9. What is distinctive about neocortex?

As we noted at the outset, the neocortex is the brain area that has expanded the most in primates, including humans. It is clearly important for human life, especially for those activities that humans distinctively perform. But, as we have stressed through this Element, other brain structures are also important. With few exceptions, the neocortex does not take over their activities but supplements them. The relevant question is what is the distinctive type of processing that occurs in the neocortex. One suggestion comes from the studies of decorticate cats discussed in section 5.4. While cats in which the neocortex is removed can live in protected environments, they would be unlikely to fare well in the world in which they confront variable conditions, including predators. Based on these studies, Buchwald and Brown (1973) proposed that the neocortex serves for detailed analysis of stimuli, extracting and representing complex and subtle information about an organism's environment and identifying relations between different bits of information. Such information is extremely useful in solving problems posed by a variable environment. In this section we investigate how the neocortex can perform these tasks.

9.1 (Artificial) neural networks and pattern extraction

The neocortex is organized in a distinctive manner that supports the hypothesis that it acts to extract subtle and complex information from sensory inputs. While many brain areas, including both the basal ganglia and the hypothalamus, are organized as interconnected nuclei, the neocortex is laid out much more systematically. As we discussed in section 2.4, Brodmann (1909/1994) differentiated areas within the neocortex based on the thickness of layers identified in stained cortical tissue. Tracing axons from neurons in one area reveals that they mostly project to selected neurons in specific other areas, resulting in relatively orderly anatomical hierarchies such as shown in Figure 14B. At the top of the figure are areas in the temporal and parietal lobe. Both streams, however, continue into the frontal cortex, reaching the far frontal area known as the *prefrontal cortex*, on which we will focus in section 9.4.

To see how such a (anatomically) hierarchically organized network could enable the extraction of information, consider *artificial neural networks* (ANNs)—computational systems that were inspired by the architecture of the neocortex. As illustrated in Figure 23, these networks consist of layers of artificial neurons, commonly referred to as *units*. A weighted connection links a unit in one layer to units in the next higher layer; in processing, the weight is multiplied by the activity value of the unit in the lower layer to determine an input to the higher-level unit. Each higher-level unit accumulates these inputs and applies a nonlinear mathematical operation to determine its activity value. Such a network will generate output activity from values supplied on its input layer and can be trained to generate desired outputs for different inputs. A common way to train ANNs (known as *backpropagation*) is to let the network generate an output from whatever weights it has and then to apply an algorithm to gradually change weights so as to reduce the difference between the actual and desired output. Over multiple iterations of training, such networks can learn to respond similarly to different instances of the patterns. For example, a network can learn to recognize pictures containing different species of

dogs. As a result, they are often characterized as *recognizing patterns*. When successful, these networks can generalize and recognize patterns when tested with novel stimuli (Bechtel & Abrahamsen, 2002; Buckner & Garson, 2019). In recent years, researchers have developed *deep-learning networks* that employ numerous layers of units between the input and output, with the weights on different layers of connections each able to be adjusted during learning to achieve better performance (Sejnowski, 2018). An intriguing finding is that when deep learning networks have been deployed to model processing of visual stimuli, they end up acquiring an organization of nodes that is similar to that found in the human visual system (Yamins & DiCarlo, 2016).

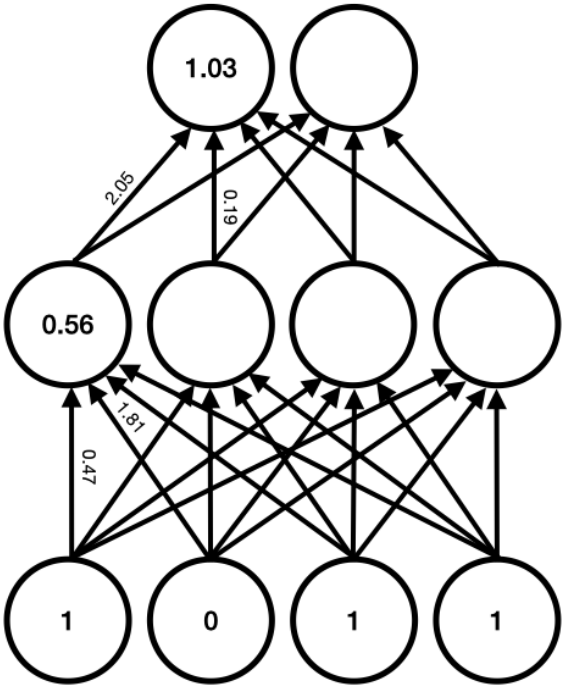

Figure 23. Simple artificial neural network. Activation values for the input units at the bottom are multiplied by weights on the connections (indicated by arrows) to determine the activation of units in higher layers. Example weights and activation values shown.

