To appear in P. Calvo and J. Symons (ed.) *Routledge companion to philosophy of psychology*. London: Routledge

What is Psychological Explanation?

William Bechtel
Cory D. Wright
Department of Philosophy and Interdisciplinary Program in Cognitive Science
University of California, San Diego

Abstract

Due to the wide array of phenomena that are of interest to them, psychologists offer highly diverse and heterogeneous types of explanations. Initially, this suggests that the question "What is psychological explanation?" has no single answer. To provide appreciation of this diversity, we begin by noting some of the more common types of explanations that psychologists provide, with particular focus on classical examples of explanations advanced in three different areas of psychology: psychophysics, physiological psychology, and information-processing psychology. To analyze what is involved in these types of explanations, we consider the ways in which lawlike representations of regularities and representations of mechanisms factor in psychological explanations. This consideration directs us to certain fundamental questions, e.g., "To what extent are laws necessary for psychological explanations?" and "What do psychologists have in mind when they appeal to mechanisms in explanation?" In answering such questions, it appears that laws do play important roles in psychological explanations, although most explanations in psychology appeal to accounts of mechanisms. Consequently, we provide a unifying account of what psychological explanation is.

1. What does *psychological explanation* refer to?

Frequently the expression *psychological explanation* is used as a catch-all term denoting any attempt to understand phenomena related to intelligent behavior. The philosophy of psychology would benefit from a more precise analytical conception of what constitutes explanation in psychology. The approach we take in this chapter focuses on what sort of explanatory practices are distinctive of experimental or scientific psychology. By noting three prototypical examples from diverse subfields of experimental psychology, we hope to provide a context for further analytical discussion. These examples point toward two very different models of explanation, which have also been discussed in the broader context of philosophy of science—namely, nomological explanation and mechanistic explanation. Two initial questions are therefore pertinent:

- To what extent are laws needed in explanations in psychology?
- What do psychologists have in mind when they appeal to mechanisms in explanation?

In some subfields of psychology, the mechanisms appealed to are characterized as *information-processing* mechanisms, which raises an additional question:

- Are information-processing mechanisms interestingly different from other mechanisms, and are there special challenges psychologists confront in advancing explanations in terms of information processing mechanisms?

To understand better what is involved in mechanistic explanation in psychology, we will specifically focus on how the decomposition and localization of psychological processes figure in psychological explanation. This will, in part, require us to examine the role that understanding the brain plays. We will then turn to a set of general questions about explanation that are answered differently in terms of nomological and mechanistic accounts:

- What are the tools employed in representing and reasoning about psychological phenomena?
- Can philosophy say anything constructive about the discovery and development of explanations?
- What sort of scope or generality applies to psychological explanation?

Lastly, we will consider the extent to which the framework of mechanistic explanation might actually provide a unifying perspective within which the role of laws can be accommodated, and whether there are forms of explanation in psychology that lie beyond the mechanistic perspective.

2. Three examples of explanation in psychology

A cursory look at the history of the discipline—particularly, in the 19th and 20th centuries—reveals a rich collage of explanatory forms and strategies that have been invoked to render intelligible our mental lives and behavioral repertoires. We will begin by giving brief sketches of three in particular, drawn respectively from psychophysics, physiological psychology, and information-processing psychology.

(*i*). *Psychophysics*. Some of the earliest roots of the experimental tradition in psychology are found in *psychophysics*, which is the subfield that attempts to identify the relation between physical features of sensory stimuli and the psychological experience of them. For example, Ernst Weber (1834) investigated the relationship between features such as the weight of an object and its perceived heaviness, and concluded that "we perceive not the difference between the things, but the ratio of this difference to the magnitude of the thing compared" (p. 172). This conclusion was later expressed as *Weber's law*, ($\Delta I / I = k$), where I is the intensity of a stimulus, ΔI is the minimum increment over I that is detectable (i.e., *just noticeable difference*), and the value of k is constant ($\approx$0.15 for loudness) except at extreme values for a given perceptual modality. Gustav Fechner (1860) extended Weber's research, adding an assumption of cumulativity so as to obtain a logarithmic function: $\Psi = c \log (I / I_0)$. On Fechner's account, the intensity of a sensation (Ψ) is proportional to the logarithm of the intensity of the stimulus (I) relative to threshold intensity (I_0). The accomplishments of Weber and Fechner were extremely important in demonstrating how to bring the same formal rigor achieved with purely physical phenomena to the domain of mental experience. They directly inspired the pioneering studies of memory by Hermann Ebbinghaus, who revealed mathematical relationships that characterize forgetting, and their efforts were perpetuated with the emergence of mathematical psychology in the 1960s.

In the mid-20th century, S. S. Stevens (1957) proposed a power function rather than a logarithmic function that made stronger—and occasionally more accurate—predictions than Fechner's law. Psychophysical laws such as Stevens are often viewed as providing explanations of individual percepts, insofar as they show that those percepts are instances of a more general regularity. Subsequently, the discovery of such laws has been viewed as an important theoretical contribution to psychology because they describe elegant and often amazingly simple and regular relations between physical and psychological phenomena. Unfortunately, the nomic regularities described by such laws are themselves left unexplained—e.g., while the relation described by Stevens' power law is regarded as ubiquitous, there have been no accounts of why it occurs.[1]

(*ii*). *Information-processing psychology*. Technological developments (e.g., telephone, computer) led theorists such as Claude Shannon and Warren Weaver to provide a mathematical characterization of information, and others such as Alan Turing and Emil Post to explore how it might be manipulated formally in mechanical systems. These and other developments (e.g., Chomsky's characterization of the requirements on any automaton that could process the syntax of a natural language), fostered psychologists' attempts to characterize mental activity as involving the processing of information (e.g., Miller, 1956). In attempting to characterize information-processing operations, psychologists reintroduced a procedure that was initially developed by Frans Cornelis Donders (1868) for measuring the time required for particular mental activities. In Donder's use of the technique, a shock was applied to either a person's left or right foot, and the person had to respond expeditiously by pressing a telegraph key with the corresponding hand. He determined that when subjects were uninformed about which hand would be shocked, they required an additional .067 second to respond.

