Chapter 29

Mechanistic Explanation in Neuroscience[1]

Catherine Stinson and Jacqueline Sullivan[2]

Abstract. In this chapter, we describe several representative epochs in the history of neuroscience in which understandings of the mechanisms of learning and memory phenomena were advanced. We use these historical developments to highlight unique features of mechanistic explanations in neuroscience and to explore some of the challenges associated with providing mechanistic explanations of cognitive phenomena.

1. Introduction

Perhaps the most striking thing you notice when thumbing through the pages of a neuroscience textbook like *Principles of Neural Science* (Kandel et al. 2012) are the elaborate diagrams of the central nervous system, brain, spinal cord, synapses, neurons and molecules. It is equally striking that whatever topic you look at, whether it be the action potential, synaptic transmission, cognition or perception, it is inevitably described and explained using the word "mechanism." You can become so accustomed to seeing and hearing about mechanisms in neuroscience that it never occurs to you to question what mechanisms in fact are, or what that choice of terminology implies. As it turns out, there are tricky philosophical problems lurking beneath the surface of mechanism talk in neuroscience.

In this chapter we explore some of the ways that mechanisms are invoked in neuroscience, and look at a selection of the philosophical problems that arise when trying to understand mechanistic explanations (several chapters in this volume go into more

detail about particular philosophical problems encountered in mechanistic explanation in neuroscience, and Chapter 6 describes in greater detail some of the history we discuss below).

We begin in Section 2 by introducing a series of historical case studies that illustrate how neuroscientists have depended on mechanistic metaphors in their efforts to understand the mind and brain, and how their mechanistic explanations have developed over time. We revisit these examples throughout the remainder of the paper. In Section 3, we use these case studies to highlight what contemporary philosophers have identified as the fundamental features of mechanisms and mechanistic explanation. In Section 4, we consider some of the methodological issues that arise in neuroscience including (1) how to integrate psychological with neural models (2) how to generalize findings in model organisms like the sea slug *Aplysia* to human learning and memory, and (3) whether to favor top-down or bottom-up methods.

2. A Short History of Neural Mechanisms of Learning and Memory

The historical examples we focus on are episodes in the search for the neural mechanisms of learning and memory, which are among the most important cognitive phenomena studied in neuroscience. These examples illustrate how mechanisms are discovered, reasoned about, represented, and how they figure in explanations.

In the 17th century, in *Treatise on Man*, the French philosopher René Descartes likened human beings to machines. Descartes drew an analogy between the movements of human beings and the movements of the automated figures in the fountains of the Royal Gardens at St. Germain. Descartes described how when visitors to the gardens step

on certain tiles, statues of Roman Gods, Goddesses and other mythical creatures move, gesture, play music, spray water, and speak. Pressing on the tiles triggers a flow of water from storage tanks beneath the fountains through a network of hidden pipes. The flow of water then causes the figures, which are connected to machinery like springs and cogs, to move. Descartes also compared these motions to those of a clock or a mill, which can be made to move continuously, not just in response to an external push.

Descartes claimed that a similar set of events takes place in the human nervous system when simple reflexes are triggered. The brain, according to Descartes, contained ventricles filled with "animal spirits" or "a very fine air or wind," which reached the ventricles via the blood (Descartes 1664/1985, 100). He believed that the ventricles were connected to networks of nerves, which he thought were mostly hollow save for a set of small fibers running their length.

According to Descartes, the nerves are connected to the brain in such a way that stimulation from the periphery, which tugs on the fibers, is communicated to the brain, triggering a response. Tugging on the fibers opens pores in the nerve, allowing animal spirits to flow from the ventricles through the nerve to the musculature, causing motion, he claimed. Descartes illustrated this with a drawing of a man placing his foot near a flame. He outlined a series of events that supposedly take place in the man's nervous system from the moment his skin contacts the flame to the moment he pulls his foot away. According to Descartes, the "tiny particles" or molecules that comprise the fire cause the area of skin that they touch to move. When the skin moves, a nerve fiber attached to it is pulled, causing a pore at the other end of the nerve to open, in turn allowing animal spirits to flow through the nerve to various muscles, causing the muscles to change shape,

and finally pulling the man's foot away from the flame. Flow of animal spirits down other nerves also causes the man's head and eyes to turn to look at the flame, he says (Descartes 1664/1985, 102).

Figure 1. Descartes's illustration of the man pulling his foot away from the flame. Reproduced from Descartes (1664/1985), out of copyright.

From this simple mechanical account of reflexes, Descartes built up a model of the nervous system to explain more complex phenomena like learning and memory. He suggested that associative memory traces—the heat of the flame and how it looks--"are

imprinted on the internal part of the brain," however, he did not have much to say about how that imprinting happens.

If we move ahead to the mid-19th century, the Russian physiologist Ivan Pavlov made the next significant advances in discovering the neural mechanisms of learning. In the process of investigating the alimentary or salivation reflex in dogs, Pavlov discovered that his canine subjects salivate not only in the presence of food, but also in the presence of stimuli that regularly precede presentation of food, such as a tone, or the experimenter entering the room. He described the first type of reflex (e.g., to food) as "inborn," involving "regular causal connections between definite external stimuli acting on the organism and its necessary reflex actions" (Pavlov 1927/2003, 16). However, he hypothesized that a second type of reflex (e.g., to a tone) involves different "mechanisms" operative in "higher nervous centres" (Pavlov 1927/2003, 25) and is "built up gradually in the course of an animal's own individual existence" (Pavlov 1927/2003, 25). In contrast to Descartes, who thought the mind influences the body through the pineal gland, Pavlov claimed non-physical or psychic causes are not responsible for either innate or conditioned reflexes. Rather, reflexes can be explained solely in terms of neural mechanisms mediating between stimuli and responses. This was in line with the views of mechanist physiologists like Hermann Helmholtz.

