**How Metaphysical Commitments Shape the Study of Psychological Mechanisms**

Eric Hochstein

*Forthcoming in: Theory & Psychology*

**Abstract**: The study of psychological and cognitive mechanisms is an interdisciplinary endeavor, requiring insights from many different domains (from electrophysiology, to psychology, to theoretical neuroscience, to computer science). In this paper, I argue that philosophy plays an essential role in this interdisciplinary project, and that effective scientific study of psychological mechanisms requires that working scientists be responsible metaphysicians. This means adopting deliberate metaphysical positions when studying mechanisms that go beyond what is empirically justified regarding the nature of the phenomenon being studied, the conditions of its occurrence, and its boundaries. Such metaphysical commitments are necessary in order to set up experimental protocols, determine which variables to manipulate under experimental conditions, and which conclusions to draw from different scientific models and theories. It is important for scientists to be aware of the metaphysical commitments they adopt, since they can easily be led astray if invoked carelessly. On the other hand, if we are cautious in the application of our metaphysical commitments, and careful with the inferences we draw from them, then they can provide new insights into how we might find connections between models and theories of mechanisms that appear incompatible.

**Keywords:** Mechanisms, models, metaphysics, integration, double dissociations, subtractive neuroimaging studies, emotions.

The pioneering neuroscientist Santiago Ramón y Cajal famously claimed that "to know the brain […] is equivalent to ascertaining the material course of thought and will, to discovering the intimate history of life in its perpetual duel with external forces" (1937). In other words, with the age of substance dualism behind us, the quest for understanding the human mind has become a quest for understanding the human *brain*. Moreover, since we know the brain to be a massively complex system composed of neurochemical mechanisms, the quest for understanding the human mind has ultimately become a quest for understanding the physical *mechanisms* that compose it.

This should not be surprising, as it brings the human mind in-line with other known biological phenomena. The history of biology has largely been a story about the search for mechanisms. Our understanding of a biological phenomenon typically involves understanding the mechanisms that

produce and sustain it. Understanding how DNA replicates, for instance, involves understanding how "the DNA double helix unwinds, exposing slightly charged bases to which complementary bases bond, producing, after several more stages, two duplicate helices" (Machamer, Darden & Craver, 2000, p. 3). Understanding how and why the poison Curare kills involves understanding how it mimics the structure of acetycholine, a neurotransmitter, and binds to certain receptors at the neuromuscular junction. This blocks the muscle from receiving electrical signals from adjacent motor neurons. As a result, muscles like the diaphragm cannot receive electrical stimulation and therefore become paralyzed, making it impossible for the animal to breathe (Craver & Darden, 2013, p. 1-14). In a similar fashion, understanding psychological phenomena like memory or attention requires understanding how the mechanisms of the brain are structured, and how they operate to generate such phenomena.

While the discovery and understanding of these psychological mechanisms has largely been thought to be a strictly empirical process, in this paper I will argue that there is an important and often overlooked *philosophical* component to the study of mechanisms in psychology and cognitive science. More specifically, I will demonstrate that scientists must adopt explicit *metaphysical* commitments when studying psychological mechanisms. These metaphysical commitments must go beyond what is empirically justified, but are nevertheless necessary in order to set up experimental protocols, determine which variables to manipulate in experimental contexts, and which conclusions to draw from our scientific models and theories. These commitments are interwoven with our experimental practices, and often implicitly guide our scientific methodologies. Moreover, appropriate study of psychological mechanisms requires that scientists be constantly vigilant of what their metaphysical commitments are, how they guide their research, when they require additional justification, and when they must be abandoned in favour of others. In this respect, being a good scientist requires being a responsible metaphysician, and engaging in appropriate philosophical theorizing.

In order to make this argument, I begin in section 1 by providing an account of what psychological mechanisms are, and the ways in which we study and discover them. In section 2, I highlight the ways in which metaphysical commitments and reasoning play an essential role in the successful study and discovery of mechanisms. In section 3, I identify particular instances in contemporary psychology and cognitive science where scientists who failed to identify or engage with the appropriate philosophical and metaphysical reasoning created problems for their experimental designs, and in their analysis of experimental results. Finally, in section 4, I argue that one of the most pressing problems in contemporary cognitive science, the lack of unification or integration among theories and models in the study of psychological mechanisms, can be overcome by paying closer attention to the metaphysical commitments that different models and theories implicitly adopt in their construction and application. By highlighting these commitments, we can find points of contact between seemingly contradictory models by identifying implicitly shared metaphysical commitments that they adopt. This allows each model to inform each other, and to contribute to the same over-arching integrative explanation of a mechanism, despite the models themselves making incompatible claims.

## Section 1: The Study of Mechanisms

### 1.1 What Are Mechanisms?

It is best to begin by saying a bit more about what mechanisms are. Talk of "mechanisms" is ubiquitous throughout the sciences. For example, in political science we might talk of the mechanism responsible for social change. Similarly, an economist might talk of the mechanism responsible for the rise in monetary inflation. However, there is a more specific sense of mechanism that is commonly employed within the context of psychology, neuroscience, and cognitive science. Put simply:

> While mechanisms are defined variously, the core idea is that they are organized systems, comprising causally relevant component parts and operations (or activities) thereof. Parts of the mechanism interact and their orchestrated operation contributes to the capacity of the mechanism. (Miłkowski, 2013, p. 3050)

Suppose, for instance, that we wish to understand the mechanism by which one neuron transmits an electrical signal to another. In this case, the mechanism involves a neuron (the presynaptic neuron) releasing a neurotransmitter which crosses the synaptic cleft between it and the adjacent neuron (the postsynaptic neuron). The chemical then binds to particular receptors on the postsynaptic neuron, which opens pores in the cell wall allowing electrical current (positively or negatively charged ions) to pass from the first neuron into the second.

Note that understanding the mechanism in this case involves decomposing the system into parts and operations. More specifically, it involves identifying particular structures (the presynaptic neuron, the postsynaptic neuron, the neurotransmitter, the receptors, the positively or negatively charged ions) whose interactions produce the phenomenon. These structures must be organized in very particular ways for the phenomenon to occur (i.e. the neurons need to be organized so that the neurotransmitters, when released, are positioned to cross the synaptic cleft and bind to the receptors). This organization allows the parts of the system to causally interact with one another in very particular ways (i.e. the neurotransmitters *bind* to receptors, *opening* the pores, allowing the ions to *enter* the neuron). This in turn produces the phenomenon of interest. Understanding a mechanism therefore requires identifying and understanding the *parts* that make it up, their *organization*, the *operations* that go on between the parts, and the resulting *phenomenon.*

With this understanding of mechanisms in hand, let us now consider how we can discover and learn about such mechanisms. What follows is a brief account of what is involved in learning about the mechanisms responsible for psychological phenomena. A more complete account of this process would need to be far more in-depth than can be reasonably provided in a paper of this length, and so for a more detailed exploration of this process, I would point readers to: Craver, 2007; Bechtel, 2008; and Craver & Darden, 2013. For the purposes of this paper, I wish only to highlight and identify some of the essential features and complications of the discovery process, and provide some historical examples to illustrate.

