global context for interpretation.

The global workspace summarizes recent experience in terms of sensory input, its integration, abstraction and inferred relevance to the context in which the underlying information was acquired. To exploit the knowledge encapsulated in such experience, the apprentice must identify and make available relevant experience. The apprentice's experience is encoded as tuples in an NTM that supports associative recall. We'll ignore the details of the encoding process to focus on how episodic memory is organized, searched and applied to solving problems.

3.3 Action Selection

In both neuroscience and artifial intelligence, reinforcement learning problems are typically modeled as Markov decision problems (MDPs). While MDPs can be solved in polynomial time, the size of the state space is often prohibitively large, making practical solution intractable [**?**]. Hierarchical reinforcement learning offers a means of reducing the computational burden by decomposing the state space resulting in a relatively small number of tractable MDPs each of which can be solved independently [**?**, **?**, **?**]. However, the problem of finding an optimal decomposition is itself intractable and hence it is necessary to resort heuristic methods and approximate solutions.

There exist a number of approaches that develop solutions to the problem of hierarchical reinforcement learning (HRL) employing various decomposition strategies, several of which we draw inspiration from [**?**, **?**, **?**, **?**, **?**, **?**, **?**] including a few that relate to biological or biologically plausible models [**?**, **?**, **?**, **?**]. It's important to keep in mind that we are dealing a partially-observable, high-dimensional, continuous state space, and an action space that includes abstract cognitive activities in addition to concrete physical activities that engage the motor system in interacting with the environment.

In the treatment here, we emphasize the problem of life-long learning as it relates to the non-stationarity of underlying process as a consequence of changes in the external environment and changes in the goals of the agent and the neural substrate available for computation during development and extending on into adulthood. In the case of a growing infant, the changes involve the appearance and maturation of critical circuits and the limitations of finite memory. In both human and machine, internal representations progress from concrete to abstract, building on a foundation grounded in the environment. This maturation in cognitive capability is accelerated by a curriculum that takes advantage of dependencies between concepts.

The network model shown in Figure 9 illustrates a system that takes as input a pattern of neural activity originating in the medial temporal and inferior parietal cortex and selects an action to perform. This particular example is meant to illustrate how episodic memory might play an expanded role in action selection. For illustration, patterns of activity serve as proxies for the state of the external environment and are represented in the figure as a sequence $s_t, s_{t-1}, s_{t-2}, \ldots$. The subnetworks labeled A and B are relatively straightforward multilayer neural networks that compute features and generate representations as their output. Network A takes as input a representation of the current state, and generates a representation of the *context* for action selection.

We'll explain the function of the box labeled M in a moment; assume for now that it generates a representation of the options available for acting in the current context. Network B then takes these suggestions as input and produces as output a representation of the selected action. The boxes labeled C, D, E and F are controllers for two differentiable neural computer (DNC) units that provide storage and access for short-term and long-term memory respectively. The controllers on the left are part of the online system for selecting actions. The controllers on the right are responsible for off-line training during which the recorded actions, along with their associated states and rewards are consolidated in long-term memory using experience replay.

The blue boxes represent stored information in the form of key-value pairs. Each key is associated