In order to identify "the precise conditions under which new conditioned reflexes are established" (Pavlov 1927/2003, 26), Pavlov and his colleagues ran many rigorously controlled experiments. On the basis of their data, Pavlov concluded that a conditioned reflex can be established if: (1) the presentation of the conditioned stimulus (e.g., a tone) precedes the unconditioned stimulus (e.g., food), (2) the two stimuli overlap in time, (3)

the animal is alert and healthy, (4) the conditioned stimulus is an environmentally familiar one to which the animal is otherwise indifferent, and (5) the investigator ensures that the only stimuli operative in the experiment are the conditioned and unconditioned stimuli.

Having reliably produced conditioned reflexes in many canine subjects, Pavlov hypothesized the physiological conditions that allow their formation: "the linking up of impulses in different areas of the brain, by the formation of new nervous connections" is the "nervous mechanism" by which "new conditioned reflexes" are formed (1927/2003, 37). More specifically, Pavlov said "it appears that the cells predominantly excited at a given time" by an unconditioned stimulus (food) "become foci attracting to themselves the nervous impulses aroused by" the conditioned stimulus (tone), and that these impulses "on repetition tend to follow the same path and so to establish conditioned reflexes" (1927/2003, 38). Pavlov illustrated what he had in mind by appeal to "telephonic installation." He explained that he could telephone his laboratory directly, or he could call the operator to connect him to the laboratory. (In those days, operators would manually connect lines by plugging cables into jacks on a switchboard.) Both methods would result in the same outcome. However, "whereas the private line provides a permanent and readily available cable" much like the neural pathway of innate reflexes, "the other line necessitates a preliminary central connection to be established" much like how the neural pathway carrying information about the conditioned stimuli must be connected to the innate pathway. Pavlov did not know precisely the location of the formation of these new connections—he thought that it was possible that it could occur

"within the cortex" or "between the cortex and subcortical areas" (Pavlov 1927/2003, 37). He also had no explanation for how such changes in neural connectivity might occur.

An explanation began to emerge at the end of the 19th century. Wilhelm His (1886), working with growing nerve cells, August Forel (1887), working on nerve cell degeneration, and the great Spanish Histologist Santiago Ramón y Cajal (1888), using Camillo Golgi's (1873) silver nitrate stain on unmyelinated nerve cells, suggested that nerve cells are independent anatomical and functional units (rather than a physically connected web of fibers as previously believed). Golgi's illustrations demonstrated many variations of the nerve cell's typical structure of cell body, single axon, and branching dendrites, in different brain regions. Cajal (1890) showed how growing neurons push their growth code outwards, and gradually form more dendrites and axon collaterals.

In 1891, Cajal discovered that sensory nerves have their dendrites in the periphery and axons projecting toward the brain, while motor cells are the opposite way around. His "law of dynamic polarization" hypothesized that conduction of impulses travel in one direction only, from dendrite to cell body to axon. In the 18th century, Luigi Galvani had established that it is electric currents, not corpuscles (i.e., animal spirits) that transmit nerve impulses. However, Cajal and many of his contemporaries believed that neurofibrils contained in nerve cells "underlie the mechanism of neuronal impulse transmission" (Cajal, 1899, p.95), harkening back to Descartes's account.

Figure 2. Cajal's drawing of Purkinje cells with basket endings. Reproduced from Cajal (1894a), out of copyright.

These combined discoveries led to speculation in the 1890s that the growth or retraction of dendritic connections and axon collateral branches might account for learning. Cajal (1894a) suggested that genius in a subject such as music might involve increased branching of certain neurons' dendrites and axons. This was pure conjecture, however. Cajal was an anatomist working almost exclusively with histological methods (slicing and staining specimens, then examining them under a light microscope), which did not lend themselves well to discovering how learning occurs, nor indeed to discovering much about how nervous impulses are communicated between nerve cells. Physiological methods were required to discover the functional import of Cajal's anatomical findings.

The English physiologist Charles Scott Sherrington first introduced the concept of the synapse in 1897 in a textbook he helped edit (Foster 1897). Based on his work on spinal reflexes, Sherrington had deduced that there is a significant delay in the speed of neural impulses where there are connections made between several nerve cells along the way to or from the spinal cord, rather than single axons travelling the whole distance. He attributed this delay to an "intercellular barrier" or membrane (Bennett 1999). Cajal had convincingly argued that axon collaterals do not directly fuse with the cells they come into contact with, but the hypothesized junction could not be seen under a light microscope. Sherrington's work suggested that the synapse acted as a valve, explaining why conduction occurred in only one direction. He also explored the relationship between inhibititory and excitatatory connections. He remarked that the synapse offered "an opportunity for some change in the nature of the nervous impulse as it passes from one cell to the other" (Foster 1897). Physiologists and pharmacologists continued, in the early 20th century, to uncover the electrical and chemical mechanisms of synaptic transmission.