ANNs are powerful systems for recognizing patterns. Since the task of vision is to extract patterns in visual stimuli, it is not surprising that perceptual processing areas in the brain are organized in the same manner. It is also easy to see how an ANN-style architecture can implement motor control: allow inputs to encode a high-level description of an action and the network can be trained to generate outputs that implement the specific motor activities required. One can extend pattern recognition beyond perceptual and motor processing to more clearly cognitive tasks that involve a sequence of inferences. Each layer in a network can be viewed as making an inference based on inputs from the previous layer (e.g., infer that an object is a bird), providing the next layer an input from which it can make a further inference (e.g., that it can fly). Accordingly, ANNs are widely used to perform reasoning and problem-solving tasks, and the fact that the neocortex is organized in a sequence of connected layers suggests that it carries out reasoning and problem-solving activities in a similar manner.

9.2 The challenge of explaining the systematicity of thinking

When theorists advanced ANNs as models of human cognitive processing in the 1980s, they were confronted by a host of objections, one of which is that ANNs cannot account for what is termed the *systematicity* of human thought. Systematicity is exemplified in arguments used to establish conclusions in such fields as mathematics, law, and science. Consider the argument:

(1) A dog is a color
(2) A color is a musical composition
(3) Therefore, a dog is a musical composition.

Even though the premises do not make sense and the conclusion is false, logicians consider the argument to be valid: if dogs were colors and colors were musical compositions, the conclusion would have to be true. It is for this reason that valid arguments serve to establish conclusions from accepted premises. What makes the argument valid is not the meaning of its words but how they are related: any argument in which the premises and the conclusion exhibit the same relations is valid. The importance of such structure (referred to as *syntax*) extends to language generally. Knowing the syntax of a language enables you to construct and understand an indefinite number of sentences with the same structure.

In the era before researchers started to invoke ANNs to explain cognitive activities, most researchers assumed that cognition worked much in the manner of logical arguments: an individual was assumed to encode thoughts in structured representations and apply rules that depended on their structure (syntax) to develop new thoughts. This ensured the systematicity of thinking. Since ANNs don't apply rules to structured representations, many theorists, including many philosophers (Fodor & Pylyshyn, 1988), argued they could not account for thinking (see Buckner & Garson, 2019, section 7).

Proponents of ANNs have advanced several responses to this challenge. One is to treat the structures exhibited in thought as patterns to be learned by a neural network realized in the neocortex. Figure 14B represents the visual processing system much like a multi-layer artificial network. Each brain region shown at higher-levels extracts additional patterns from the patterns recognized at lower-levels. Brain areas in the central and anterior inferotemporal cortex, at the top of the ventral stream (shown on the right of Figure 14B and labeled *inferotemporal stream*), respond to patterns corresponding to abstract categories such as shape, color, or faces. Some areas in this region also respond differentially to categories of objects (e.g., dogs, houses). Barsalou (2008) has suggested how this process of identifying more abstract patterns from more concrete ones can be extended to relational categories such as "on top of" or "a type of."[1] A further challenge is to explain the ability to connect the states in the network that represent these categories in flexible (and sometimes arbitrary) ways to

[1] An important part of Barsalou's project is to show that the categories we use in thinking, including abstract ones, are perceptually grounded.

capture systematic relations as illustrated in the above example of a valid but nonsensical argument. O'Reilly et al. (2014) has crafted an ANNs that can represent many such relations and employs a simulation of the basal ganglia (section 5.4) to enable flexible combination of these relations with other mental concepts.