A particularly elegant use of reaction times to assess mental operations occurs in an early study in which Saul Sternberg investigated the retrieval of information from short-term memory (retention of information over a period of seconds to minutes during which there is no interruption). He required subjects to study a list of digits (e.g., 6, 9, 2, 4), and afterwards asked them to determine whether a particular digit (e.g., 5) was on the list, measuring the time it took them to respond. Sternberg evaluated three hypotheses: (a) subjects mentally examine all the items in memory simultaneously, (b) subjects examine them serially, stopping when they encounter the target item, and (c) subjects examine them serially, but complete the list regardless of whether they encounter the target. These hypotheses predicted different reaction times. If all items were examined simultaneously, the length of time required to answer should remain constant regardless of the length of the list. If subjects examined items serially and stopped when they reached the target, then the times should be longer for longer lists, but the times for positive responses should be shorter than for negative responses (as the target item would typically be encountered before the list was completed). Finally, if subjects examined items serially but completed the list before responding, then the reaction times should be the same for both positive and negative responses, and longer the longer the list. Sternberg discovered that it was the last prediction that was true, providing positive evidence for hypothesis (c). More specifically, it took subjects on average 392.7 + 37.9 *s* milliseconds to respond (where *s* was the number of items on the list to be remembered).

---

[1] Interestingly, Weber himself, in applying the relation he discovered to line length, noted that it was inconsistent with a possible perceptual process in which "the mind […] counts the nerve endings touched in the retina" to assess length; but he offered no alternative procedure that would give rise to the relation he discovered.

(*iii*). *Physiological psychology*. A third main type of psychological explanation involves characterizing the physiological processes governing psychological phenomena. There are many examples of physiologically-based explanations, from the opponent-processing of visual input to the neural oscillatory patterns involved in sleep and wakefulness. One of the more intriguing is the electrophysiological explanation of pleasure and reward, which was inaugurated by James Olds and Peter Milner's (1954) serendipitous discovery that electrical stimulation of certain anatomical sites of rats' brain causes them to work extremely hard, suggesting they found the stimulation rewarding. Electrophysiological self-stimulation studies were immediately performed on a variety of animals, from snails to monkeys, and within 10 years it had become entirely clear just how rewarding the stimulation was: Olds (1955) found that rats would self-stimulate over 50,000 times in a 26-hour period, and Valenstein and Beer (1964) reported that rats averaged 29.2 self-stimulating responses per minute for three weeks without respite. In human studies, subjects were found to self-stimulate various midbrain structures in ½ second trains up to 400 times per hour (Bishop, Elder, & Heath, 1963; Heath, 1963). Olds and Milner quickly discovered that self-stimulation behavior can be elicited from stimulation spanning the full length of the medial forebrain bundle (MFB), and so suggested that the activation patterns of this pathway directly mediate both the hedonic effects of, and complex behavioral responses to, all pleasures and rewarding stimuli. The MFB is a major neuronal band traversing the brain ventrally from the brainstem to numerous sets of limbic and cortical structures across the cerebrum—most notably, the ventral tegmental area (VTA), lateral hypothalamus (LH), nucleus accumbens (NAc), and the frontal and prefrontal cortex (pFC)—and has also been proposed as a major structure in the final common pathway of reinforcement and approach behavior (Wise 1989: 410).

Eventually, Olds and Milner's initial explanation that reward is governed by the activity of this neuronal band did not do justice the complexity found in this band, or to the mesocorticolimbic system more generally. Valenstein and Campbell (1966) showed that self-stimulation behavior was not disrupted even when virtually all of the MFB was ablated. So, although these fibers are clearly involved in producing self-stimulation behavior governing animals' experience of reward, the initial 1-to-1 identification of reward function with MFB activity was a vast oversimplification. This suggested the need for more complex localizations in contiguous neural systems and subsystems. For example, the initial localization to the MFB was made much more precise with the advent of dopamine-selective (DA) neurotoxins such as 6-hydroxydopamine HBr, which can be used to lesion particular MFB substructures and thus cause the animal to become oblivious to the hedonic valence of most rewards. Better explanations also require increasingly complex decompositions of a given mechanistic activity into component operations and suboperations, followed by an assessment of the extent of their interaction and integration. Complex localization usually evolves from a better understanding of the organization of those systems, taking into account the fact that mechanistic systems are not always serial, but employ nonlinear cycles, parallel processing, positive and negative feedback loops, etc. As a result of increasingly precise complex decompositions and localizations, explanations of reward have converged on several highly-interconnected anatomical structures constituted, in part, by the long and fast myelinated DA axons projecting afferently from the VTA to the NAc.[2]

[2] Other catacholaminergic neurotransmitter systems were explored early on—particularly, norepinephrine (NE) fibers in substructures of the MFB (which is also a major point of convergence for NE systems). Yet, by showing that lesions to the dorsolateral noradrenergic bundle and other NE sites did not abolish self-stimulation

2. From laws to information-processing mechanisms

Traditionally, accounts of explanation in philosophy of science have given pride of place to laws and lawlike generalizations. On the deductive-nomological account of explanation, a given phenomenon is explained by showing that a statement of it can be derived from one or more laws and specification of initial conditions. For example, to explain the period (time required to complete a swing, e.g., from the far left back to the far left) of a given pendulum, a physicist would derive the value from the pendulum's length (*l)* and the law discovered by Galileo

$$T = 2\pi\sqrt{\frac{l}{g}}$$

in which *g* is gravitational force. There has been a great deal of philosophical discussion of just what is required of a law, what they actually describe, and how they explain. Most accounts hold that a law is at least a universally qualified conditional statement which specifies what *must* happen when the initial conditions are satisfied. Psychophysics exemplifies this strategy by appealing to laws such as those advanced by Weber, Fechner, or Stevens to explain the percepts that people experience as a result of experiencing stimuli with specific features.