Advances in *neurophysiological* theorizing and methodology in the first half of the 20th century were instrumental in connecting this developing knowledge about synaptic transmission to the phenomena of learning and memory. One such advance came in the form of a simple neurophysiological postulate put forward by Donald Hebb in *The Organization of Behavior* (1949). Synthesizing a broad selection of research from psychology (e.g., Pavlov), neuroanatomy (e.g., Cajal), and neurophysiology (e.g. Sherrington, Lorente de Nó), Hebb hypothesized that just as associative learning at the level of behaving organisms required the repetition and contiguity of stimuli or stimuli

and responses, so too, did the permanent metabolic changes or growth processes thought to underlie learning, require the contiguous and repetitive excitation of the neurons carrying information about those stimuli and/or responses.

More specifically, Hebb claimed that "when an axon of [a] cell A is near enough to excite a cell B and repeatedly or persistently takes part in firing it, some growth process or metabolic change takes place in one or both cells such that A's efficiency, as one of the cells that fires B, is increased" (Hebb 1949/2002, 62). Although the main idea at the heart of Hebb's postulate was not new, as Hebb acknowledged, the postulate provided insight into the kinds of methods that could be employed by physiologists to determine if neurons were plastic in the way that Hebb's predecessors, like Pavlov and Cajal had claimed.

By the 1960's, neurophysiologists had a working model of the neuron as consisting of 1) an *input* component (the dendrites), 2) an *integrative* component (the axon hillock), 3) a *conductile* component (the axon) and 4) an *output* component – the synaptic terminal from which neurotransmitter is released (Kandel and Spencer 1968, 69-70). However, what was missing was "a crucial experiment identifying specifically a change occurring in neural tissues as learning takes place" (Hilgard 1956, 481). Such crucial experiments came much later in the form of Nobel Prize winner Eric Kandel and colleagues' development of a simplified preparation for studying the cellular and molecular mechanisms of simple forms of associative and non-associative learning in the invertebrate sea mollusc, *Aplysia californica*.

Aplysia has a defensive reflex known as the *gill-siphon withdrawal reflex*. When a tactile stimulus is applied to the animal's siphon---a small fleshy spout located above the

gill that expels seawater and waste---it retracts or withdraws the siphon and the gill. In one early set of experiments, Kandel and colleagues experimentally isolated the sensory neurons that carry stimulus information from the siphon, the motor neurons to which these sensory neurons project, and a set of excitatory and inhibitory interneurons that receive input from the sensory neurons and project to the motor neurons. By isolating the neurons that comprise this simple circuit (and other circuits to which it was connected), Kandel and colleagues were able to identify specific cellular and molecular changes that accompany a set of simple forms of associative and non-associative learning in *Aplysia*. More specifically, they studied a form of learning known as *sensitization*. In a sensitization experiment, an investigator begins by applying a tactile stimulus (e.g., a Q-tip) to an Aplysia's gill or siphon, so as to measure the extent and duration of the withdrawal reflex. The experimenter then delivers a set of noxious shocks to the organism's tail. Following these shocks, she reapplies the tactile stimulus and again measures the extent and duration of the withdrawal reflex. An increase in duration of the withdrawal reflex prior to the tail shocks compared to after the tail shocks is taken as indicative that the animal has learned that there is a noxious stimulus in its environment. Textbooks like *Principles of Neural Science (5th Edition*, 2012) reveal in detailed diagrams what we now know about the changes in the strength of the synaptic connections that underlie this form of learning and that they are mediated by specific changes in cellular and molecular activity.

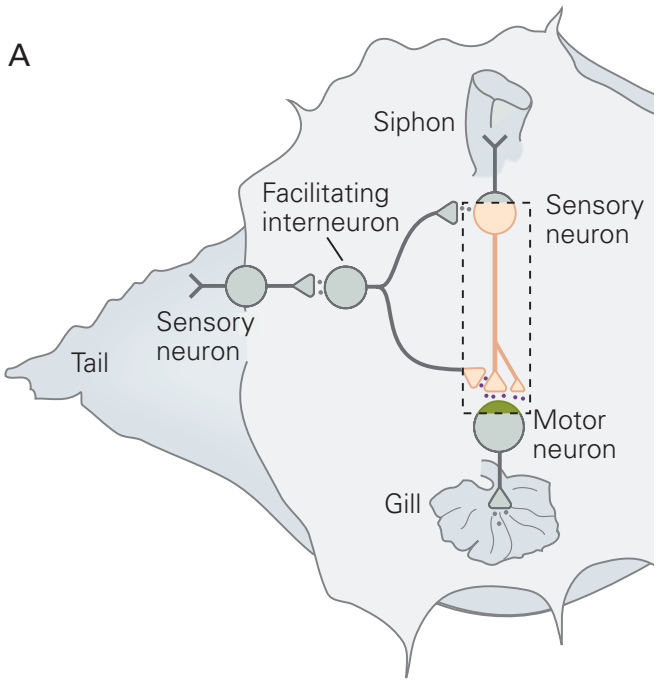

Figure 3. The simple neuronal circuit involved in sensitization of the gill-siphon withdrawal reflex in *Aplysia*. (Kandel, E.R., Schwartz, J.H., Jessell, T.M., Siegelbaum, S.A., Hudspeth, A.J. *Principles of Neural Science*, 5th Edition, 2012, McGraw-Hill Education. Reproduced with permission of McGraw-Hill Education).

These historical examples should give an idea of some of the ways that scientists discover and explain the mechanisms of the brain. In the following sections we'll revisit these cases, highlighting the fundamental features of mechanistic explanations, and considering some of methodological issues that arise in neuroscience.