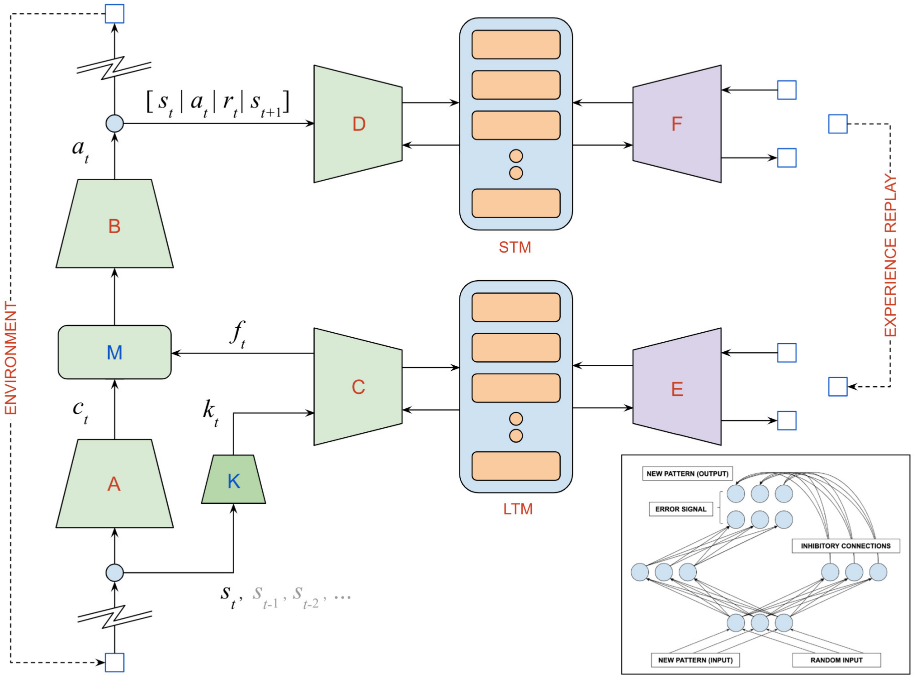

Figure 9: The network shown here takes as input a pattern of activation originating in the temporal and parietal lobes and selects an action to perform. The subnetworks labeled A and B are relatively straightforward multilayer neural networks that compute features and generate representations as their output. Network A takes as input a representation of the current state, and generates a representation of the context for action selection. Network K is an embedding network that takes as input a sequence of states corresponding to recent activity and generates as output a unique key associated with a subspace of the full MDP state space that includes the current state. The box labeled M corresponds to a location in working memory. The networks C, D, E and F are controllers for two differentiable neural computer (DNC) peripherals that provide storage and access for short-term and long-term memory respectively. The long-term memory is used to store the weights for networks that encode architecturally identical networks — only the weights are different — providing specialized expertise in restricted domains corresponding to subspaces of the full MDP state space. The model operates in two modes. In each cycle during the *online mode*, the C controller loads the selected expert network into location M where it is fed the output of A and produces the input to B. In this mode, the short-term memory is used to record activity traces that are subsequently used in the *offline* mode to update the networks stored in long-term memory. The network in the lower-right inset implements a version of *pseudo rehearsal* as a means of mitigating catastrophic forgetting [**?**].

with a subset or *subspace* of the set of all states that represents a restricted domain of expertise for selecting actions. The value for this key is a function implemented as a network trained as an expert for the associated subspace. K is an embedding network that takes as input a sequence of states corresponding to recent activity and generates as output a unique key associated with a subspace of the current state. A given state can belong to more than one subspace and the particular key selected at any given point in time depends on the current state and the immediately previous states in a fixed window. The order of the states matters.

In the online phase, the embedding network retrieves this key which it forwards to the controller labeled C that uses it to retrieve the expert for the relevant subspace. The box labeled M corresponds to a location in working memory and in each online cycle the C controller loads the expert subsystem in location M of working memory where it can be utilized to compute a set of options appropriate for the current state. During off-line periods the system uses the recorded sequences of activity to run some variant of experience replay to update the relevant expert subsystems stored in long term memory [?, ?, ?].

The training that occurs offline involves adjusting the weights of networks using relatively small samples and so runs the risk of catastrophic interference in transfer learning [?]. One way in which we hope to ameliorate the adverse consequences of catastrophic interference is by defining separate networks for separate subspaces. The embedding space method mentioned in describing K is designed to isolate expertise by identifying states that tend to occur together. The hope is that the actions exercised is such states will tend to be interrelated and hence they should be represented using the same network to facilitate their coordination.

Of course temporal proximity in their occurrence doesn't guarantee they serve the same task since we are always getting distracted or interrupted requiring us to interleave tasks that have very little to do with one another. It may be possible to segment activity streams into coherent tasks in a similar way to how we segment conversations involving multiple speakers [?, ?]. Alternatively, there has been some success with the method of *pseudo rehearsal* which consists of retraining existing networks by interleaving new examples with synthetic-examples produced by randomly activating the existing network [?, ?, ?, ?, ?, ?].

In this model the STM roughly corresponds to the hippocampus as the storage system for episodic memory. The LTM resembles the cerebellum in the way that it essentially compiles prior activity to construct a set of programs each of which spans some portion of the overall state space. As described above, the STM is only used for temporary storage awaiting off-line replay to consolidate recent memories. An alternative is to maintain a much larger collection of episodic memories that can be used in a manner similar to that suggested in Gershman and Daw who posit that we routinely draw upon our stored memories in the hippocampus to figure out what to do in novel situations not covered by our other sources of procedural knowledge [?]. See Box B for more detail concerning episodic memory and experience replay.