Beyond psychophysics, however, there are few subfields of psychology in which researchers have established relations between variables that are referred to as *laws*. Psychologists appeal to and discuss laws relatively infrequently. A recent bibliometric study of abstracts from the PsycLit database from 1900–1999 (> 1.4 million) revealed an average of only .0022 citations for the term *law*; as Figure 1 shows, that number has continuously dwindled over the last few decades—down to .001 from 1990–1999 (Teigen, 2002).

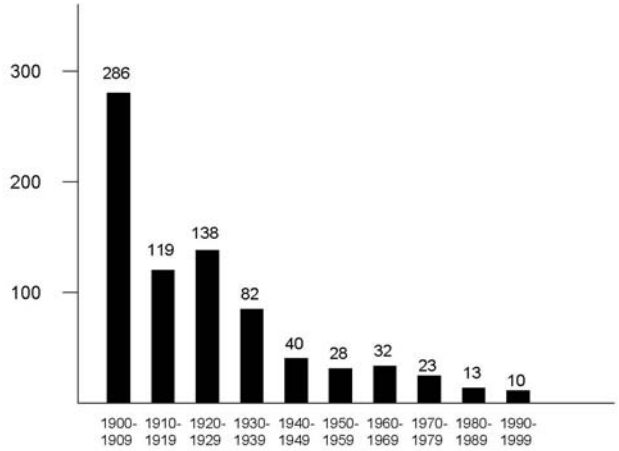

Figure 1. Occurrence of *law* in PsychLit abstracts per 10,000 entries (redrawn from Teigen, 2002).

---

behavior from the locus coeruleus, Clavier *et al*. (1976) showed that NE was not as significant in mediating reward function as previously thought. Later reinterpretations of NE data (e.g., Wise 1978) confirmed that this transmitter system responsible for playing the role attributed to NE was actually DA.

As Robert Cummins (2000) pointed out, such relations tend to be called *effects*, not *laws*. Indeed, there are numerous examples of such effects discussed throughout psychological literature. One well-known example is the Garcia effect, which is the tendency of animals to avoid foods eaten prior to experiencing nausea, even if the proximal cause of nausea was something other than the food. The difference between referring to *laws* and *effects* is not merely terminological. As, Cummins argued—and as we noted in the case of psychophysical laws—appeals to effects are typically not explanatory. Instead, they serve to describe phenomena that in turn require explanation and elucidation—i.e., the explanandum.

Within accounts of explanation, philosophers of science often distinguish between *empirical* (or *observational*) and *theoretical* laws. Since effects are descriptions of the relations between empirically measured variables, they do seem to correspond to empirical laws. How, then, can psychologists explain these empirical laws? The strategy described in many philosophical accounts is to explain empirical laws by deriving them from theoretical laws. In classical physics, for example, Newton's theoretical laws concerning forces can be invoked to explain the empirical law describing the pendulum stated above. The challenge in applying this strategy to psychology is that unclear what the theoretical laws are to which one might appeal in explanations. An alternative is to appeal to the laws of more basic sciences (e.g., neurophysiology). Unfortunately, this approach is likewise problematic, as there are even fewer examples of relations called *laws* in physiology or biology.

Marcel Weber (2005) has argued compellingly that the laws invoked in explaining physiological phenomena are drawn from physics and chemistry. The laws of physics and chemistry are certainly not irrelevant to psychological phenomena. Psychological phenomena are realized in brains comprised of neurons that generate action potentials, and the electrical currents that constitute the action potential are governed by principles such as Ohm's law. However, psychologists seldom allude or explicitly appeal to such basic physical laws in their explanations (for reasons addressed below). Rather, when psychologists (as well as physiologists and many other investigators in the life sciences) offer explanations that go beyond the empirical laws or effects they identify, they frequently suggest that such explanations model a *mechanism*—i.e., a composite system whose activity is responsible for the target phenomenon (Wright & Bechtel, 2006; Bechtel, in press). Thus, in the example from physiological psychology, after Olds and Milner seemingly found a neural locus for reward, other investigators began to identify other components of the brain's reward system and develop accounts of their role in reward function. Information-processing psychology is also engaged in a quest to identify mechanisms, as in the example from Sternberg, although the mechanisms are of a special kind—namely, those that process information.

The appeal to mechanisms in explanation, despite being a staple of both biology and psychology, has received little attention in philosophy until the last two decades. Recently, a variety of philosophers turned their attention to the conceptual analysis of what is meant by *mechanism* and *mechanistic explanation*. A mechanism is simply a composite system organized in such a way that the coordinated operations of the component parts constitute the mechanistic activity identified with the explanandum. Hence, a central feature of mechanisms is that they are mereological: the mechanism as a whole is comprised of component parts, the orchestrated operation of which constitute its function(s). Not infrequently, the parts of a mechanism are

themselves mechanisms consisting of component parts at lower levels, which implies that mechanisms are intrinsically hierarchical. We will return to this issue later.

The primary challenge confronting researchers advancing a mechanistic explanation is how to model the appropriate decomposition of that composite system into its component parts and their operations, and to determine how those parts and operations are organized. In giving a mechanistic explanation, investigators must represent the system (either in their heads or in terms of external symbols such as diagrams or physical models) and make inferences about how the parts' operations suffice to account for the mechanism's activity, and how that activity thereby is the phenomenon to be explained. Often these inferences result from simulating the mechanism (whether mentally, or by invoking computer programs or computer models, etc.).