3. Philosophical Accounts of Mechanisms in Neuroscience

One aim of philosophy of science is to understand the structure of science. Another is to account for scientific progress. Up until the latter half of the 20th century,

the examples considered in philosophy of science were taken primarily from the history of physics. This exclusive focus led to an understanding of science that conceived of its history as involving the discovery of laws (e.g., planetary motion, gravitational attraction) and the development of grand unifying theories (e.g., relativity theory). By the middle of the twentieth century, philosophers characterized scientific explanations as arguments, where statements of laws and initial conditions were taken to logically imply the observations to be explained or predicted. Different branches of science were thought to be hierarchically organized with those that studied the most fundamental things (e.g., particles) at the bottom and those that studied the least fundamental things (e.g., societies) at the top. Branches of science were regarded as compartmentalized, and progress within a given branch (e.g., psychology) was not taken to rely on developments within other branches (e.g., neurophysiology). Progress between branches was taken to involve "intertheoretic reduction"—the reduction of theories in "higher-level" sciences like biology to theories in "lower-level" sciences like physics (see Oppenheim and Putnam 1958, Nagel 1961) (See also Chapter 16).

The history of scientific research on learning and memory that we described above defies these characterizations in a number of ways. Descartes, Pavlov, Cajal, Hebb and Kandel were neither aiming to discover large-scale scientific theories, nor to reduce those theories to physical ones. The kinds of explanations for learning and memory phenomena they sought combined findings and insights from different areas of science including anatomy, physiology, psychology and later biochemistry. These diverse branches of science all aimed at understanding learning and memory from different angles and seemed to be making progress interactively rather than independently.

An alternative account of scientific explanation has recently been proposed that provides a more congenial analysis of the discovery strategies and markers of progress described in these historical cases, as well as in the biological sciences more generally. This account focuses on the role of *mechanisms* in scientific explanation. (See for example Bechtel and Richardson 1993/2000; Craver 2007; Glennan 1996; Illari and Williamson 2012; Machamer, Darden and Craver 2000.)

The first crucial step in providing a mechanistic explanation is to identify the phenomenon to be explained (See for example Glennan 1996; Bechtel 2008; Craver and Darden 2001). In each case considered above, the phenomenon for which a mechanism is sought is precisely delineated. Consider how Descartes conceives of a reflex; it begins with a stimulus—a man putting his foot into a fire—and ends when the man looks at the fire and retracts his foot from the flame. Similarly, Pavlov's conditioned reflexes begin with repeated and contiguous presentation of UCS and CS and end with elicitation of the conditioned response (i.e., salivation) to the conditioned stimulus. Sherrington, Hebb and Kandel postulate a very specific set of inputs—repetition and contiguity in firing of two cells that comprise a synapse—and a very specific output—a change in the way that the two cells communicate. As Peter Machamer, Lindley Darden and Carl Craver (MDC 2000, 3) claim, the phenomena of interest in mechanistic explanations have clear starting points or set-up conditions and clear endpoints or termination conditions.

Another important feature of mechanistic explanations is that they "account for the behavior of a system in terms of the functions performed by its parts and the interactions between these parts" (Bechtel and Richardson 1993/2000, 17) rather than in terms of general laws or theories. In Descartes's example, *molecules, nerve fibers, pores,*

animal spirits and muscles are all parts or entities of a human organism. *Tugging, opening, flowing, moving* are all activities in which these parts or entities engage in the production of reflex behaviors. In Cajal's anatomical work, *axon collaterals, dendrites, dendritic spines* and *growth cones* are the entities. Sherrington, Hebb, and Kandel later contributed knowledge about the activities of those entities.

William Bechtel and Robert Richardson (1993/2000) emphasize that a central heuristic strategy operative in developing mechanistic explanations is the decomposition of the phenomenon into its component parts and their operations. Decomposing a system in this fashion and explaining its behaviors mechanistically is not something that can be accomplished in a single area of science. As the "New Mechanists" emphasize, it requires input from many different areas of science. In the process, rather than one branch or area of science being reduced to another, input from different areas of science is "integrated into descriptions of multi-level mechanisms" (e.g., Craver 2007).

Cajal's work is a prime example of the decomposition strategy at work. It was critical in convincing anatomists of the "neuron doctrine," which extended cell theory to neural tissues, stating that the brain is made up of anatomically discrete cellular units. Cajal showed how neurons of different types, such as the basket and Purkinje cells of the cerebellum, are connected in organized patterns. He also worked to discover the anatomical properties of the neuron's sub-parts like dendrites, axon collaterals, growth cones, and dendritic spines. Cajal's discoveries of the anatomical properties of neurons and their component parts, in combination with Sherrington and Pavlov's discoveries about how these parts function, shaped the development of Hebb's postulate, which

informed Kandel's work in developing simplified preparations that decomposed reflex operations in *Aplysia* to a simple neuronal circuit and its component parts.

Bechtel and Richardson also note that the mechanistic explanatory strategy is often constrained by available technology and that scientists "will appeal analogically to the principles they know to be operative in artificial contrivances as well as in natural systems that are already understood" (1993/2000, 17) in order to provide mechanistic explanations. Descartes's appeal to the mechanical statues in St. Germain to explain reflex action and Pavlov's appeal to a telephone switchboard to explain how conditioned reflexes come about are clear examples of how the available technology of a time period can shape how investigators conceive of a mechanism.