Box B: Replaying Experience, Consolidating Memory

When we encounter a new experience in the environment, we do not act independently of the past, but rather, our past experiences substantially inform our present decisions. Here, we introduce the basic principles of *hippocampal replay* and a few key ways in which it has motivated reinforcement learning algorithms.

Replay is the process by which hippocampal representations of previous experiences are sequentially reactivated [?]. Studies show that cells in the rodent hippocampus replay past experiences to stabilize behaviorally relevant memories [?, ?]. Though initially observed in spatial tasks, recent work suggests that non-spatial task states are also replayed, and that this phenomenon is common in humans [?].

In the reinforcement learning literature, the *experience replay* algorithm was introduced as an analogical framework in online learning agents [?]. Transitions containing state, action, and reward information are sequentially stored in memory and sampled randomly for learning. Randomly replaying old memories not only allows decorrelation of consecutive experiences encountered during data collection, but also enables reuse of training data, increasing sample efficiency, and encourages resampling of rare experiences, potentially alleviating forgetting.

A relatively well-studied question is *what to replay*. Some studies suggest the correlation of replay frequency with *novelty* approximated by temporal difference (TD) error [?], and others with high *reward* [?]. In particular, dopaminergic release, which encodes both novelty and reward [?], enhances *sharp wave-ripple* activation — the basic unit of replay. Yet other studies show that experiences more *vulnerable to forgetting* are more likely to be replayed [?]. While the exact selection algorithm is unknown, the observed association with novelty inspired the *prioritized experience replay* algorithm which samples experiences with probabilities weighted by their TD errors and is now consistently preferred to the originally proposed uniform sampling variant [?].

The significantly less studied question is *what happens during replay*. Besides re-learning of experiences, many neuroscientists support the idea that replay also serves as a substrate for *memory consolidation* — the gradual integration of new experiences processed into existing knowledge representations in the neocortex [?, ?, ?, ?, ?], as to stabilize memories against interference. The idea is that replaying information stored in memory will encourage synaptic consolidation processes.

While we lack a precise understanding of the underlying mechanisms of consolidation in the brain, in our architecture we frame consolidation as the process by which experiences are used to update expert subsystems stored in long-term memory. We propose an adaptive replay algorithm whereby experiences with contexts similar to the current context are replayed and thus preferentially consolidated into long-term storage. Since action selection directly depends on the relevant expert network drawn from long-term memory, we can ensure to maximally update the currently relevant expert network with existing memories related to its corresponding context. This algorithm is partly inspired by the result by [?] whereby they observed that when a rat pauses at a branching point in a maze, it replays representations of trajectories in the past with similar context to drive its present decision-making.

There exist many other cognitively inspired variants of experience replay. One example is *hindsight experience replay* [?], where the agent pretends that whatever state it reaches had been the goal state from the start and learns from the experience regardless of whether it actually succeeded, just as humans can learn from undesirable outcomes.

In this case, the LTM stores what can be thought of as subroutines or libraries for solving routine problems. Used in the manner described in Gershman and Daw [?], the DNC labeled STM more closely captures the functionality of the hippocampus in combining short-term and long-term episodic memories with specific procedural knowledge based on past experience that may or may not be common enough to warrant compiling as a standalone library. The dentate gyrus is best known for its ability to separate patterns to avoid mistaking one pattern for another. Less well understood is a possible complementary role that involves integrating similar patterns.

The ability to draw upon episodic memory to select what to do in situations similar to those encountered in the past provides a simple form of one-shot learning. It could enable us to make predictions, perform hypothetical reasoning and put ourselves in someone else's shoes assuming that our ability to retrieve memories allows us match situations that we find ourselves that we haven't experienced, but know from someone else's experience. It might avoid some of the problems with interference if the process of integrating new procedural knowledge with old could be spread out over longer periods if, say, each time you encounter a similar situation you make only minor adjustments to the weights of the associated subspace expert network.

3.4 Executive Control

There is a growing consensus and a fair bit of evidence to support the hypothesis that the human frontal cortex is in charge of executive control, goal-directed planning and abstract thinking. There are differences in opinion about how these cognitive processes are implemented and how they coordinate their activities with that of the rest of the brain. One thing that seems clear is that the frontal cortex and in particular the prefrontal cortex employs many of the same strategies as do networks elsewhere in the brain, both cortical and subcortical.