We will return shortly to the strategies by which psychologists develop mechanistic explanations; in the meantime, note that the mechanisms appealed to by many psychologists are of a distinctive kind. Rather than serving to transform chemical substances as in basic physiology (e.g., the process of synthesizing proteins from free amino acids), many of the mechanisms appealed to in psychology are those that serve to regulate behavior or process information. Reward mechanisms, for example, figure in control systems responsible for the information relevant to approach and consummatory behavior. Other psychological mechanisms, especially in primates, are a stage removed from the actual regulation of behavior, and are instead involved in tasks like planning future behaviors or securing information about the world. The advent of digital computers provided a model for understanding how mechanisms process information. Some of the physical processes within the mechanism serve as representations for other entities and processes (i.e., they have as their content these other entities or processes) and the manner in which these states are operated on is appropriate to their content. Specifying how representations have content has been a major concern for philosophers of psychology (Dretske, 1981; Millikan, 1984); but, for the purposes of characterizing information-processing mechanisms, the key point is that those mechanisms use the representations to coordinate the organism's behavior with respect to or in light of the represented features of its environment.[3]

3. Decomposing the mind into operations and localizing them in the brain

The major tasks in developing mechanistic explanations in psychology are to identify the parts of a mechanism, determine their operations, discern their organization, and finally, represent how these things constitute the system's relationship to the target explanandum. There are many ways of decomposing a mechanism into parts; but the explanatorily-interesting parts are those that figure in the operations—i.e., *operative parts*. These are the parts that either perform the operations or are operated on in an operation. Both identifying the parts and determining the operations involved in a mechanism can be challenging activities. Many ways of decomposing a mechanism into parts fail to identify operative ones. Neuroanatomists in the 19th century put considerable effort into identifying and naming the gyri of the cortex; but sulci and gyri turned out to be the indirect result of folding the sheet of cortex, and not to be operative parts. The areas identified via cytoarchitecture by Korbinian Brodmann and numerous other investigators at the

[3] (see Newell, 1980, for a developed account of such mechanisms as physical symbol systems).

beginning of the 20th century more closely approximated functional areas, although they often turned out to contain distinct component parts within them.

If anything, identifying operations is even more difficult. Smoothly functioning mechanisms are organized such that the various operations engage each other in coordinated manner. Behaviorists eschewed such projects, doubting that psychologists' attempts to identify operations inside the head would invariably invoke the same imprecise mentalistic vocabulary that was used to label the overall activity. Where behaviorists and others declined to tread, researchers in the newly-emergent tradition of cognitive psychology in the mid-1950s pushed forward in the attempt to reverse engineer the mind. They approached their task by hypothesizing types of operations that transformed representations so as to produce the overall information-processing activity. As the example from Sternberg reveals, they often drew inspiration from the then-new field of computer engineering, and postulated activities of storing, retrieving, and operating on representations. (These operations in computers were in turn inspired by the activities of humans who were employed to compute functions and did so by repeatedly reading and writing symbols from a page, performing simple computational operations on these symbols, and then writing the results on the page. The representations appealed to by psychologists when they adapted this account to explain human activities were themselves hypothetical constructs postulated to account for the overall performance.)

From the 1950s until the 1990s, researchers in cognitive psychology had to pursue their efforts to identify mental operations with little guidance from neuroscience, primarily because there were few tools for characterizing brain regions that realize psychological phenomena. An oft-invoked resource was the analysis of patients with brain damage. From their deficits in overall performance, researchers advanced inferences as to what the damaged area contributed to normal function. For example, on the basis of deficits in articulate speech resulting from a damaged area in the left prefrontal cortex, Paul Broca (1861) famously proposed that this area was responsible for articulate speech. Broca's studies, and most studies within the field known as *neuropsychology*, have relied on damage occurring either from accidents, tumors, or strokes, which typically do not respect the boundaries of operative parts in the brain. An influential case of a cognitive deficit in the early development of cognitive psychology involved William Scoville's removal of the hippocampus and surrounding areas of the medial temporal lobe in an epileptic patient H.M (Scoville & Milner, 1957). The surgery successful reduced H.M.'s seizures, but left him with severe amnesia for post-surgical events (anterograde amnesia) and—in a more graded fashion—for events in the years prior to surgery (graded retrograde amnesia). Although H.M. exhibited profound amnesia for events in his life and was unable to learn new facts, he could learn new skills (albeit not remembering, and therefore denying, that he had learned new skills). This case provided powerful support for a distinction between declarative or explicit memory and procedural or implicit memory, and also initiated a program of research directed at determining hippocampal contributions to the acquisition of new declarative memories. That H.M. retained memories from much earlier in his life indicated that the hippocampus was not the locus of long-term storage and the graded retrograde amnesia, and further indicated that the consolidation of long-term memories was protracted, lasting several years.

Broca's case and that of H.M. reflect attempts to relate operative parts in the brain to the operations they perform—what is often called *localization*. Naturally or even surgically-induced lesions in humans typically do not involve a precisely delineated brain area, rendering localization claims difficult. When cognitive psychologists began advancing decompositions of mental function, neuroscientists working with other species (especially cats and monkeys) were developing more precise tools for localizing operations in the brain. In addition to increasingly precise surgically-induced lesions, neuroscientists developed techniques either to record from or induce electrical activity into individual neurons. This strategy proved especially effective in the case of visual processing. By systematically varying visual stimuli so as to determine what features of a stimulus would generate action potentials in a given cell, investigators have identified brain regions in which cells seemingly process particular features of visual stimuli such as motion, shape, etc. It is possible, for example, to identify cells that respond to the perceived color or motion of the stimulus, not the wavelength or actual motion of the stimulus. Importantly, this research enabled researchers to determine both the brain regions and perceptual operations involved. Although many details remain to be resolved, this research has generated detailed models of the mechanisms involved in visual processing (van Essen & Gallant, 1994).