This raises another important feature of mechanistic explanations detailed by Darden (2002): they are gradually discovered over time (see also Chapter 19). In terms of their empirical support, candidate mechanisms can have the status of "how-possibly" "how-plausibly" or "how-actually" explanations (Craver 2007). In terms of their completeness, mechanistic explanations start out as sketches (MDC 2000) that have gaps in their productive continuity or black boxes left to be filled in with detail. Sketches are revised, filled in, and fit into their surrounding contexts, until they eventually gain the status of adequately complete mechanistic explanations, or are rejected as false starts. The activities and sub-entities that mediate the connections between neurons were black boxes for Cajal before Sherrington developed the concept of the synapse. While Hebb put forward a "how-possibly" mechanism for permanent changes in communication between neurons, this was not yet considered an adequate explanation. Later work by Kandel and

colleagues was directed at understanding "how-actually" such changes come about during real learning events.

The process of discovery sometimes requires more substantial revisions to how the phenomenon was originally individuated and circumscribed (See Bechtel and Richardson 1993/2001; Bechtel 2008; Craver 2007, 2009). Experimentation may result in discoveries that prompt a revision to the original taxonomy of kinds of phenomena identified in a given field of research. For example, it may be discovered that what was once considered one phenomenon (e.g., memory) includes at least two forms (e.g., declarative and procedural). Experimentation may reveal that we were looking in the wrong place for a mechanism, or that a single mechanism performs what were originally thought to be two separate functions (e.g., see Eichenbaum and Cohen (2014) on the hippocampus's dual role in memory and navigation). It is well known today that rather than being the conduit between the body and immaterial soul, the pineal gland produces and secretes melatonin, which is involved in the modulation of circadian rhythms in the vertebrate brain.

These historical vignettes as a whole demonstrate another key feature of mechanistic explanations: they are multi-level. From Pavlov to Kandel, for example, we move from observations of mid-scale entities and activities like dogs, bells, and salivation, to micro-scale entities and activities like ions and neurotransmitter release. Mechanistic explanations involve entities and activities at multiple scales, some of which are sub-mechanisms that constitute higher-level components. Even in Descartes's description of the mechanism of the reflex, the behavior of the whole organism is explained by appeal

to some of its constituent parts and their sub-parts, like nerves, nerve fibers, pores, and animal spirits.

A related and notable characteristic of mechanistic explanations is that they are not byproducts of a single area of science. Rather, they rely for their development on information emanating from multiple different areas of science that study entities and activities at varying scales (cf. Darden and Maull 1977, Craver 2007). Consider Descartes' explanation of the reflex—it combined a corpuscular theory of matter, with a rudimentary understanding of the anatomy of the nervous system prevalent in his day, and a theory of animal spirits originating with Galen. Although Pavlov thought that physiology could advance an understanding of the mechanisms of conditioned reflexes without appeal to psychology, he recognized that it could not do so in the absence of advances in anatomy and cell biology. Cajal's histological preparations could not reveal the functional nature of the connections between contiguous neurons without the addition of physiological work, which Sherrington later contributed. Kandel and colleagues' research into the mechanisms of simple forms of non-asssociative learning in *Aplysia* combines anatomical, electrophysiological, biochemical, behavioral and pharmacological techniques.

4. Discovering Mechanisms: Open Philosophical Problems

The last two features of mechanistic explanation we mentioned—their multi-level nature, and the fact that they integrate results from various branches of science—are very much at odds with traditional thinking about scientific explanation. As mentioned briefly, in the mid-20th century, scientific phenomena at different scales, and the fields of science that study them, were thought to be related to one another in terms of reduction.

Chemistry, for instance, was supposed to occupy itself with a circumscribed range of chemical phenomena, which the methods of chemistry alone were appropriate for investigating. Furthermore, all of chemistry, it was thought, would eventually prove to be reducible to physics in the way that heat is reducible to the average kinetic energy of physical particles. Higher-level sciences, according to this way of thinking, may serve pragmatic and heuristic purposes along the way to finding the fundamental theory, but eventually should turn out to be superfluous.

Mental phenomena have long posed a challenge to this picture; many philosophers (and others) want to deny that the mind is reducible to more fundamental physical entities and activities. The multi-level nature of mechanistic explanations is meant to provide an alternative to reduction. All of the levels in a mechanism, from low to high, contribute to it performing its function. Going down lower does not provide a more fundamental understanding, even if it might provide finer grained details; in fact, scientists sometimes purposely focus their investigations at higher levels, because that's where the functions they're interested in are performed.

Many questions remain about how exactly this plays out in practice. Craver (2007) describes a picture of "integrative unity" in which a psychological capacity, such as spatial memory, is brought about by anatomically differentiated parts of the brain (area CA1 of the hippocampus), its physiological component parts (neural networks, neurons, synapses), and activities (firing, transmitter release), which in turn are composed of smaller scale parts (receptors, molecules) and their activities (activation, phosphorylation). If Craver is right, we should be able to fit the results from our historical vignettes into a hierarchy of mechanisms with Pavlov's conditioned reflexes at the top,

Hebb's associative synaptic mechanisms slightly below, Cajal's anatomical picture of the neuron and Sherrington's physiological insights into synapses another step down, then finally Kandel's molecular mechanisms of learning at the bottom.