## 1.2 Learning about Mechanisms

In order to understand the mechanism responsible for a particular psychological phenomenon, we must begin with an understanding of what that phenomenon is. For example, if we wish to understand the mechanisms responsible for episodic memory, then we must begin with a characterization of what episodic memory is supposed to be, when it appears, and under what conditions. Knowing when the phenomenon occurs, or fails to occur, is often the first step in discovering the underlying mechanism that produces it. As William Bechtel rightly points out:

> To begin with, if scientists are theorizing about a mechanism to explain a particular kind of behavior, it is indispensable to begin with a good characterization of the behavior. Otherwise, they may produce a proposal for a possible mechanism that does not in fact exist, and whose behavior would not correspond to anything that actually happens. Moreover, once a mechanism is proposed, the evidence for or against it comes not just from investigations of internal

> operations but from whether it actually can account for factors that are known to affect the behavior. (2005, p. 323)

This means we often study mechanisms by reverse engineering them. By characterizing in detail the phenomenon, and then using this as a guide for determining what mechanisms are operating when the phenomenon is in effect, or absent when it is not. The more detailed our account of the phenomenon, the more constraints it places on what sorts of mechanisms are capable of producing that phenomenon under those known conditions.

Consider the study of the action potential in the mid-20th century. The now famous Hodgkin & Huxley model of the squid giant axon (1952) was originally created with the intention of identifying and describing the mechanism by which the action potential of the neuron fired. The model ultimately failed to accomplish this goal, but it was able to mathematically characterize the time course of the action potential by identifying relevant time and voltage dependencies. With this mathematical characterization of the phenomenon in hand, scientists were then able to use these dependencies as a guide for future research by treating them as constraints that any potential mechanistic account of the phenomenon needed to conform to. In the 1960s and 70s, "textbook descriptions of the mechanisms underlying the action potential were heavily based on the Hodgkin and Huxley papers. Membranes were typically introduced with an equivalent circuit diagram and were then followed by the Hodgkin-Huxley sodium conductance curves" (Trumpler, 1997, p. 63).

As another example, consider the study of the Thompson Effect. The Thompson Effect is a psychological phenomenon whereby we discriminate the speed of an object based not on its actual movement but on the visual contrast between it and its surrounding environment. The greater the visual contrast, the faster the object appears to be moving. Alan Stocker and Eero Simoncelli were able to construct a mathematical model which could accurately predict how fast an object would appear to be

moving to an observer given its level of visual contrast (Stocker & Simoncelli, 2006). This model did not identify any of the mechanisms responsible for the Thompson Effect, but by describing and predicting the phenomenon in a detailed way, it pointed the way towards what mechanisms were likely involved. They claim that:

> The form of the contrast-dependent measurement noise in our model suggests that the locus of representation for measurements m is likely to be cortical. Neurons in area MT are a natural choice: they are highly motion selective, and their responses have been directly linked to perception. (Stocker and Simoncelli, 2006, p. 583)

Of course, using a detailed account of the phenomenon as a guide for identifying the relevant underlying mechanisms is not nearly as straightforward a task as it may initially appear. Discovering and characterizing the parts, organization, or operations of a complex mechanism, even when we have a detailed account of the phenomenon, often requires the resources, methodologies, and experimental practices of many different scientific domains. Each domain provides limited information about the system that can inform, constrain, and guide the research of the others.

To illustrate, let us return to the study of the action potential. The Hodgkin & Huxley model characterized the time course of the action potential, but did not provide any account of the parts or operations of the mechanism within the neuron that produced them. Ultimately, an account of what these structural and organizational features of the neuron were had to be *inferred* from the results of many different studies conducted by scientists working in different domains. These studies, in conjunction with the behavioural data of the Hodgkin & Huxley model, allowed scientists to triangulate on what the mechanism of the action potential was likely to be. For instance, in the early 1970s, the biologists Singer and Nicholson proposed that a part in the mechanism of the action potential was a

protein pore embedded in the membrane of the cell that allowed for changes in its conductance (1972). Part of their reasoning for this was due to results being generated by different scientists working in electrophysiology. As Maria Trumpler argues:

> Electrophysiologists had shown that several neurotoxins, most notably Tetrodotoxin (TTX), not only blocked the sodium conductance but produced a binding curve of the same form as that established for other proteins. Thus it was plausible that the sodium conductance might be a protein pore embedded in the membrane, with at least one part permanently accessible to the external solution (1997, p.71)

Put simply, certain behavioural changes we see in neurons when they are exposed to tetrodotoxin are consistent with behavioural changes we see in those neurons when they bind to various proteins, but in this case it likewise stops electric current from entering the neuron. This suggests that the conductance of the neuron may be due to a protein pore, which was being blocked by the tetrodotoxin. Here, the results of experiments generated by electrophysiologists, in addition to the conductance curves identified by the Hodgkin & Huxley model, were both needed to help biologists point the way towards identifying and understanding one of the crucial components of the mechanism for the action potential (protein pores). This nicely highlights how learning about the parts, operations, organization, and phenomenon of a mechanism requires the resources of different domains of science, each of which provide guidance and constraints on the others. Other examples include the discovery of the mechanisms responsible for learning (Craver 2007, p.243), long term potentiation (Miłkowski 2016a), and protein synthesis (Machamer, Darden & Craver 2000), which all required insights from multiple distinct scientific domains and methodologies.

With this established, I now intend to argue that philosophy is another domain which contributes to the study of psychological mechanisms in important ways. More specifically, that this process of discovery and understanding only works when scientists take on board numerous *metaphysical* commitments about the world that often go beyond what is empirically warranted, but which are nonetheless necessary in order for the experimental study of mechanistic systems to take place.

## Section 2: Mechanisms, Methodology & Metaphysics

As we've seen, the study of a complex mechanism often begins with a detailed characterization of the phenomenon produced by that mechanism, which then points the way to possible underlying structures and causes. However, in order for us to effectively study the phenomenon itself, scientists must begin by taking on metaphysical commitments regarding what the phenomenon is and what its boundaries are. These commitments are not initially determined by empirical study, since they must be presupposed in order to set up our experimental conditions *needed for* relevant empirical study. Such commitments often require appropriate justification and argumentation, since different commitments would licence or justify incompatible sets of experimental protocols and procedures. Moreover, keeping track of these commitments throughout our empirical investigations, and recognizing when to revise or adopt different commitments as new data emerges, is key to our study and discovery of psychological mechanisms. In other words, scientists must learn to be good metaphysicians in order for the study of psychological mechanisms to be successful.

To illustrate, suppose we wish to identify and understand the psychological mechanisms responsible for emotions like anger, sadness, or joy. This requires starting with the assumption that there is indeed a coherent and well-delineated metaphysical phenomenon that corresponds to anger,

sadness, or joy. We must be committed to metaphysical facts about what the phenomenon is supposed to be, that it is there in the world to be studied, and that there are clear instances of its manifestation. We cannot look for the mechanisms for anger without starting with assumptions regarding the existence and occurrence of anger. The psychologist Lisa Feldman Barrett highlights this fact when she notes that there is...