In particular, circuits in the prefrontal cortex recapitulate the coarse-to-fine, concrete-to-abstract feature hierarchies that we see in the sensory, motor and somatosensory cortex. They exhibit the profuse reciprocal recurrent connections between levels of abstraction that enable us to generalize on the basis of relatively small amounts of information, learn to make accurate predictions in an unsupervised manner depending on observations and interactions with the environment to ground our conclusions, and that provide the foundation for constructing a rich repertoire of representations that serve decision-making.

The neural correlates of abstract thinking, including the circuits that enable us to solve practical problems as well as pursue pure mathematics, are generally agreed to be located in the prefrontal cortex with reciprocal connections throughout the rest of the cerebral cortex, the cerebellar cortex and subcortical regions including the basal ganglia, hippocampal formation and parts of the limbic system involved with emotion, motivation and episodic memory. See Box C for more detail regarding abstraction, hierarchy and executive oversight in the prefrontal cortex.

Box C: Hierarchy, Abstraction and Executive Control

The prefrontal cortex (PFC) is generally considered to be responsible for executive cognitive control and enabling the synthesis of novel behavior. Here, we briefly review prefrontal anatomy, development and physiology, focusing on three key executive cognitive functions: *attentional set*, *working memory* and *action selection*. For each function, we suggest how our current understanding might lead to new architectures and algorithms for AI systems.

The PFC sits atop a group of hierarchically organized sensory and motor areas in the cortex enforced through reciprocal anatomical connections [?]. This arrangement, referred to as *Fuster's hierarchy*, motivates computational models of the PFC that posit the development of highly abstract representations of the sensorimotor context that can be used understand what we perceive and direct how we act [?]. In addition to connections between layers, Fuster's hierarchy stipulates reciprocal connections between sensory and motor areas of cortex at the same level of abstraction within each layer of the hierarchy. This intralayer connectivity between perception and motor suggests that action representations feed back into and enhance perception, a principle codified in the notion of *corollary discharge* [?].

Beyond the model of network interactions, intralayer connectivity in Fuster's hierarchy suggests that each layer of the hierarchy along with the layers below but excluding those above, forms a self-contained perception-action loop. Evidently the neocortex undergoes a series of developmental stages with the PFC among the last areas to mature [?]. This implies training of a complex agent may need to unfold in a manner akin to greedy layer-wise deep network training [?, ?], with developmentally-staged, abstraction-comparable, layer-wise learning of the coupled sensorimotor features.

Attentional set (ASET) refers to the preparation of downstream perception and motor cortices for expected stimuli or action. ASET is exhibited in *cued-attention tasks* where, in anticipating a visual stimuli, PFC and V4 will be active before the stimuli is given [?]. ASET suggests existing inhibitory attention masks may be augmented with additive excitatory attention, allowing a neural network to reduce bottom-up input needed for neuron stimulation or cause neuron firing in the absence of sensory input altogether. Allowing the controller to generate new patterns via network activation even suggests a new model of imagination, with improvements in both sensory synthesis [?] and planning [?].

Working memory is the maintenance of recent stimuli for subsequent action planning. Working memory consists of groups of coupled neural circuits in the PFC called *stripes* that are connected to potential target stimuli in sensory cortex and access controlled by circuits in the basal ganglia. Computational models of working memory [?] include implementations similar to the recurrent memory circuit of an LSTM cell [?], more exotic architectures involving stacked LSTMs [?] and multiple memory stripes manipulated by a central controller.

In *action selection*, the PFC generates many actions that are approved or denied by the basal ganglia; both basal ganglia and orbitomedial PFC receive dopaminergic afferents originating in midbrain structures, providing a reward signal that reinforces learning [?]. Computational models of dopaminergic systems [?] point to an architecture similar to existing actor-critic models [?]; a key improvement is the modeling of *reward inhibition*, whereby learning ceases for repetitive stimuli. Suppression of reward to prevent response overfitting could aid in tackling other problems such as reward hacking [?], catastrophic forgetting, and lifelong learning [?], all challenges in effectively managing the learning process.

With respect to hierarchical goal-based planning, there is growing evidence pointing to a set of adjoining regions in the prefrontal cortex that are responsible for how abstract plans are initially selected, subsequently refined and finally realized as concrete actions. These same regions also appear to be involved in relational reasoning from simple binary relations to higher-order relationships.