Some of the entities and activities involved do fit together as parts to wholes, such as Kandel's molecular mechanisms, which describe parts of Cajal and Sherrington's neurons and synapses. However, it is not clear that the levels will always connect in such a tidy way, especially at the higher levels. Craver's account seems to presuppose that psychological and neural mechanisms are part of the same ontological hierarchy, yet psychological mechanisms do not necessarily have neural mechanisms as parts (Stinson, 2016).

Consider, for example, an information-processing mechanism that explains how an organism learns to respond to stimuli like burning flames or noxious shocks. That mechanism needs to store the relationship between stimulus and response in some memory medium. Reflexes mediated by nerve fibers, as Descartes imagined, can't do the whole job, because we can learn not only to pull our foot away from a flame, but also to do many other things with our limbs in response to many other kinds of stimuli. The nerve fiber doesn't have enough bandwidth to represent all of these learned relationships. This notion of bandwidth is an abstract concept that doesn't appeal specifically to any parts of the stimulus-response system, and yet it provides a psychological-level, mechanistic explanation of why Descartes's fibers can't be the whole story.

Another issue is that what look like the natural boundaries of a phenomenon from the perspective of one science (including start an finish conditions, and the way components are picked out) might not match up with what look like the natural

boundaries of the same phenomenon from the perspective of another science. In Stinson (2016) one of us argues that the science of memory has this problem. From the perspective of psychology, it seems clear that memory encoding, storage, and recall are distinct processes, for example. Yet from the perspective of neuroscience, there do not appear to be clear distinctions between these memory processes. When you try to integrate the mechanisms of memory studied in these two sciences, you do not find a neat relationship where neural mechanisms turn out to be the parts of psychological mechanisms. At the neural level, encoding, storage, and recall are all intertwined, so neural mechanisms of memory don't turn out to be related to psychological mechanisms of memory as parts to wholes.

Thus, rather than different areas of science like psychology and neuroscience being seamlessly integrated into unified mechanistic explanations, we often find explanations that cross levels in the mind-brain sciences to be messy and partial. Craver illustrates his mosaic unity with images like the one on the left of Figure 4. Instead, we suggest that the inter-field relationships we should expect will look more like the more complex image on the right.

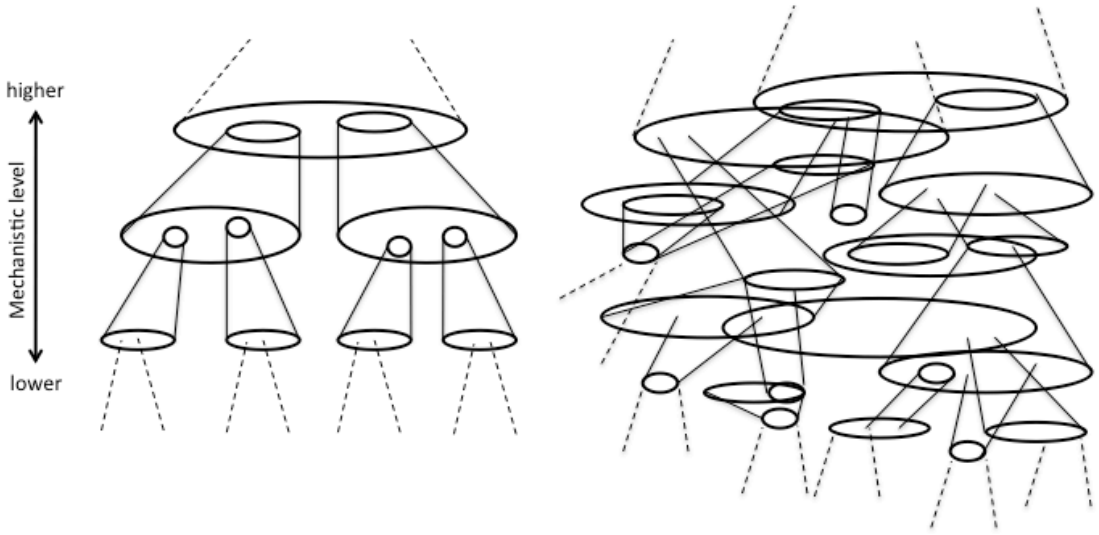

Figure 4. A comparison of Craver's (2007) view of inter-level relations [left] between mechanisms, and Stinson's (2016) [right]. Copyright 2016 Catherine Stinson. Used with permission.

Scientists working in different fields conceive of their phenomena of interest in different ways, experimentally investigate phenomena in different ways, ask different research questions, use different methods aimed at vastly different scales, and investigate these phenomena in different species. For example, Psychologists have historically been characterized as interested in providing explanations of cognitive capacities by functional analysis (e.g., Cummins 1983; Fodor 1968). Many psychologists believe this requires a clear specification and decomposition of abstract cognitive processes (like learning) involved in psychological tasks. When tasks are regarded as inappropriate for individuating a discrete function, they are often refined. However, neurobiologists investigating learning in invertebrates (like Kandel) and physiologists who investigate learning in non-human mammals (rodents, dogs (like Pavlov) are often not interested in

individuating the cognitive processes engaged during training in learning paradigms. While not worrying about abstract cognitive processes makes good sense in the case of *Aplysia*, which has a simple nervous system and can be studied using reduced preparations, ignoring the component cognitive processes that may be involved when rodents are trained in learning paradigms will render the connections investigators would like to make between cellular and molecular mechanisms and cognitive capacities tenuous at best (See Sullivan 2009, 2010, 2016).