To date, models of how the neocortex can implement systematic thought are hypothetical proposals, not grounded in details of neural activity or connectivity. What they show is that it is possible for a structure like the neocortex, assisted by the basal ganglia, to produce systematic cognition. A couple of considerations should be kept in mind, however: human thinking is not perfectly systematic (we make inferential errors) and we often employ other types of reasoning, such as reasoning by analogy and metaphor. Further, many animals that are generally not thought to engage in high-level cognitive reasoning also have an extensive neocortex. The ways in which humans use the neocortex, especially the prefrontal cortex, may reflect in part the cultures in which we live and modes of learning supported by those cultures. One thing these cultures make available are languages, which are themselves powerful representational tools both for logical reasoning as well as analogical and metaphoric reasoning. The networks in our brains, especially those in the neocortex, have learned to accommodate the structures available in the languages we acquire, and this may be a significant part of the explanation of our ability to engage in systematic thinking.

9.3 Differences between the neocortex and artificial neural networks

ANNs provide a powerful framework for modeling important features of the neocortex, but there are significant respects in which the neocortex is different from ANNs. First, in most ANNs, processing connections are only feedforward. Error is propagated backwards during learning, but not in processing. Yet, in the neocortex there are at least as many *recurrent projections*, projections from anatomically higher-level processing areas to sensory inputs, as forward projections. Although the full function of these recurrent projections is not understood, they allow a response in a higher-processing region, activated by whatever means, to activate other patterns in the lower-level input areas that are frequently associated with that response.

The recurrent activation of lower layers from higher layers provides an explanation of what is referred to as *top-down* processing, according to which the concepts one applies to stimuli affects how one sees them. These is compelling evidence that we engage in such processing. In a classic experiment, Bruner and Postman (1949) flashed playing cards to participants and asked them to name them. Among the cards were abnormal cards, such as a red four of spades. Participants would regularly report a normal card, e.g., a four of hearts, although sometimes noting that something seemed to be wrong with the card (but unable to say what). Feedback from higher-visual areas on earlier-visual areas overrides the input from the senses. It can also explain abilities such as visually imagining a bird when one hears the word "bird" (Kosslyn, 1994) or reporting seeing features of a bird that were not visible in a given presentation.

The prevalence of recurrent projections in the neocortex has led some neuroscientists (Friston, 2010) and philosophers (Clark, 2013) to advance *predictive coding*, an account of neural processing that reverses the more traditional account that starts from activation of the senses and proceeds to recognition of objects. Instead, these theorists propose that higher processing areas make predictions about subsequent sensory input. For example, if one looks down after viewing a person's face, one expects that one will see a human torso. If the prediction is true, no sensory information is processed and the neural mechanism increases its confidence in making such a prediction in the future. But if the sensory input violates the prediction—one sees the torso of a bear—sensory information is processed further. If violations of expectations are frequent enough, one learns from them and makes different predictions in the future.

A second feature distinguishes processing in the neocortex from that in ANNs: all regions in the neocortex are also interconnected with nuclei in the thalamus and the basal ganglia (section 5.4), both receiving inputs from them and sending outputs to them. In many cases, these projections form loops that have functional significance. For example, as we have noted in section 6.3, negative (also positive) feedback loops generate oscillations such as those registered with EEG (section 3.3). The result is that there are ongoing oscillations in the neocortex at many different frequencies (Buzsáki, 2006). These affect how information is processed (discussed further in Bechtel, 2019). As just one example, when subthreshold oscillations in two different brain regions are synchronized, inputs from one region are more likely to generate action potentials in the other region. Recently, researchers have identified traveling waves—patterns of oscillation that move from region to region in the neocortex—and advanced evidence that these modulate such things as sensitivity to perceptual stimuli (Davis, Muller, Martinez-Trujillo, Sejnowski, & Reynolds, 2020). Another example involves loops that include the basal ganglia, which may be particularly important in controlling processing in the neocortex. As we discussed in section 5.4, the basal ganglia by default inhibit other brain regions. As a result of loops with the thalamus and neocortex, the basal ganglia decide which activity in the neocortex is released from inhibition and allowed to continue.