Research on non-human animals, however, provided little insight into the mental activities of greatest interest to cognitive psychologists (e.g., reasoning, problem-solving, language processing, memory). Cognitive psychologists therefore had to proceed by first hypothesizing how a mechanism might perform the requisite activity, and then test the hypothesis with predictions using more indirect measures such as reaction times and errors made in normal performance. Although their explanatory aim was to identify the task-relevant mental operations, cognitive psychologists more often succeeded in establishing differences between psychological phenomena and showing that they rely on different mental operations without thereby specifying them. For example, in addition to the distinction between declarative and procedural memory, Endel Tulving (1983) advanced a distinction within declarative memory between memory for factual information, including facts about oneself (*semantic memory*) and memory that involves reliving episodes in one's own life (*episodic memory*). Tulving proposed that episodic and semantic memory (plus other types of memory) are due to separate memory systems (Schacter & Tulving, 1994). Others psychologists (Roediger, Buckner, & McDermott, 1999) have questioned whether some of the operations involved in performing different memory tasks are actually the same; unfortunately, little progress has been made in further articulating the nature of these operations.

In the 1990s, cognitive psychologists gained access to a new research technique that altered their potential to use information from the brain in understanding mental operations. Functional neuroimaging, either via positron emission tomography (PET) or functional magnetic resonance imaging (fMRI), provided a means to identify the brain areas in which increased blood flow accompanied a particular psychological phenomenon. Although the actual relationship between neural processing and blood flow has not been established, increased blood flow is commonly assumed to indicate increased metabolism, which in turn indicates increased neural processing in the region.[4]

[4] (See Raichle & Minton, 2006, for discussion and a proposal for the causal relation).

Neuroimaging has provided a means to localize component operations involved in performing cognitive tasks with brain regions in which they are performed, and so has elicited much interest (both popular and academic). However, it is important to consider exactly how neuroimaging figures in psychological explanation. It often seems that the goal of neuroimaging is simply to establish where in the brain cognitive activities are occurring; but this does little to advance the explanatory charge of psychology. It is useful to consider again what neuroscientists accomplished using single-cell recording in the case of visual processing. Their goal was not simply to determine what areas of the brain are involved in vision (though it is suggestive that over ⅓ of the primate brain is involved in visual processing); rather, the functional decomposition of the brain was advanced by determining what operations individual brain areas performed. Recent neuroimaging, for instance showing how the same brain areas are involved in both mnemonic and perceptual tasks suggests that the areas are performing operations required for both, which directs inquiry to questions about what common operations may be involved in perceiving and remembering.

Consequently, learning the various areas involved in task performance plays a heuristic role in investigating the nature of the operations performed. The heuristic also works in the opposite direction: ideas about the operations to be performed can guide inquiry toward the brain areas that perform them. Moreover, the heuristic benefit is often achieved by playing the proposed decompositions into parts and operations off one another, providing grist for revising each account. Accordingly, Bechtel and McCauley (1999) characterize localization in terms of a *heuristic identity theory*. Whereas the classical identity theory in philosophy of mind treated the identification of a mental operation with a particular brain area as an end in itself, the heuristic identity theory treats the goal of identity claims instrumentally—i.e., as advancing the project of discovering mechanisms by helping to understand how component parts and operations are involved in psychological phenomena.

4. Producing and generalizing psychological explanations

The dominant account in 20th century epistemology represented knowledge linguaformally, with propositions reporting sensory experiences justifying, by their logical relations, those propositions that do not. A similar account was adopted in philosophy of science, whereby statements about events are derived from—and thus putatively explained by—knowledge of laws and initial conditions. Moreover, success in deriving not just what had happened but what would happen (*prediction*) was taken as providing evidence that one had fixed upon the correct laws. On this account, which was exemplified in explanations in psychophysics, understanding a scientific explanation required understanding the laws and reasoning appropriate to derive or subsume consequences. Hence, the challenge in discovery was to formulate laws, and this—many philosophers contended—was not a matter of logic and hence not something for which philosophy could provide an account. A few philosophers and artificial intelligence researchers dissented from this pessimistic assessment, and attempted to develop procedures for discovering laws (see, e.g., Holland, Holyoak, Nisbett, & Thagard, 1986; Langley, Simon, Bradshaw, & Zytkow, 1987; Thagard, 1988).

Insofar as laws are taken to have the form of universal generalizations, this perspective offered a straightforward way of generalizing from one case to additional cases—i.e., additional cases were simply instances covered by the same generalization. As we noted, for nomological accounts of explanation, laws provide the basic explanatory resource; in turn, explaining a law then amounts to the ability to derive it from more fundamental laws. These more fundamental laws involved generalizations of greater scope, and hence the gradual process of axiomatizing theories so that specific laws were seen to be instances of more general and fundamental laws could be envisaged as, leading to grand unification in science (Nagel, 1961).

Mechanistic accounts of explanation provide a very different understanding of what producing and generalizing explanations involves. At the core of mechanistic accounts are models that represent the nature of the component parts and the operations they perform, plus how these are organized such that the phenomenon of interest can be understood as the overall mechanistic activity. Operative parts are spatially situated, and their spatial relation to other operative parts often figures critically in how they behave. While it is possible to describe such spatial layouts propositionally, the use of diagrams, maps and other visual displays often facilitates problem-solving and is easier to process. Operations involve component parts interacting with other parts; and while these too can be described propositionally, it is often easier to portray the changes, e.g., with video. Even a static diagram that uses arrows to indicate what part is acting on another often provides superior understanding. Accordingly, early cognitive theories were often presented in box-and-arrow diagrams, with boxes representing mental operations; arrows designating operations were sometimes overlaid on pictures of the brain to indicate both the neural components and their operations (Figure 2).