> ...wide acceptance [among psychologists and neuroscientists] of assumptions that are not warranted by the available empirical evidence. These assumptions can be summarized by one core idea: Certain emotions (at least those referred to in Western cultures by the words "anger," "sadness," "fear," "disgust," and "happiness") are given to us by nature. That is, they are natural kinds, or phenomena that exist independent of our perception of them. Each emotion is thought to produce coordinated changes in sensory, perceptual, motor, and physiological functions that, when measured, provide evidence of that emotion's existence. The natural-kind view of emotion has been productive in defining the boundaries for the scientific study of emotion and continues to guide scientific discourse: It underlies the major questions, the experimental designs, and the interpretation of empirical findings that characterize emotion research as a domain of scientific inquiry. (2006, p. 28-29)

For the moment, let us leave aside the question of whether this is the *correct* metaphysical story to tell about what basic emotions are (namely, "a nonarbitrary grouping of instances that occur in the world […] given by nature and is discovered, not created, by the human mind" Barrett, 2006, p. 29). What is more important to note for now is that in order to study emotion, we must take on board *some* metaphysical commitments about what emotions are, and what their boundaries are. It is these

commitments that allow us to define the conditions for the scientific study of emotion, underlies the experimental designs we use, and the ways in which we interpret our empirical findings.

To illustrate, suppose we wish to study the mechanism responsible for anger. If we assume that anger is a distinct psychological phenomenon, then we might try to use various neuroimaging techniques to see which brain regions become active when a given subject is angry, and whether that subject remains angry when those regions are not active (this might include using things like PET or fMRI scans to observe blood flow to various parts of the brain). This may allow us to localize the parts and organizations that constitute the mechanism for anger. On the other hand, suppose we assume that anger is *not* a single unified phenomenon, and is instead an ill-defined category that is arbitrarily lumping together numerous different and completely unrelated psychological phenomena. We would now not conduct the same experiment as before. The fact that certain brain regions are active when we would *describe* someone as being "angry" tells us nothing about what mechanisms are in fact at work, or what those brain regions do. And so depending on the metaphysical commitments we take on board regarding what the nature of the phenomenon is, and when it occurs, this will lead to radically different accounts of what the underlying mechanisms of the system are thought to be, and *how we ought to study them*. Yet, we cannot proceed without *some* metaphysical story in-hand; otherwise we have no means of setting up experimental protocols at all, or determining which variables to intervene on. In this respect, scientists must go beyond what is empirically licensed by the available evidence from the outset and make deliberate metaphysical claims and arguments about the phenomenon in the world in order to scientifically investigate it.

To build on this point further, consider the sorts of neuroimaging techniques mentioned above (like PET and fMRI scans). Attempts to localize and identify which neurological mechanisms are responsible for which kinds of psychological phenomena using such techniques frequently involve the use of subtractive neuroimaging studies. These studies require taking two sets of scans of a subject. The

first set of scans are taken when the subject is passive; not engaging in any sort psychological task. This provides a baseline for comparison. We then take a further set of scans when the subject is engaging in the relevant psychological activity (or undergoing the relevant experience) we wish to explain. We then subtract the findings of the first set from the second and observe any differences. Any brain regions that appear to be active when the psychological phenomenon is present, but not when the subject is at rest, is thereby thought to contain the relevant mechanism responsible for the phenomenon.

Yet in order for these sorts of studies to provide the relevant empirical data regarding the location of the target mechanism, working scientists are first required to adopt explicit metaphysical commitments regarding structural and functional features of the brain that have not yet been empirically determined. Specifically, that the brain is primarily composed of feed-forward modules, and that the neural activity identified in the subtraction studies can point to the relevant localizable module responsible for the phenomenon. Without these metaphysical assumptions, we cannot infer from increased activity in a particular brain region that the mechanism responsible for the psychological phenomenon is to be found there. As Van Orden & Paap note:

> Briefly, one must begin with [the assumption that they have the] "true" theory of cognition's components, and assume that corresponding functional and anatomical modules exist in the brain. The true theory is necessary to ensure that experimental and control images differ by the single component of interest. Additionally, the brain must be composed of feed-forward modules to ensure that the component of interest makes no qualitative changes "upstream" on the shared components of experimental and control tasks. Finally, each contrasted task must invoke the minimum set of components for successful task performance. If any one of these assumptions is false the enterprise fails.

> Assuming we knew the actual components of cognition, and the modularity assumption were true, then neuroimaging studies could reliably localize cognitive components in brain regions. One cannot disconfirm false models, however, because subtractions always highlight some brain region. Nevertheless, if different laboratories draw experimental and control tasks from different models (only one of which could be "true"), or if the modularity assumption is false, then images may diverge rather than converge. (1997, p. S87-S88)

Van Orden & Paap go on to argue that many of these metaphysical commitments are in fact false, and thus call into question the scientific value of certain neuroimaging studies (see also: Uttal 2001). For our immediate purposes, however, we need not be concerned with whether such metaphysical commitments are true or not. Instead, it is important only to highlight the fact that our empirical investigation of the mechanisms underlying psychological phenomena using subtractive neuroimaging studies first *requires* that working scientist adopt metaphysical commitments about the system producing it. Without some metaphysical commitment one way or another, we are unable to draw relevant conclusions about psychological mechanisms from such studies. Different sets of metaphysical commitments results in wildly divergent interpretations of the data produced by the techniques. In this regard, the very act of engaging in this sort of empirical investigation is heavily dependent on the metaphysical commitments that scientists adopt, and the justifications they provide in favour of them.

Even the assumption that we can identify the well-delineated parts and operations of the mechanisms responsible for psychological phenomena contains an empirically-contentious metaphysical assumption that the mechanisms which exist in nature always have clearly definable parts, operations, and boundaries. Yet often our empirical studies do not yield a clean or straightforward decomposition of a psychological phenomenon into well-delineated parts and boundaries (Bechtel 2015; Austin 2017; Chirimuuta 2017). Mutual manipulation experiments, which are designed to empirically demarcate parts

of mechanisms from their environment[1], frequently include too much or too little. As Christopher Austin notes,

> Though [mutual manipulability seems] *prima facie* plausible, a bi-directional boundary building test based on counterfactual discrimination may be problematic in the biological realm. For instance, in one direction of dependency, it may be too restrictive, and generate false negatives: the holistic, mechanism-level activity of complex biological systems is often impervious to minor alterations in the activities of their constituents – a phenomenon known as *robustness*. In the other direction, it may be too permissive, and generate false positives: a large swathe of organismal features (both morphological and behavioural) bear counterfactual dependence relations to extra-organismal, environmental stimuli, as evidenced by the well-known phenomenon of *phenotypic plasticity.* (Austin 2017, p.417)

A clearly defined boundary between mechanisms and their environments frequently requires a choice on the part of working scientists to metaphysically demarcate where the mechanism officially begins and ends, and this often cannot be straightforwardly determined through strictly empirical means. This is why Bechtel insists that...

> It is the cognitive activities of investigators that picked out some entities as the parts constituting the [target mechanism] and screened off or ignored the effects of other entities on these parts. [...] In particular, the scientists have deemed specific entities as particularly

[1] These experiments attempt to determine whether a given structure can be manipulated independently of the occurrence of the phenomenon. If a manipulation of the structure has no effect on the occurrence of the phenomenon, then it is likely not a part of the mechanism producing it. But if we cannot manipulate the phenomenon without likewise manipulating the structure, then it is likely a part of the mechanism itself.