9.4 Cognitive control and the prefrontal cortex

An important feature of human cognition is what is referred to as *cognitive control*: the ability to resist habitual or emotionally salient behaviors in order to act in more context-appropriate ways (often ways that are expected to be more beneficial in the long term) (Miller & Cohen, 2001). For example, when a European or North American drives in Thailand, she needs to resist her habitual responses to drive on the right side and follow instead Thailand's rule of driving on the left. As another example, if we promised our best friend that we'd show up for his party at 11pm, we might have to suppress our physical and emotional exhaustion to honor our promise. Such activities are common among humans. Even when not required to do so, children often share their candies fairly with their friends, suppressing their desire to eat more themselves.

Cognitive control draws upon the processing capacities of a part of the neocortex that we haven't discussed much so far, the prefrontal cortex, which occupies the front part of the frontal lobe (the rear portion primarily contains areas involved in processing motor commands).

Like posterior areas of the brain involved in vision, the prefrontal areas comprise multiple different processing areas that have been associated, largely through single-cell recording studies in monkeys and neuroimaging studies in humans (see section 3.3), with a variety of capacities. These are organized into processing streams that extend those involved in visual processing shown in Figure 14B. The dorsal "where" or action-oriented stream gives rise to areas that represent actions, often complex actions, and the individual's evaluations of those actions. As one moves forward in the prefrontal cortex, the areas first encountered code for learned associations between sensory stimuli and motor responses. Areas yet further forward code more abstract rules between contexts and classes of actions, including social and moral norms (Carlson & Crockett, 2018) and facilitate thinking about hypothetical actions and future states of the world. In contrast, the continuation of the ventral stream represents increasingly abstract features of objects as well as evaluations of these objects.

Recurrent projections are especially prevalent in the prefrontal cortex. These enable individual areas to maintain active states for prolonged periods after the initiating stimulus has ceased. This provides for a form of temporary memory known as *working memory*, which is taken to be particularly important for carrying out complex actions or actions after brief delays. Goldman-Rakic (1995) demonstrated how animals could retain information in these circuits until needed to perform the action. Fuster, Bodner, and Kroger (2000) emphasize how these circuits support the integration of information from different modalities needed for temporal structuring of behavior. However, the active maintenance of information needs to be coupled with flexible and appropriate update of information to be adaptive—when a social rule is no longer appropriate in guiding behaviors in a particular context, the cognitive system needs to be able to shut down the associated neural activities and replace them with ones that represent the currently appropriate social rules. Interconnections between prefrontal areas and the basal ganglia are important in doing this (section 5.4).

These capacities, especially the ability of prefrontal areas to encode and maintain active representations of goals, rules, and values, suggest how cognitive control is possible. As these areas are connected to other brain areas, these representations can, for example, activate the relevant sensory inputs, memory representations, and motor outputs needed to perform context-appropriate actions while inhibiting actions that might be triggered directly by sensory stimuli (Miller & Cohen, 2001).

We should note a few features to this account of cognitive control. First, cognitive control involves multiple levels in the various hierarchies in the neocortex: a lower-level motor response can be controlled by contextual information at a higher-level, which can in turn be modulated by superordinate contextual information. For example, children are often told to speak in a softer voice while indoors. They often learn quickly, though, that this rule only applies with the presence of adult supervision. The context of having an adult nearby, then, further contextualizes the indoor rule (Badre & Nee, 2018). Second, in addition to top-down control signals travelling from anatomically higher-levels to lower-levels in a given information processing hierarchy, recent empirical literature reveals control signals that are bottom-up and lateral, enabling controllers across different levels and hierarchies to constrain each other

(Cisek & Kalaska, 2010). We turn to violations of hierarchical control organization in the next section.

9.5 Summary

The cortex, especially the neocortex, is organized differently than the rest of the brain. Within the neocortex there is a hierarchy of processing areas in which neurons in one area project to those in subsequent ones. ANNs, modeled on this pattern of organization, suggest that the neocortex is a powerful pattern recognition system. We sketched how this structure can account for the systematicity exhibited in human cognition. We also emphasized the importance of recurrent projections in the neocortex and the connections of regions throughout the neocortex with subcortical structures, especially the thalamus and basal ganglia, and sketched how these enable humans to exert cognitive control.