(a) (b)

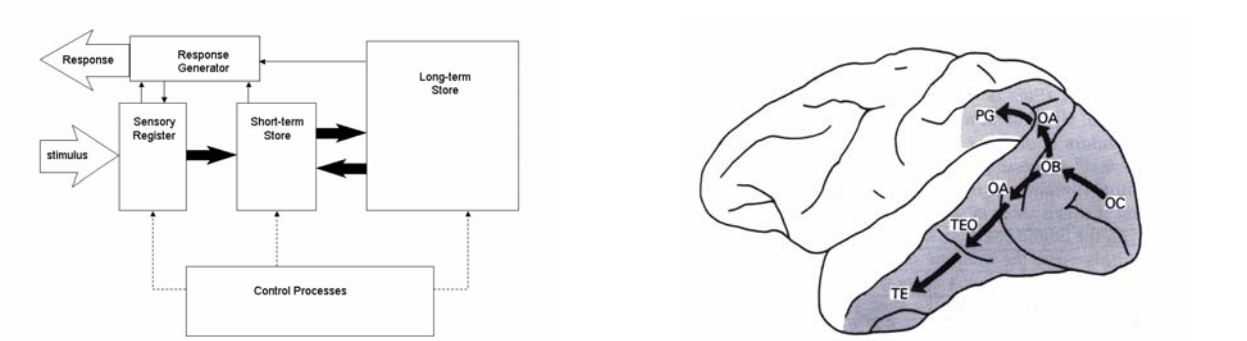

Figure 2. (a) Box and arrow diagram of the Atkinson and Shiffrin (1968) model of memory processes and (b) arrows overlaid on brain areas in Mishkin, Ungerleider, and Macko's (1983) identification of two visual systems in primate brains.

While there is little direct evidence about what occurs inside people as they reason about mechanisms, it seems likely that they often create visual or tactile models of mechanisms, which they then transform to simulate the mechanism's activity (see Waskan, 2006). By emphasizing the role of visual representation, we note that linguistic representations often function somewhat differently without thereby negating their contributions to explanation. Consider, e.g., figure captions in journal articles, which provide interpretations of aspects of the visual representation and help focus attention. Insofar as people understand how a mechanism works through

simulation (whether mentally or orthographically), it is the ability to model mechanisms that provides the dynamics of understanding—not traditional logic.

Whereas philosophers focused on nomological explanation had little to say about the discovery of laws, there is much to say about the discovery of mechanisms. Insofar as decomposing mechanisms into operative parts and organization is crucial to eventually understanding them, one can examine the ways in which scientists generate such decompositions. Experimental techniques play a crucial role in decomposition, and various techniques both provide important types of information (but also are potentially misleading). Scientists also face challenges in developing appropriate concepts for operative parts, and here they often reason by analogy from a domain where decomposition has been successful to a novel domain.

In the nomological tradition, generalization was straight-forward: the law generalized to all cases that satisfied its antecedent. Models of mechanisms, in contrast, are developed to account for specific cases chosen for study. In biology, researchers often employ model organisms or preparations, and work out detailed explanations of the mechanisms active in that organism or preparation. For example, the giant squid axon provided the basis for much of the research in electrophysiology that led to understanding the action potential of neurons. The giant squid axon was chosen not because it was typical, but because it was possible to conduct experiments on it given its size and the crudeness of the available electrodes. Other neurons will differ, and it is never clear in advance of conducting experiments on them how much the model of the mechanism will have to be revised to be applicable, much less whether a very different sort of account will be required. Cognitive psychologists, for the most part, limit their investigations to one species—namely, humans, and particularly college students—as their model system. They face questions about whether mechanistic models developed for this population will apply to others differing in age, education, cultural heritage, etc. So, generalization with mechanisms does not involve simply applying a law to other instances of the specified natural kind, but engaging in inquiry to determine what modifications are required to characterize the similar yet somewhat different mechanism active in another person or in different circumstances in the same person.

5. Mechanism as a unifying framework for understanding psychological explanation

There is perhaps no greater disciplinary crucible than psychology for working out the issues and problems of scientific explanation, if only because of the sheer range of explanatory practices and strategies that figure in research on psychological phenomena. We have indicated some examples of psychological explanation in which individual phenomena were putatively explained by showing them to be instances of general laws, as well as examples that involved representing target phenomena as mechanistic activities. In the previous section, we identified a number of contrasts between these two kinds of cases; in this final section, we conclude by showing that appeals to laws and appeals to mechanisms are quite compatible, and that the mechanistic framework allows for both. Consequently, it provides a unifying framework for answering the question, “What is psychological explanation?”

We have already alluded to one strategy for incorporating laws into the mechanistic framework; laws (*effects* in psychology) often provide precise descriptions of the phenomenon to be

explained. A great deal of research goes into establishing psychological phenomena—e.g., identifying variables and specifying as precisely as possible (ideally, mathematically) the relations between these variables that characterize the phenomena. But description and explanation can certainly diverge. Explanation requires developing a model of the mechanism in terms of parts, operations, and organization. Stuart Glennan (1996) proposed a second role for laws in mechanistic explanations: mechanistic explanations both explain lawlike regularities and appeal to other lawlike regularities to characterize the operations (he speaks of *interactions*) constituting the mechanistic activity. It is occasionally possible to describe the operations in mathematical detail, but operations in psychological explanations are rarely characterized in terms of laws. Yet, appropriately modified, Glennan's proposal may be acceptable; after what may be several iterations of decomposing a mechanism's component parts and operations, investigators may reach processes that fall under the scope of physical or chemical laws (e.g., Ohm's law and the Nernst equations in the case of the action potential of neurons).