A related problem with this unity picture is the issue of comparing mechanisms across species. Kandel's sea slugs are vastly different from humans, yet it is assumed that the results of experiments undertaken in one species are generalizable to others. Pavlov and Kandel are interested in mechanisms of human learning, but perform their experiments on canines and invertebrates. For both ethical and practical reasons, the systems scientists have historically and continue to use are model organisms like dogs, frogs, birds, rodents, sea slugs and fruit flies. While certain cellular and molecular mechanisms are conserved across species, there are obvious differences between sea slugs and humans that prohibit direct inference from one to the other.

Although neuropsychological research on patients with localized brain damage and fMRI experiments involving human beings have shed some light on the loci of specific types of learning and memory in the human brain, we continue to lack a mechanistic understanding of human learning; the explanations we currently have are patchy at best. It is supposed that advances in imaging technologies will eventually enable a visualization of the loci and mechanisms of human learning. However, before such discoveries are to be feasible, scientists require better methods for individuating

learning phenomena in human beings and non-human mammals. It is not simply that scientists lack the available imaging technologies; it is that many experiments in the cognitive neurosciences lack the rigor of work with model organisms that have smaller repertoires of behaviors. It is more difficult to design tasks that tease apart discrete kinds of learning in human beings than in *Aplysia*. One reason for this difficulty is that human beings might use multiple strategies to perform a cognitive task, and it's not always possible to predict the range of strategies that might be used, or to detect whether subjects are using the expected one.

Despite these difficulties in drawing connections between experimental findings using protocols from different fields, and in phylogenetically distant species, it is necessary for mechanistic explanations in neuroscience to find ways of bridging these gaps. As we mentioned earlier, Cajal's histological experiments could only go so far. He was able to get a fairly accurate picture of the anatomy of the neuron, but the structure alone couldn't reveal how neurons communicate. Pavlov could only figure out the functional characteristics of reinforcement learning using his experimental methods. His functional picture could not reveal what sorts of structures might give rise to reinforcement learning. As a general rule, neither bottom-up (from structure to function) nor top-down (from function to structure) methods in isolation can get us all the way to understanding mechanisms. Instead what is needed is a multi-level approach, with researchers simultaneously using many strategies to investigate different phenomena, alongside some efforts at linking the results of these together, i.e., something very much like how neuroscientific research is in fact pursued.

Scientists approach the problem of understanding the brain at various levels because there are robust regularities at various levels, both in neuroscience and in the life sciences more generally. There are some phenomena that we feel compelled to think of as real or natural kinds, like molecules, cells, organs, organisms, and species, even when we can't give them tidy definitions in terms of their component parts. We think that neurons are a genuine kind of thing despite the fact that (contra Cajal) nerve cells sometimes do fuse together in ways that challenge their anatomical and physiological independence. Organisms often end up in symbiotic relationships with other organisms, like our gut microbiota, without which we couldn't live, challenging the independence of organisms. The action potential depends on a membrane, ion channels, and extra and intra-cellular ions, but it only exists only within a narrow range of conditions.

A complex biological system like the brain will likely prove impossible to fit into a neat hierarchy of nested parts, because the borders of mechanisms are fuzzy. This does not mean that we can't ever have an integrated science that links together different levels. There are connections to be made between the results from different experimental paradigms, experiments on different species, models of different phenomena, and different models of the same phenomenon. Many of these connections will be partial, and the integrated picture will be patchy (see Schaffner 2006, Stinson 2016).

5. Conclusion

The historical case studies we considered span several centuries, but a common aim in each case was discovering mechanisms. Constructing multi-level mechanistic explanations involves intensive collaboration across different branches of science, and

involves many challenges, both pragmatic and methodological. Available technologies, training in experimental methods, choice of model organisms, levels of investigation, and inter-field collaborators all can either ensure success or act as barriers to progress. Integrating the discoveries from various fields where the phenomena are circumscribed in different ways requires piecing together results in complex ways, and carefully considering when and how results can be generalized to different contexts.

References

Bechtel, W. (2008). *Mental Mechanisms: Philosophical Perspectives on Cognitive Neuroscience*. New York: Taylor and Francis.

Bechtel, W. and Richardson, R. (1993/2000). *Discovering complexity: Decomposition and localization as strategies in scientific research*. Princeton, N.J.: Princeton University Press.

Bennett, M.R. (1999). "The early history of the synapse: From Plato to Sherrington" *Brain Research Bulletin*, Vol. 50, No. 2, pp. 95–118.

Cimino, G. (1999). Reticular theory versus neuron theory in the work of Camillo Golgi. *Physis,* 36(2), 431-472.

Craver, C. (2007). *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience*. Oxford, UK: Oxford University Press.

Craver, C. & Darden, L. (2001). Discovering mechanisms in neurobiology: The case of spatial memory, in P.K. Machamer, R. Grush, and P. McLaughlin (eds.), *Theory and Method in the Neurosciences*. Pittsburgh: University of Pittsburgh Press.

Cummins, Robert. (1983). *The Nature of Psychological Explanation*. Cambridge, MA: MIT Press.

Darden, L. (2002). Strategies for Discovering Mechanisms: Schema Instantiation, Modular Subassembly, Forward/Backward Chaining. *Philosophy of Science* 69 (3), S354–S365.

Darden, L. and N. Maull. (1977). Interfield Theories. *Philosophy of Science* 43:44-64.

Descartes, R. (1664/1985). *Treatise on Man* in *The Philosophical Writings of Descartes Volume I*, translated by John Cottingham, Robert Stoothoff and Dugald Murdoch. Cambridge, UK: Cambridge University Press.