> relevant to explaining the phenomenon they were interested in and have included them in the mechanism. This involves creating boundaries in an interconnected world. (2015, p. 88)

This is likewise why Austin suggests that "the compositional stability that individuates mechanisms is merely a heuristic necessity applicable only to models of mechanisms: the biological realm is not mereologically dissected into frozen collections of unalterable clockwork, even if our models of that realm must be." (Austin 2017).

This is not to suggest that psychological phenomena do not have mechanistic explanations, only that there is often not a clear or straightforward way of empirically demarcating the boundaries of where such mechanisms begin and end. However, despite this, there can often be good *metaphysical* reasons to clearly demarcate mechanisms at one point instead of another, even if there are not decisive empirical ones. As Bechtel argues,

> For different explanatory purposes researchers may draw these boundaries in different locations or at different time points. These choices, though, while not simply responsible to pre-existing boundaries, are not entirely arbitrary. [...] [W]hile real-world networks are highly interconnected, there are clusters within them that are semi-independent of the rest and productively posited to be the mechanisms responsible for specific phenomena. (2015, p. 85)

More importantly, scientifically studying the neurological and physiological structures and organizations that produce and sustain psychological phenomena frequently *requires* that scientists work with clearly defined boundaries to systems in order to set up experimental protocols. Thus, a metaphysical commitment to a clearly defined boundary becomes an essential part of investigating mechanisms. Mazvitta Chirimuuta (2017), for instance, claims that "the positing of boundaries is a useful way to

simplify the explanandum. It enables neuroscientists to bracket some of the known facts about the brain's messy, Heraclitean nature." (p.1150). Bechtel likewise insists that imposing boundaries "serves an important role in the project of developing mechanistic explanations" (2015, p. 92). Meanwhile, Chris Eliasmith (2009) notes that without deciding on a system's boundaries, it becomes impossible to effectively investigate it under experimental conditions.

In all of these examples, engaging in appropriate scientific methodology first requires engaging in metaphysical theory and argumentation, and that the results of these endeavors radically shape the experimental practices that result. All of this is not to suggest that the initial metaphysical commitments that scientists choose to adopt are somehow immune from future revisions, only that we *must* engage in metaphysical deliberation and reasoning in order to conduct the required empirical investigations. Of course, our metaphysical commitments might eventually need to be revised after a great deal of empirical work has taken place. We might find that the theories and models which use certain metaphysical commitments as their foundation eventually run into dead ends, or provide results or conclusions that are impossible or incoherent. This can require changing our metaphysical understanding of what the phenomenon is. In this regard, the study of psychological mechanisms requires being responsible metaphysicians: keeping track of which metaphysical commitments are required to empirically investigate the phenomenon, and which of our commitments may need further justification or revision.

Should we be concerned by the fact that scientists must be good metaphysicians in order to do their empirical research? Not at all. The very shift during the cognitive revolution from behaviourism to representational theories of mental phenomena demonstrates how a shift in metaphysical commitments have altered the way in which scientists can and do study psychological mechanisms. The study of mechanisms is a deeply interdisciplinary endeavour. The conceptual resources and methodological practices of many different disciplines play an important part in the discovery, study,

and understanding of psychological or cognitive mechanisms. Philosophy is simply one of these disciplines. Like with every discipline, however, it must be carried out responsibly. Being *careless* with one's metaphysical commitments can easily lead one astray. Scientists need to be cautious about the sorts of metaphysical commitments they adopt, and careful about the inferences they draw from those commitments. The importance of metaphysics in studying psychological mechanisms can be made even more apparent when we examine cases in which problems have emerged in psychology and cognitive science precisely because a lack of care was taken by psychologists in tending to their metaphysical commitments, or failing to sufficiently justify them.

## Section 3: When Lack of Metaphysical Care Breeds Adversity

To better illustrate just how the metaphysical commitments of working scientists influence the scientific study of psychological mechanisms, let us consider the ways in which inattention to such commitments have led to complications in the study of such mechanisms.

### 3.1 Double Dissociation and Mechanism Localization

One of the most common methods by which cognitive scientists try to map psychological phenomena onto neurological mechanisms is by way of double dissociation studies. If damage to one brain region seems to affect a subject's ability to carry out one sort of psychological task, but not another, and damage to a different region affects their ability to carry out the second task, but not the first, then we have reason to think those tasks are produced by distinct mechanisms localized in different brain regions. The localization of one mechanism can be inferred from the subject's deficit in the given task when there is damage to one brain region, but not to the other. Meanwhile, the positing of a different

mechanism for the second phenomenon can be inferred from its continued function after lesioning of the first region, but not after lesioning of the second. These sorts of cases are the primary method employed in Cognitive Neuropsychology, a field dedicated to bridging our understanding of psychological phenomena with our mechanistic understanding of the brain (Patterson & Plaut 2009).

While this may seem intuitive, the problem with this method of inferring distinct localized psychological mechanisms from instances of psychological disfunction is that we must adopt certain contentious metaphysical commitments about the brain for the inference to hold. One such commitment is that the brain is massively modular, since this is required in order to justify the inference from the subject's deficit in a given task to the idea that the mechanism uniquely responsible for that task resides in the damaged brain region. A second is that we can use such cases of double dissociation to reliably say something about the functional organization of human brains *in general*. Yet there is strong evidence that we should be highly suspicious of both these metaphysical assumptions (see: Goldberg 1995; Van Orden & Paap 1997; Samuels 1998; Cowie & Woodward 2004; Poldrack 2006; Hohwy 2007; Patterson & Plaut 2009; Anderson 2007, 2008, 2010; Eliasmith 2013; Palecek 2017).

For our purposes, the fact worth noting here is not so much that the metaphysical commitments adopted by cognitive neuropsychologists may be contentious, but that many cognitive neuropsychologists *were not aware that they were presupposing these commitments*, *or did not believe they needed any justification*. As Patterson & Plaut (2009) point out, "the assumption that cognitive abilities have the identical functional organization in all brains […] was not always made explicit and⁄or defended", and that when it came to criticisms of this assumption, "perhaps the surprise is how relatively infrequent such criticism was." (p.42) They similarly note that published concerns about the empirical plausibility of massive modularity "were infrequent and had little impact on the cognitive neuropsychology mainstream" (p.44). The lack of appropriate attention paid to these implicit metaphysical commitments have resulted in many cognitive neuropsychologists failing to notice that

various key inferences being drawn from double dissociation studies were extremely misleading or not consistent with other empirical findings.