This iterative nature of decomposition and its role in the refinement of mechanistic models deserves further elucidation. The discovery of operative parts often inspires further investigation into how those operations are performed. This practice of taking mechanisms apart into their components, and, in turn, the components into their components, is clearly reductionistic; yet, reduction in the context of mechanism has a rather different flavor than philosophical accounts of reduction emphasizing derivation of laws from more fundamental laws. Eventually, such accounts envisage deriving all laws of the special sciences from fundamental laws of physics. Such derivations can only go through, though, with the specification of boundary conditions, whose source is seldom identified. Given that lawful regularities are often explained by models of mechanisms, we can see that descriptions of boundary conditions provide a specification of the components and organization of the mechanism. At best, lower-level laws describe some operative parts—not their presence or configuration. Accordingly, lower-level laws fall far short of providing all the information needed to derive the higher-level regularities, which are better explained using mechanistic models. For example, the organization of components parts and operations, both spatially and temporarily, are crucial to a mechanism's activities, and this is not provided simply by lower-level laws or even knowledge of the component parts and operations themselves.

Additionally, a given mechanistic activity is always constrained by its environmental conditions. And because mechanisms are composite hierarchical systems with myriad mereological part/whole relations, their component parts operate at a lower level than—and are organized differently from—the level at which the mechanism as a whole produces the target phenomenon (e.g., as a final common pathway). Again, mechanisms are often themselves a component part in yet a higher-level mechanism, and the regularities resulting from the organization and situatedness of that higher-level mechanism constrain the activities of the initial component mechanism. Hence, the process of both decomposing *and* composing systemic structures and functions across various levels is a fundamental part of the mechanistic framework. Accordingly, while mechanistic explanations are in part reductionistic, they also accommodate the emergence of higher levels of organization and the need for autonomous inquiry into the regularities found amongst the denizens of these higher levels. So, as mechanists have consistently pointed out, the inherently reductionistic elements of mechanistic explanation need not threaten the explanatory

autonomy of higher-level psychological explanations—indeed, it depends on them to situate the mechanism in context (Wright & Bechtel, 2006; Wright, 2007).

By accommodating both a reductionist and an emergentist perspective, mechanistic explanation provides a unifying framework that integrates a variety of explanatory projects in psychology. Many psychological explanations focus on particular aspects of the behavior of whole agents. To explain these, psychologists try to identify the operations involved and increasingly localize these to brain regions where they are performed. In other subfields of psychology, e.g., social psychology, the focus is not just on the behavioral propensities of agents but also the social situations in which these are realized; and increasingly, investigators interested in cognitive capacities are also concern with the embodied and situated context of the agent performing these activities. As noted, environmental contexts often figure centrally in determining the activities of mental mechanisms, and therefore have a non-trivial role in being represented in the explanans of a mechanistic explanation. Turning downwards, researchers are increasingly discovering the importance of the chemical milieu in the brain to aspects of behavior and mood. To understand how, e.g., the chemistry of an individual's brain affects the person's performance on memory and problem-solving tasks, researchers need to understand how the particular chemistry affects the component operations (or the components of the components) of the responsible mechanism. Besides neurochemistry, other subfields of neuroscience are increasing understanding how component mechanisms operate in psychological agents. Findings in both neuroscience and the social sciences are highly pertinent to the development of psychological explanations. Nonetheless, for phenomena falling within the purview of psychology, the relevant mechanisms are those that scientific psychology investigates. Psychological explanation is integrated into the explanations offered in related sciences, but retains its own identity.

Having suggested how mechanistic accounts of explanation can incorporate the insights of their nomological counterparts, there remains the question of whether mechanistic explanation exhausts psychological explanation. Two endeavors in contemporary psychology suggest alternatives to mechanistic explanation or the need to expand the characterization of mechanistic explanation. The first involves the application of tools of dynamical systems theory to psychological phenomena (Thelen & Smith, 1994; Kelso, 1995; Port & van Gelder, 1995). Much in the spirit of nomological explanation, advocates of dynamical models appeal to equations relating variables that characterize the behavior of mental systems, and—at times—seem to deny the utility or even possibility of decomposing systems into component parts, operations, and organization (van Gelder, 1995; van Orden, Pennington, & Stone, 2001). Whether such explanations are a genuine alternative primarily turns on the nature of the variables in dynamical accounts. If these variables characterize operative parts, then dynamical models may be subsumed under mechanistic accounts of explanation, although they rightly refocus attention on the complex dynamics resulting from components that interact in complex ways over time. The second endeavor involves appeals to evolution—especially natural selection—to explain psychological phenomena. As with biological explanation, there appears to be a tension between providing mechanistic explanations of a current system (what Mayr termed *proximal explanation*) and an evolutionary explanation (*ultimate explanation*) (Mayr, 1982). The very terms *proximal* and *ultimate* imply some prioritization, but many in evolutionary biology have come to recognize the critical constraints that understanding specific mechanisms, especially mechanisms of development, places on processes such as natural selection. Natural selection is

often itself characterized as a mechanism of evolution, but there is currently active discussion as to whether explanation in terms of natural selection fits extant philosophical accounts of mechanisms (Skipper & Millstein, 2005) or requires a different account of explanation altogether.

References

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In K. W. Spence & J. T. Spence (Eds.), *The Psychology of Learning and Motivation: Advances in Research and Theory* (Vol. 2, pp. 89-195). New York: Academic.

Bechtel, W. (in press). *Mental mechanisms: Philosophical perspectives on the sciences of cognition and the brain*. Mahwah, NJ: Erlbaum.

Bechtel, W., & McCauley, R. N. (1999). Heuristic identity theory (or back to the future): The mind-body problem against the background of research strategies in cognitive neuroscience. In M. Hahn & S. C. Stoness (Eds.), *Proceedings of the 21st Annual Meeting of the Cognitive Science Society* (pp. 67-72). Mahwah, NJ: Lawrence Erlbaum Associates.

Bishop, M. P., Elder, S. T., & Heath, R. G. (1963). Intracranial self-stimulation in man. *Science, 140*, 394-396.