These theoretical observations combined with behavioral studies and fMRI recordings have led to a number of computational models of hierarchical planning that exhibit similar patterns of cognitive activity. In particular, cognitive neuroscientists have developed models of how such abstract hierarchical reasoning in the prefrontal cortex is related to what we know about how the basal ganglia and areas of the limbic system involved in motivation contribute to action selection [?].

The network shown on the right in Figure 10 consists of three subnetworks that roughly align with the lateral frontal polar cortex (bottom), dorsolateral prefrontal cortex (middle) and anterior premotor cortex (top) as shown in the figure. Each of the subnetworks is composed of three elements: a recurrent multilayer perceptron constructed of interleaved convolutional and max-pooling layers shown in orange, a multilayer attention network shown in green and a masking layer in blue that

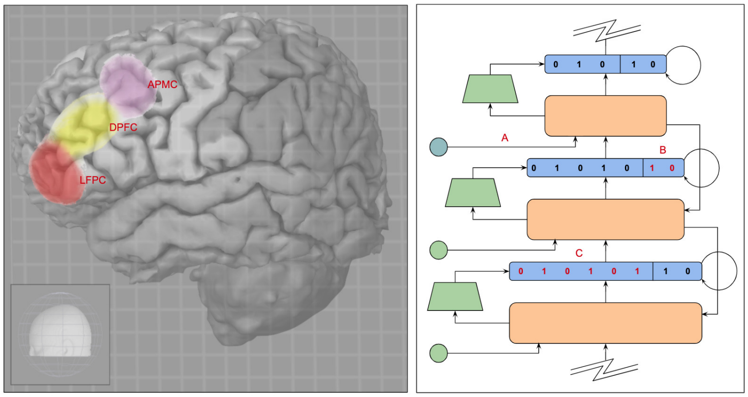

Figure 10: The panel on the left highlights three areas of the prefrontal cortex shown in the figure from left to right (rostro-caudal) and referred to in the text as the *lateral frontal polar cortex* (LFPC) *dorsolateral prefrontal cortex* (DPFC) and *anterior premotor cortex* (APMC). According to the theory first articulated by Joaquín Fuster and subsequently refined David Badre [**?**], Mark D'Esposito [**?**] and Etienne Koechlin *et al* [**?**] and their colleagues, as actions are specified from abstract plans to concrete responses, progressively posterior regions of lateral frontal cortex are responsible for integrating more concrete information over more proximate time intervals. This process of progressive articulation does not correspond to different stages of execution so much as to how actions are selected, maintained and inhibited at multiple levels of abstraction [**?**]. The panel on the right shows a simple neural-network model of the brain regions aligned with the rostro-caudal axis of the frontal cortex and hypothesized to account for how action representations are selected, maintained and inhibited at multiple levels of abstraction. The neural-network model is described in more detail in the main text, but a few points are in order here: A — different abstraction layers may include input from other sources, e.g., natural language embeddings, that are only required at particular levels of abstraction; B — each recurrent level of the abstraction hierarchy includes state variables encoding information that would typically appear on the call stack in a conventional computer architecture; C — attentional layers mask (suppress) input that is not determined to be relevant to decision making at a given time and level of abstraction resulting in a sparse context vector.

selectively suppresses a subset of the outputs of the convolutional stack in accordance with the output of the attention network.

Input to each of the three subnetworks includes areas of associative activity throughout the sensory and motor cortex as well as areas corresponding to higher-level abstractions located in the frontal cortex responsible for abstract thought and subcortical regions responsible for motivation. While not emphasized here, the active maintenance in working memory of information originating from these sources is critical for the cognitive activities that these networks support [**?**, **?**]. The outputs are fed to a network (not shown) that serves as the interface for the peripheral motor system (the fully instrumented integrated development environment (FIDE) in the case of the programmer's apprentice) which could play the role of the basal ganglia and cerebellum, but could also be considerably simpler depending on the application.

Figure 10 is just a sketch employing familiar neural network components to make the point that building these architectures out of standard components is not the most significant challenge. The real challenge is in training them as part of larger system with lots of moving parts. The expectation here, as in the model sketched in Figure 9, is that end-to-end training with stochastic gradient descent isn't going to work, and that training will likely require some form of layer-by-layer developmentally-staged curriculum learning [**?**, **?**, **?**, **?**] and a strategy for holding some weights fixed while adjusting other weights to account for new information and avoid problems like catastrophic forgetting.