The intuitive inference that takes one from a double dissociation to the conclusion of localized mechanisms is extremely seductive to many psychologists and cognitive neuropsychologists precisely because they have unexamined metaphysical commitments that shape how they evaluate such cases. For instance, if damage to a particular brain region does not yield a clear deficit in a given psychological function, the immediate conclusion that the mechanism which carries out this function must be located elsewhere is only intuitive if one implicitly assumes that the mechanisms of the brain are modular and not deeply integrated and interconnected. If the later is the case, as an increasing number of cognitive scientists argue (e.g. Anderson 2007, 2008, 2010; Patterson & Plaut 2009; Eliasmith 2013; Palecek 2017), then it may well be the case that damage to the brain region *does* damage the mechanism responsible for the target function or ability, but that other brain regions compensated for the damage allowing the ability to be carried out in a different way (see: Seidenberg 1988). In other words, "when a complex and adaptive system like the brain is damaged, it may still manage to perform a task but in a fashion not all that informative about normal function" (Patterson & Plaut 2009, p. 44). Meanwhile, the loss of function when a different brain region is damaged likewise does not necessarily licence the conclusion that the mechanism is localized there. Instead, it may reflect the fact that certain essential links in a deeply interconnected network may have been severed, cutting off a small part of the mechanism from it's larger distributed whole (as opposed to the mechanism itself being localized in the second area). As Michael Anderson points out, "it is possible for focal lesions to cause specific functional deficits in non-modular systems, and double-dissociations do not by themselves support any inference about the underlying functional architecture of the brain" (2010, p.248). And so the inference that is typically drawn to the localization of distinct mechanisms is not one that obviously follows from the results of any double dissociation case. Yet in virtue of not paying attention to their underlying metaphysical

commitments, many psychologists and cognitive neuropsychologists can often be unaware of just how much those commitments influence their evaluation of the data and the inferences they make.

To make matters worse, evidence that is emerging from other domains of cognitive science which point to a more integrative picture of brain are being largely overlooked precisely because the implicit metaphysical commitment to modularity has led many to ignore other accounts of representation and information processing that have been developed. This is precisely why Patterson & Plaut caution that…

> …cognitive neuropsychology has become increasingly detached from other areas within cognitive science and neuroscience in large part because the modularity assumption licenses a lack of consideration of representations and processes, and it is exactly in this respect that cognitive science can make a critical contribution to cognitive neuropsychology. […] Perhaps it would be more fruitful to start with a theoretical framework grounded in interactivity and then explore the extent to which it can give rise not only to normal cognitive behavior but also to the types of selective deficits observed in neuropsychological research. (Patterson & Plaut 2009, p.47)

The point of all this is not to suggest that localization of functions to brain regions based on double dissociation studies is always necessarily inappropriate. Only that failing to pay attention to the implicit metaphysical commitments underlying their evaluation of such cases has resulted in many psychologists being unknowingly drawn to misleading conclusions that may not be licenced by the evidence. And so being a responsible metaphysician becomes key here. Knowing what one's metaphysical commitments are is essential since they may need further justification than we have provided to licence the inferences being drawn, or they may need revising if new evidence demands it.

As Goldberg (1995) argues: "This is not to say that all instances of isolated strong dissociations are theoretically useless. This is to say, however, that they must be approached with a degree of wariness, pending the demonstration of their high prevalence in the presence of a particular lesion location, or/and converging evidence from other sources'' (p. 195). Here we can see exactly how a lack of care in tending to one's metaphysical commitments have led to problems in one of the most common methods in learning about psychological mechanisms.

### 3.2 Dedicated Mechanisms for Core Emotions

For a very different example, let us consider again the study of emotion. Some of the most compelling evidence for the idea that certain core emotions like happiness, fear, anger, and sadness are natural kinds (i.e. that each corresponds to a distinct innate physiological mechanism that all humans share), is that certain emotions are recognized across cultures and are universally associated with certain kinds of facial or behavioural expressions (e.g. Ekman et al. 1969, Ekman & Friesen 1971; Izard 1971; Ekman et al. 1987; Ekman 1992; Russell 1994; Elfenbein & Ambady 2002). This universality in facial and behavioural expressions, and our cross-cultural agreement in grouping such expressions together under core emotion categories, suggests that we share the same set of basic emotion mechanisms which produce those same facial and behavioural expressions in us all (which is why we seem to be able to easily group them accordingly across cultures).

Some of the most influential experiments intended to show this demonstrate that pairing different sorts of facial expressions with certain core emotional concepts in different languages results in widespread cross-cultural agreement as to which emotions are to be paired with which facial expressions (Ekman et al. 1969, Ekman & Friesen 1971; Izard 1971; Ekman et al. 1987; Ekman 1992). Other experiments involved telling subjects a short story, and asking them to vocalize in their native

language an emotional expression in response to the story being told. Again, substantial agreement was found (Sauters et al. 2010, 2015; Cordaro et al. 2016).

However, one of the central objections to these studies has been that they only appear to provide evidence for shared cross-cultural emotion recognition because working scientists unintentionally build certain contentious metaphysical commitments that they already believe into the experimental protocols of the experiments without realizing it. For instance, Lisa Barrett argues that there is *not* in fact genuine universal agreement that certain facial or behavioural expressions correspond to core emotion types, nor evidence that core emotions correspond to distinct types of physiological mechanisms (Barrett 2006, 2017). Instead, she argues that it is the emotional *concepts* and their learned usage that dictates how and when certain facial or behavioural expressions tend to be categorized together as falling under one emotion type as opposed to another. Without such emotion concepts structuring a subject's categorization of facial or behavioural expressions, subjects do not tend to group the relevant facial or behavioural expressions together in the same way. As a result, it is not at all clear that there are shared kinds of facial or behavioural expressions that correspond to core emotions across cultures, or that they are universally recognized as such. This undercuts the force of the evidence for the natural kind status of core emotions, and the idea that there therefore must be dedicated mechanisms for the core emotions that all humans share. There is no mechanistic explanation for core emotions to be found, since they are socially and linguistically constructed and not distinct phenomena which are the product of distinct mechanisms.

But if this is the case, then why do so many studies show universal agreement across different linguistic cultures in groupings of facial expressions, or in the emotional evaluation of stories? The reason, according to Barrett, is that such studies typically force subjects to classify facial expressions or story vignettes based on *a specific selection of emotional concepts corresponding to core emotional categories*. In this respect, subjects are being taught to group expressions in the appropriate way

according to the relevant core emotional concepts before the experiments are carried out. When subjects did not have the pre-existing emotional concepts, they were taught them for the purposes of the experiment.

For instance, Barrett notes of experiments conducted by Sauters et al. (2010) that:

> After the Himba participants heard an emotion story but before they listened to any sound pairs, they were asked to describe how the target person in the story was feeling. To help them in this task, Sauter and colleagues "allowed participants to listen several times to a given recorded story if needed, *until they could explain the intended emotion in their own words*." Whenever Himba participants described something other than the English emotion concept, they received negative feedback and were told to try again. Test subjects who were unable to provide the expected description were disqualified from the experiment. In effect, Himba participants were not permitted to listen to any sounds, let alone pick the ones that matched the story, until they had learned the corresponding English emotion conceptions. (2017, p.50)

Meanwhile, when others tried to carry out the same experiments *without* first training subjects in the use of the leading emotional categories, then the cross-cultural consensus in grouping facial expressions, or responding to stories, was lost almost entirely (see: Lindquist et al. 2006; Widen et al. 2011; Crivelli et al. 2015; Crivelli et al. 2016; Barrett 2017). This suggests that the apparent consensus was due to training subjects in the appropriate concepts, and not due to some set of innate emotion mechanisms that we all share. Barrett notes that this problem is prevalent in the majority of studies that claim to argue for cross-cultural consensus of emotional recognition of core emotions, since they all…

> …use the basic emotion method, which you have just seen contains a secret stash of concept knowledge about emotion. If humans actually had an inborn ability to recognize emotional expressions, then removing the emotion words from the method should not matter… but it did, every single time. There is very little doubt that emotion words have a powerful influence in experiments, instantly casting into doubt the conclusions of every study ever performed that used the basic emotion method. (2017, p.52)

Barrett ultimately concludes from this that "emotion concepts are the secret ingredient behind the success of the basic emotion method. These concepts make certain facial configurations appear universally recognizable as emotional expressions when, in fact, they are not." (2017, p.51).