Broca, P. (1861). Remarques sur le siége de la faculté du langage articulé, suivies d'une observation d'aphemie (perte de la parole). *Bulletin de la Société Anatomique, 6*, 343-357.

Cummins, R. (2000). "How does it work?" versus "what are the laws?": Two conceptions of psychological explanation. In F. Keil & R. Wilson (Eds.), *Explanation and cognition* (pp. 117-144). Cambridge, MA: MIT Press.

Donders, F. C. (1868). Over de snelheid van psychische processen. Onderzoekingen gedaan in het Physiologisch Laboratorium der Utrechtsche Hoogeschool: 1868-1869. *Tweede Reeks, 2*, 92-120.

Dretske, F. I. (1981). *Knowledge and the flow of information*. Cambridge, MA: MIT Press/Bradford Books.

Fechner, G. T. (1860). *Elemente der Psychophysik*. Leipzig: Breitkopf und Härtel.

Heath, R. G. (1963). Electrical self-stimulation of the brain in man. *American Journal of Psychiatry, 120*, 571-577.

Holland, J. H., Holyoak, K. J., Nisbett, R. E., & Thagard, P. R. (1986). *Induction: Processes of inference, learning and discovery*. Cambridge, MA: MIT.

Kelso, J. A. S. (1995). *Dynamic patterns: The self organization of brain and behavior*. Cambridge, MA: MIT Press.

Langley, P., Simon, H. A., Bradshaw, G. L., & Zytkow, J. M. (1987). *Scientific discovery: Computational explorations of the creative process*. Cambridge: MIT Press.

Mayr, E. (1982). *The Growth of Biological Thought*: Harvard.

Miller, G. A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review, 63*, 81-97.

Millikan, R. G. (1984). *Language, thought, and other biological categories*. Cambridge, MA: MIT Press.

Mishkin, M., Ungerleider, L. G., & Macko, K. A. (1983). Object vision and spatial vision: Two cortical pathways. *Trends in Neurosciences, 6*, 414-417

Nagel, E. (1961). *The structure of science*. New York: Harcourt, Brace.

Newell, A. (1980). Physical symbol systems. *Cognitive Science, 4*, 135-183.

Olds, J. (1955). A physiological studyof reward. In D. McClelland (Ed.), *Studies of motivation* (pp. 134-143). New York: Appleton.

Olds, J., & Milner, P. (1954). Positive reinforcement produced by electrical stimulation of septal area and other regions of rat brain. *Journal of Comparative and Physiological Psychology, 47*, 419-429.

Port, R., & van Gelder, T. (1995). *It's about time*. Cambridge, MA: MIT Press.

Raichle, M. E., & Minton, M. A. (2006). Brain work and brain imaging. *Annual Review of Neuroscience, 29*, 449-476.

Roediger, H. L., Buckner, R. L., & McDermott, K. B. (1999). Components of processing. In J. K. Foster & M. Jelicic (Eds.), *Memory: Systems, process, or function* (pp. 32-65). Oxford: Oxford University Press.

Schacter, D. L., & Tulving, E. (1994). What are the memory systems of 1994? In D. L. Schacter & E. Tulving (Eds.), *Memory systems 1994* (pp. 1-38). Cambridge, MA: MIT Press.

Scoville, W. B., & Milner, B. (1957). Loss of recent memory after bilateral hippocampal lesions. *Journal of Neurology, Neurosurgery, and Psychiatry, 20*, 11-21.

Skipper, R. A., & Millstein, R. L. (2005). Thinking about evolutionary mechanisms: Natural selection. *Studies in History and Philosophy of Biological and Biomedical Sciences, 36*, 327-347.

Stevens, S. S. (1957). On the psychophysical law. *Psychological Review, 64*, 153-181.

Teigen, K. H. (2002). One hundred years of laws in psychology. *American Journal of Psychology, 115*, 103-118.

Thagard, P. (1988). *Computational philosophy of science*. Cambridge, MA: MIT Press/Bradford Books.

Thelen, E., & Smith, L. (1994). *A dynamical systems approach to the development of cognition and action*. Cambridge, MA: MIT Press.

Tulving, E. (1983). *Elements of episodic memory*. New York: Oxford University Press.

Valentstein, E. S., & Beer, B. (1964). Continuous opportunity for reinforcing brain stimulation. *Journal of the Experimental Analysis of Behavior, 7*, 183-184.

Valentstein, E. S., & Campbell, J. F. (1966). Medial forebrain bundle-lateral hypothalamic area and reinforcing brain stimulation. *American Journal of Physiology, 210*, 270-274.

van Essen, D. C., & Gallant, J. L. (1994). Neural mechanisms of form and motion processing in the primate visual system. *Neuron, 13*, 1-10.

van Gelder, T. (1995). What might cognition be, if not computation. *The Journal of Philosophy, 92*, 345-381.

van Orden, G. C., Pennington, B. F., & Stone, G. O. (2001). What do double dissociations prove? Inductive methods and isolable systems. *Cognitive Science, 25*, 111-172.

Waskan, J. (2006). *Models and cognition*. Cambridge, MA: MIT Press.

Weber, E. H. (1834). *De pulsu, resorpitione, auditu et tactu: Annotationes anatomicae et physiologicae*. Leipzig: Koehlor.

Weber, M. (2005). *Philosophy of experimental biology*. Cambridge: Cambridge University Press.

Wright, C. D. (2007). Is psychological explanation going extinct? In M. Schouten & H. Looren de Jong (Eds.), *The matter of the mind: Philosophical essays on psychology, neuroscience, and reduction*. Oxford: Blackwell.

Wright, C. D., & Bechtel, W. (2006). Mechanisms and psychological explanation. In P. Thagard (Ed.), *Philosophy of psychology and cognitive science* (pp. 31-77): Elsevier.