Eichenbaum H, Cohen N. (2014). Can We Reconcile the Declarative Memory and Spatial Navigation Views on Hippocampal Function? *Neuron* 83(4):764– 770.

Fodor, J. (1974). "Special Sciences or The Disunity of Science as a Working Hypothesis." *Synthese* 28: 97-115.

Fodor, J. (1968). *Psychological Explanation: An Introduction to the Philosophy of Psychology*. New York: Random House.

Forel, A. (1937/1991) Out of my life and work. In G.M. Shepherd (Ed.), *Foundations of the neuron doctrine* (pp. 115-116). New York: Oxford University Press.

Foster, M.; with Sherrington, C. S. (1897). *A textbook of physiology, part three: The central nervous system*, 7th ed. London: Macmillan and Co. Ltd.

Glennan, S. (1996). "Mechanisms and the Nature of Causation." *Erkenntnis* 44: 49-71.

Golgi, C. (1873/1991). On the Structure of the Gray Matter of the Brain. *Foundations of the Neuron Doctrine,* Gordon M. Shepherd, tr. Oxford University Press, pp84-88.

Hebb, D. (1949/2002). *The Organization of Behavior: A Neuropsychological Theory*. Mahwah, NJ: Lawrence Erlbaum.

Hilgard, E.R. (1940). *Theories of Learning*, 2nd ed. New York: Appleton-Century-Crofts.

His, W. (1886/1991). On the structure of the human spinal cord and nerve roots. In G.M. Shepherd (Ed.), *Foundations of the neuron doctrine* (pp. 106-110). New York: Oxford University Press.

Illari, P. and Williamson, J. (2012). What is a mechanism? Thinking about mechanisms across the sciences. *European Journal for the Philosophy of Science* 2(1): 119-135.

Kandel, E.R. and W.A. Spencer. (1968). Cellular neurophysiological approaches in the study of learning. *Physiological Review* 48(1): 65-134.

Nagel, E. (1961). *The structure of science: Problems in the logic of scientific explanation*. London, UK: Harcourt, Brace & World.

Oppenheim, Paul, and Hilary Putnam. (1958). "The Unity of Science as a Working Hypothesis." In *Minnesota Studies in the Philosophy of Science*, ed. Herbert Feigl, Grover Maxwell and Michael Scriven, 3-36. Minneapolis: Minnesota University Press.

Pavlov, I. (1927/1960). *Conditioned Reflexes*. Mineola, NY: Dover Publications.

Machamer, Darden and Craver. (2000). Thinking about mechanisms. *Philosophy of Science* 67 (1): 1-25.

Piccinini, Gualtiero and Carl Craver. (2011). "Integrating Psychology and Neuroscience: Functional Analysis as Mechanism Sketches." *Synthese*, 183(3): 283-311.

Ramón y Cajal, S. (1888/1991). Structure of the nervous system of birds. In G.M.

Shepherd (Ed.), *Foundations of the neuron doctrine* (pp. 141-148). New York: Oxford University Press.

------- (1894a/1991). The fine structure of the nervous centers (Croonian lecture). In G.M. Shepherd (Ed.), *Foundations of the neuron doctrine* (pp. 239-253). New York: Oxford University Press.

------- (1894b/1990). *New ideas on the structure of the nervous system in man and vertebrates*. Neely Swanson and Larry W. Swanson (Trans.). Cambridge, MA: The MIT Press.

------- (1891/1988) On the structure of the cerebral cortex of certain mammals. In J. DeFelipe and E.G. Jones (Eds.), *Cajal on the cerebral cortex: An annotated translation of the complete writings* (pp. 23-54). New York: Oxford University Press.

Schaffner, K. F. (2006). Reduction: The Cheshire Cat Problem and a Return to Roots. *Synthese*, 151(3):377–402.

Shepherd, G.M. (1991). *Foundations of the neuron doctrine.* New York: Oxford University Press.

Stinson, Catherine (2016) Mechanisms in Neuroscience: Ripping Nature at Its Seams *Synthese* 193:1585–1614.

Stinson, Catherine. (forthcoming) You Say Schemas, I Say Schemata: Abstraction in MDC Mechanisms. In *Eppur si Muove: Doing Philosophy of Science with Peter Machamer*. Marcus Adams, Zvi Biener, Uljana Feest and Jacqueline Sullivan, eds., Springer.

Sullivan, Jacqueline. (2009). The Multiplicity of Experimental Protocols: A Challenge to Reductionist and Nonreductionist Models of the Unity of Science. *Synthese* 167: 511-539.

Sullivan, Jacqueline. (2010). Reconsidering Spatial Memory and the Morris Water Maze. *Synthese* 177(2): 261-283.

Sullivan, Jacqueline. (2016). Construct Stabilization and the Unity of the Mind-Brain Sciences. *Philosophy of Science.*

Wimsatt, William. (2007). *Re-Engineering Philosophy for Limited Beings: Piecewise Approximations to Reality*. Cambridge, MA: Harvard University Press.

Wimsatt, William. (1976). Reductionism, Levels of Organization and the Mind-Body Problem, in G. Globus, I. Savodnik, and G. Maxwell, eds., *Consciousness and the Brain*, New York: Plenum, pp. 199-267.

---

[1] The authors would like to thank Stuart Glennan and Phyllis Illari for very helpful comments on an earlier draft of this chapter.

[2]Co-authors had equivalent input and are listed in alphabetical order.