My intention here is not to argue that there is no compelling evidence for the natural kind view of emotions (see, for example: McCaffrey, 2016; Celeghin et al., 2017). The point instead is to note that by not paying attention to their metaphysical commitments, many psychologists studying emotion may have unwittingly built assumptions of natural kinds *into their experimental designs* when forcing their subjects to learn the appropriate core emotion concepts. This in turn produced potentially distorted results which seemed to confirm the metaphysical commitments that were already implicitly guiding their methodology. On the other hand, if we adopt different metaphysical commitments regarding emotions, then the same experimental results (e.g. the apparent cross-cultural agreement of associating facial and behavioural expressions with core emotion terms) can be explained by the structuring of the emotion concept itself, and not by any universally shared physiological mechanisms underlying core emotions.

Here then we see again the dangers of not tending to one's metaphysical commitments. Innocent and well-intentioned experimental procedures can result in potentially distortive data if scientists are not paying attention to the metaphysical commitments that underlie the construction and

application of such procedures. By making these metaphysical commitments explicit, understanding how they structure our methodological practices, and understanding when and how they may need to be better justified, revised, or reconsidered, we can avoid being unintentionally pulled down misleading inferential avenues without noticing. In other words, we must be cautious about the sorts of metaphysical commitments we take on board, and when they ought to be replaced or modified. We likewise need to be careful about the sorts of inferences we draw from the different metaphysical commitments we adopt. These can radically reshape how we study psychological mechanisms, the way in which such mechanisms are defined, and the methods we use to study them.

Of course, just as carelessness with our metaphysical commitments can lead us astray, so too can learning to be responsible metaphysicians provide new insights into contemporary scientific debates. In the section to follow, I will discuss how learning to be good metaphysicians can help us to overcome certain problems regarding the lack of unity or integration currently plaguing the scientific study of psychological mechanisms.

## Section 4: Metaphysical Commitments and Unification

Given that the study of mechanisms is a deeply interdisciplinary process, one of the biggest obstacles currently facing the scientific study of psychological mechanisms is how the information from these different domains, which each invoke distinct concepts, theories, and models, can be effectively integrated to create a coherent account or explanation of the mechanism. Poldrack et al., for instance, claim that cognitive science "faces an increasingly critical challenge: How can we integrate knowledge from an exploding number of studies across multiple methodologies in order to characterize how mental processes are implemented in the brain?" (2011, p.1) Jacqueline Sullivan likewise claims that "over the past two decades scientists and philosophers have noted that rampant conceptual and

methodological pluralisms in psychology and neuroscience are impediments to conceptual and explanatory progress" (2017, p. 132).

Part of what complicates this problem is that different domains of science, given their limited scope and resources, often must simplify and idealize aspects of the phenomenon that fall outside their particular jurisdiction or focus. Neurobiologists studying the mechanisms responsible for psychological phenomena like sexual orientation often do not have the resources or time to include the information provided by behavioural genetics, despite such information providing essential insights into various components of the mechanism. Instead, the models from neurobiology often abstract away from, simplify, or idealize genetic details out of practical necessity, just as models in behavioural genetics will similarly idealize or simplify neurobiological details (for details, see: Longino 2006, 2013). This means that:

> Each approach employs methodologies that require particular ways of understanding the causal space. Some phenomena regarded as causally active in one approach are simply not included in another. These differential selections result in incongruous causal spaces. (Longino, 2006, p. 118)

This makes attempts to integrate the various models exceedingly difficult. The various models from different scientific domains will not fit together neatly since they each idealize, distort, or simplify different features of the target mechanism in incompatible or contradictory ways (see: Mitchel 2002; Hochstein 2016a, 2016b).

All of this means that finding a way to directly integrate the different models, methodologies, and theories from different domains into one single gigantic unified model or theory is likely not a realistic option. The conflicting simplifying assumptions made by the various models and theories would

result in a Frankenstein's monster of a model full of contradictory claims. And so how do we generate a coherent understanding of a complex mechanism, one which incorporates the insights of all these different incompatible methods and models?

It is here that being responsible metaphysicians can help us. Understanding how seemingly incompatible models from different scientific domains or research traditions can inform, constrain, and influence one another requires paying close attention to what metaphysical commitments are implicitly built into those models, or used as background conditions in their application. The various simplified models we created have particular metaphysical assumptions as their foundation. Knowing what those assumptions are is key to being able to find points of contact between the different models so as to understand how we can draw inferences from one to another.

Take, as an example, the Spaun model developed by Eliasmith et al. (2012). Spaun is a large-scale brain model which includes approximately 2.5 million simulated neurons. Typically, large scale brain models focus on characterizing organizational features of neural populations and their causal interactions, but do not identify how such neural activity connects to complex psychological behaviours. Spaun bridges this gap by demonstrating how neural mechanisms can coordinate to carry out, and flexibly switch between, a host of different complex psychological tasks. It does this by identifying certain principles by which spiking neurons implement neural representations which can then be easily manipulated to carry out a range of different computations needed for the different kinds of tasks (Eliasmith et al., 2012; Eliasmith, 2013).

In order to effectively represent such principles however, the model must work with numerous simplifying and idealizing assumptions. First, the model works with point neurons, representing neurons as having no morphological or physical characteristics and instead as mere mathematical points with appropriate input/output relations. Second, the model treats neural inputs as if they were a linear combination of synaptic currents when real neural inputs are not. Third, the model is not nearly as

adaptive as a real brain. Fourth, Spaun is unable to learn completely new tasks, which real brains can. And fifth, the variability in real spiking neurons are not always reflected in the variability of spiking neurons within the model (for more on the idealized nature of Spaun, see: Eliasmith et al., 2012; Stöckel et al., 2017). How then do we integrate the principles identified by the Spaun model with other more accurate biological models that characterize neuron morphology, electrophysiological models which more accurately capture neural input/output relations, and psychological models that characterize the complex psychological behaviour that Spaun is attempting to mechanistically explain?

This can be done by keeping track of the metaphysical commitments used in the creation and application of Spaun, and how they relate to the metaphysical commitments implicit in the application and creation of models from different scientific domains. For instance, while Spaun itself includes no morphological details of the neurons it represents, the construction of Spaun uses morphological details provided by other models as constraints on how their simulated neurons ought to behave. In other words, consistency with the behaviour of more accurate biophysical models was built into the behaviour of Spaun's simulated neurons, even though the biophysical details themselves were not included in the model. As Eliasmith et al. note, the "model embodies neuroanatomical and neurophysiological constraints, making it directly comparable to neural data at many levels of analysis." (2012).

So while Spaun itself idealizes or distorts certain anatomical details identified by other scientific models, the construction and application of Spaun is still *implicitly metaphysically committed* to those details. By understanding what metaphysical commitments the Spaun model implicitly adopts, we can determine exactly how other seemingly incompatible models may share points of agreement, and thus how to draw inferences across models. To flesh this idea out, consider that while more physiologically detailed models make claims about neuron morphology that contradicts the way in which Spaun characterizes its point neurons, the creators of Spaun never *intended* their model to accurately represent neuron morphology. In this regard, their metaphysical commitments to the biophysical details

not explicitly included in their model allowed them to treat point neurons as idealizations while still creating neuron behaviour that is consistent with more biophysically accurate models. In essence, the implicit metaphysical commitments upon which Spaun were built align with the metaphysical commitments about neuron morphology explicitly stated in more accurate biophysical models, despite the fact that Spaun itself idealizes and distorts such details. By keeping track of these metaphysical commitments, we can understand how models which make contradictory claims can find points of contact by sharing certain metaphysical commitments which allows each to inform, and constrain, the other.[2]

What all this means is that integration and unification in cognitive science will not necessarily be a matter of amalgamating all our data into one giant model or theory. Instead, it may involve understanding how collections of incompatible models can each contribute to the same coherent understanding of a complex mechanism by drawing inferences across their shared implicit metaphysical commitments. By identifying these shared commitments, we can find points of contact between the

[2] Of course, not all obstacles to integration can be easily addressed by keeping track of our metaphysical commitments, or using them to find points of contact between models. In some cases, the metaphysical commitments underlying the different models may themselves be in conflict. In such cases, the dispute is not about how to integrate the information generated from different models or methodologies, but about the underlying metaphysical commitments of the scientists who construct and apply those models. For instance, in the case of Spaun, disputes have arisen regarding exactly which sorts of underlying physiological details are truly essential for the metaphysical production of large-scale dynamics or cognitive behaviours and whether the construction and application of Spaun is implicitly or explicitly committed to the appropriate details (Eliasmith & Trujillo 2014; Miłkowski 2016b). Such metaphysical disputes provide a genuine challenge to integration that goes beyond merely finding points of contact between different idealized models. Other obstacles to integration are more methodological in nature. Jacqueline Sullivan, for example, argues that distinct experimental protocols used by different neuroscientific labs have called into question whether the findings of one lab can be effectively applied to the findings of another (Sullivan 2009). This too is not something the account provided here can easily solve. This account is sadly not a cure-all for the problems facing unification or integration in the scientific study of psychological mechanisms. It does, however, provide a guide for how some seemingly incompatible models and theories might find points of agreement in order for us to begin to build connections between them. As such, it brings us a step closer to understanding how various kinds of seemingly contradictory models can contribute to same underlying mechanistic explanation, even if it can't provide a solution to all problems that such a project faces.

different models which act as bridges by which we can use data gathered from one model to inform others.

## Conclusion

The discovery and study of the mechanisms responsible for psychological phenomena is an interdisciplinary process, requiring the resources and methodologies of many different domains. What I have argued in this paper is that one important such domain is philosophy. The study of mechanisms requires engaging in responsible metaphysics: understanding when to make metaphysical claims, how such claims are intertwined with our empirical and experimental practices, and when such claims need to be revised. Others have argued for the importance of philosophy to cognitive science for different reasons (Thagard, 2009), however less attention has been given to the metaphysical commitments required to empirically study cognitive mechanisms.

Being aware of what our metaphysical commitments are, and how they are employed, is extremely important to how we build models, run experiments, and test hypotheses about mechanisms in psychology. We must be cautious of what metaphysical commitments we adopt and why, and careful about the sorts of inferences we draw from those commitments. Carelessness can result in complications and additional problems for the scientific study of mental mechanisms, but conversely getting clear on the metaphysical commitments underlying different scientific models can help us get one step closer to understanding how the plurality of models and theories needed to study psychological mechanisms can relate to one another.

## References:

Anderson, M. (2007) Evolution of cognitive function via redeployment of brain areas. *The Neuroscientist* 13:13–21.

Anderson, M. (2008) Circuit sharing and the implementation of intelligent systems. *Connection Science* 20 (4):239–51.

Anderson, M. (2010). Neural reuse: A fundamental organizational principle of the brain. *Behaviorial and Brian Science* 33: 245-313.

Austin, C. (2017) The Philosophy of Biology. *Analysis* 77 (2): 412-432.

Barrett, L. (2006). Are Emotions Natural Kinds? *Perspectives on Psychological Science, 1,* 28-58.

Barrett, L. (2017). *How Emotions are Made*. Boston: Houghton Mifflin Harcourt.

Bechtel, W. (2005). The Challenge of Characterizing Operations in the Mechanisms Underlying Behavior. *Journal of the Experimental Analysis of Behavior, 84,* 313-325.

Bechtel, W. (2008). *Mental Mechanisms: Philosophical Perspectives on Cognitive Neuroscience*. New York: Lawrence Erlbaum Associates.

Bechtel, W. (2015). Can mechanistic explanation be reconciled with scale-free constitution and dynamics? *Studies in History and Philosophy of Biological and Biomedical Sciences, 53,* 84-93.

Cajal, S. R. (1937). *Recollections of My Life*. Philadelphia: American Philosophical Society.

Celeghin, A., Diano, M., Bagnis, A., Viola, M., Tamietoo, M. (2017). Basic Emotions in Human Neuroscience: Neuroimaging and Beyond. *Frontiers in Psychology* 8.1432

Chirimuuta, M. (2017). Crash Testing an Engineering Framework in Neuroscience: Does the Idea of Robustness Break Down? *Philosophy of Science* 84 (5): 1140-1151.

Cowie, F. & Woodward, J. (2004). The Mind is Not (Just) a System of Modules Shaped (Just) by Natural Selection. in C. Hitchcock (Ed.), *Great Debates in Philosophy: Philosophy of Science*. New York and Oxford: Blackwell. 312-334.

Cordaro, D., Keltner, D., Tshering, S., Wangchuk, D., Flynn, L. (2016). The Voice Conveys Emotion in Ten Globalized Cultures and One Remote Village in Bhutan. *Emotion* 16 (1): 117-128.

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

Craver, C. & Darden, L. (2013). *In Search of Mechanisms*. Chicago: The University of Chicago Press.

Crivelli, C., Carrera, P., Fernandez-Dols, J-M. (2015). Are Smiles a Sign of Happiness? Spontaneous Expressions of Judo Winners. *Evolution and Human Behavior* 36 (1): 52-58.

Crivelli, C., Jarillo, S., Russell, J., ernandez-Dols, J-M. (2016). Reading Emotions from Faces in Two Indigenous Societies. *Journal of Experimental Psychology* 145 (7): 830-843.

Elfenbein, H. & Ambady, N. (2002). On the Universality and Cultural Specificity of Emotion Recognition: A Meta-Analysis. *Psychological Bulletin* 128 (2): 203-235.

Ekman, P. (1992). An Argument for Basic Emotions. *Cognition and Emotion* 6: 169-200.

Ekman, P., & Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology* 17: 124-129.

Ekman, P., Sorenson, E., Friesen, W. (1969). Pan-Cultural Elements in Facial Displays of Emotion. *Science* 164 (4616): 1208-1210.

Ekman, P., Friesen, W., O'Sullivan, M., Chan, A., Diacoyanni-Tarlatzis, I., Heider, K., Krause, R., LeCompte, W., Pitcairn, T., and Ricci-Bitti, P. (1987). Universal and Cultural Differences in the Judgments of Facial Expressions of Emotion. *Journal of Personality and Social Psychology* 53 (4): 712-717.

Elfenbein, H. & Ambady, N. (2002). On the Universality and Cultural Specificity of Emotion Recognition: A Meta-Analysis. *Psychological Bulletin* 128 (2): 203-235.

Eliasmith, C. (2009). Dynamics, Control, and Cognition. In P. Robbins and M. Aydede (Eds.), *Cambridge Handbook of Situated Cognition*. Cambridge University Press.

Eliasmith, C. (2013). *How to build a brain: A neural architecture for biological cognition*. Oxford: Oxford University Press.

Eliasmith, C., Stewart, T., Choo, X., Bekolay, T., DeWolf, T., Tang, Y., & Rasmussen, D. (2012). A large-scale model of the functioning brain. *Science, 338 (6111)*, 1202-1205.

Eliasmith, C., & Trujillo, O. (2014). The use and abuse of large-scale brain models. Current Opinion in Neurobiology, 25, 1–6.

Goldberg, E. (1995). Rise and fall of modular orthodoxy. *Journal of Clinical and Experimental Neuropsychology* 17: 193–208.

Hochstein, E. (2016a). One mechanism, many models: a distributed theory of mechanistic explanation. *Synthese, 193*, 1387–1407

Hochstein, E. (2016b). Giving up on Convergence and Autonomy: Why the Theories of Psychology and Neuroscience are Codependent as well as Irreconcilable. *Studies in History and Philosophy of Science* 56: 135-144.

Hodgkin AL, Huxley AF (1952) A quantitative description of membrane current and its application to conduction and excitation in nerve. *J Physiol* 117:500–544.

Hohwy, J. (2007). Functional integration and the mind. *Synthese* 159: 315-328.

Izard, C. (1971). *The Face of Emotion*. East Norwalk, CT: Appleton-Century-Crofts

Lindquist, K., Barrett, L., Bliss-Moreau, E., and Russell J. (2006). Language and the Perception of Emotion. *Emotion* 6 (1): 125-138.

Longino, H. (2006). Theoretical Pluralism and the Scientific Study of Behavior. In S. Kellert, H. Longino, and C.K. Waters (Eds.), *Scientific Pluralism* (pp. 102-132). Minneapolis: University of Minnesota Press.

Longino, H. (2013). *Studying Human Behavior: How Scientists Investigate Aggression and Sexuality*. Chicago: The University of Chicago Press.

Machamer, P., Darden, L., & Craver, C. (2000). Thinking about mechanisms. *Philosophy of Science*, *67*, 1–25.

McCaffrey, J. (2016, November). *Does the Brain Respect Basic Emotion Theory? Pattern Classification and Correspondence in fMRI Research*. Paper presented at meeting of the Philosophy of Science Association, Atlanta, Georgia.

Milkowski, M. (2013). A mechanistic account of computational explanation in cognitive science. In *Proceedings of the annual meeting of the Cognitive Science Society*.

Miłkowski, M. (2016a). Integrating cognitive (neuro)science using mechanisms. *AVANT*, Vol. VI (2): 45-67.

Milkowski, M. (2016b). Explanatory completeness and idealization in large brain simulations: a mechanistic perspective. *Synthese* 193: 1457–1478

Mitchell, S. (2002). Integrative Pluralism. *Biology and Philosophy*, *17*, 55-70.

Palecek, M. (2017). Modularity of Mind: Is it Time to Abandon This Ship?. *Philosophy of the Social Sciences* 47 (2): 132-144.

Patterson, K. and Plaut, D. (2009). "Shallow Draughts Intoxicate the Brain": Lessons from Cognitive Science for Cognitive Neuropsychology. *Topics in Cognitive Science* 1 (1): 39-58.

Poldrack, R. A. (2006) Can cognitive processes be inferred from neuroimaging data? *Trends in Cognitive Sciences* 10: 59–63.

Poldrack R., A. Kittur, D. Kalar, E. Miller, C Seppa, Y. Gil, D. Parker, F. Sabb, Bilder, R. (2011). The Cognitive Atlas: Toward a Knowledge Foundation for Cognitive Neuroscience. *Frontiers in Neuroinformatics*, *5*, 17.

Russell, R. (1994). Is There Universal Recognition of Emotion from Facial Expressions? A Review of the Cross-Cultural Studies. *Psychological Bulletin* 115 (1): 102-141.

Samuels, R. (1998). Evolutionary Psychology and the Massive Modularity Hypothesis. *British Journal for the Philosophy of Science* 49: 575-602

Sauter, D., Eisner, F., Ekman, P., and Scott, S. (2010). Cross-Cultural Recognition of Basic Emotions Through Nonverbal Emotional Vocalizations. *Proceedings of the National Academy of Sciences* 107 (6): 2408-2412.

Sauter, D., Eisner, F., Ekman, P., and Scott, S. (2015). Emotional Vocalizations Are Recognized Across Cultures Regardless of the Valence of Distractors. *Psychological Science* 26 (3): 354-356.

Seidenberg, M. S. (1988). Cognitive neuropsychology and language: The state of the art. *Cognitive Neuropsychology* 5: 403–426.

Singer, S. & Nicolson, G. (1972). The fluid mosaic model of the structure of cell membranes. *Science, 175 (4023)*, 720–731.

Stöckel, A., Voelker, A. R., & Eliasmith, C. (2017). *Point neurons with conductance-based synapses in the neural engineering framework* (Tech. Rep.). Waterloo, ON: Centre for Theoretical Neuroscience.

Stocker, A. & Simoncelli, E. (2006). Noise Characteristics and Prior Expectations in Human Visual Speed Perception. *Nature Neuroscience, 9*, 578–585.

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

Sullivan, J. (2017). Coordinated pluralism as a means to facilitate integrative taxonomies of cognition. *Philosophical Explorations, 20*, 129-145.

Thagard, P. (2009). Why Cognitive Science Needs Philosophy and Vice Versa. *Topics in Cognitive Science, 1:* 237-254.

Trumpler, M. (1997). Techniques of intervention and forms of representation of sodium-channel proteins in nerve cell membranes. *Journal of History of Biology, 30*, 55–89.

Uttal, W. (2001). The New Phrenology The Limits of Localizing Cognitive Processes in the Brain. Cambridge, Massachusetts: The MIT Press.

Van Orden, G., and Paap, K. (1997). Functional Neuroimages Fail to Discover Pieces of Mind in the Parts of the Brain. *Philosophy of Science* 64: S85-S94

Widen, S., Christy, A., Hewett, K., and Russel, J. (2011). Do Proposed Facial Expressions of Contempt, Shame, Embarrassment, and Compassion Communicate the Predicted Emotion? *Cognition and Emotion* 25 (5): 898-906.