CellPress

Neuron
Perspective

Neuroscience Needs Behavior: Correcting a Reductionist Bias

John W. Krakauer,[1,*] Asif A. Ghazanfar,[2] Alex Gomez-Marin,[3] Malcolm A. MacIver,[4] and David Poeppel[5,6]
[1]Departments of Neurology, and Neuroscience, Johns Hopkins University, Baltimore, MD 21287, USA
[2]Princeton Neuroscience Institute, Departments of Psychology and Ecology & Evolutionary Biology, Princeton University, Princeton, NJ 08540 USA
[3]Instituto de Neurociencias, Consejo Superior de Investigaciones Científicas & Universidad Miguel Hernández, Sant Joan d'Alacant, 03550 Alicante, Spain
[4]Neuroscience and Robotics Laboratory, Department of Neurobiology, Department of Mechanical Engineering, Northwestern University, Evanston, IL 60208, USA
[5]Department of Psychology, New York University, New York, NY 10003, USA
[6]Neuroscience Department, Max-Planck Institute for Empirical Aesthetics, 60322 Frankfurt, Germany
*Correspondence: jkrakau1@jhmi.edu
http://dx.doi.org/10.1016/j.neuron.2016.12.041

There are ever more compelling tools available for neuroscience research, ranging from selective genetic targeting to optogenetic circuit control to mapping whole connectomes. These approaches are coupled with a deep-seated, often tacit, belief in the reductionist program for understanding the link between the brain and behavior. The aim of this program is causal explanation through neural manipulations that allow testing of necessity and sufficiency claims. We argue, however, that another equally important approach seeks an alternative form of understanding through careful theoretical and experimental decomposition of behavior. Specifically, the detailed analysis of tasks and of the behavior they elicit is best suited for discovering component processes and their underlying algorithms. In most cases, we argue that study of the neural implementation of behavior is best investigated *after* such behavioral work. Thus, we advocate a more pluralistic notion of neuroscience when it comes to the brain-behavior relationship: behavioral work provides understanding, whereas neural interventions test causality.

Introduction

Advances in technology have allowed the study of neurons, including their component parts and molecular machinery, to an unprecedented degree. This work promises to yield much new information about brain structure and physiology independent of behavior—for example, the biophysics of receptors or details of spatial summation in dendrites. In addition, new methods such as optogenetics allow some causal relationships between brain and behavior to be determined. Here we will argue, however, that detailed examination of brain parts or their selective perturbation is not sufficient to understand how the brain generates behavior (Figure 1). One reason is that we have no prior knowledge of what the *relevant* level of brain organization is for any given behavior (Figure 1A). When this concern is coupled with the brain's deep degeneracy, it becomes apparent that the causal manipulation approach is not sufficient for gaining a full understanding of the brain's role in behavior (Marom et al., 2009). The same behavior may result from alternative circuit configurations (Marder and Goaillard, 2006), from different circuits altogether or the same circuit may generate different behaviors (Katz, 2016) (Figures 1D and 1E). This concern has been voiced before in a variety of ways (Anderson, 1972; Marr, 1982/2010; Oatley, 1978), but we think that it is useful to revisit and reframe the arguments at a time of understandable excitement about ever more effective interventionist approaches in neuroscience.

An analogy from computer science that has both historical and conceptual appeal is the distinction between software and hardware; whereby the software represents "what" the brain (or one of its modules) is doing, and the hardware represents "how" it is doing it. Sternberg has stated it as the "distinction between processors and the processes that they implement" (Sternberg, 2011, p. 158). The core question we address here is whether the processes governing behavior are best inferred from examination of the processors. In a nice irony, the computer science analogy has come full circle with a provocative study that applied numerous neuroscience techniques to a single microprocessor (analogous to a brain) in an attempt to understand how it controls three classic videogames (analogous to behaviors) (Jonas and Kording, 2017). Crucial to the experiment was that the answer was known a priori: the processor's operations can be drawn as an algorithmic flow chart. The sobering result was that performing interventionist neuroscience on the processor could not explain how the processor worked. We have more to say about this study later.

Neuroscience is replete with cases that illustrate the fundamental epistemological difficulty of deriving processes from processors. For example, in the case of the roundworm (*Caenorhabditis elegans*), we know the genome, the cell types, and the connectome—every cell and its connections (Bargmann, 1998; White et al., 1986). Despite this wealth of knowledge, our understanding of how all this structure maps onto the worm's behavior remains frustratingly incomplete. Thus, it is

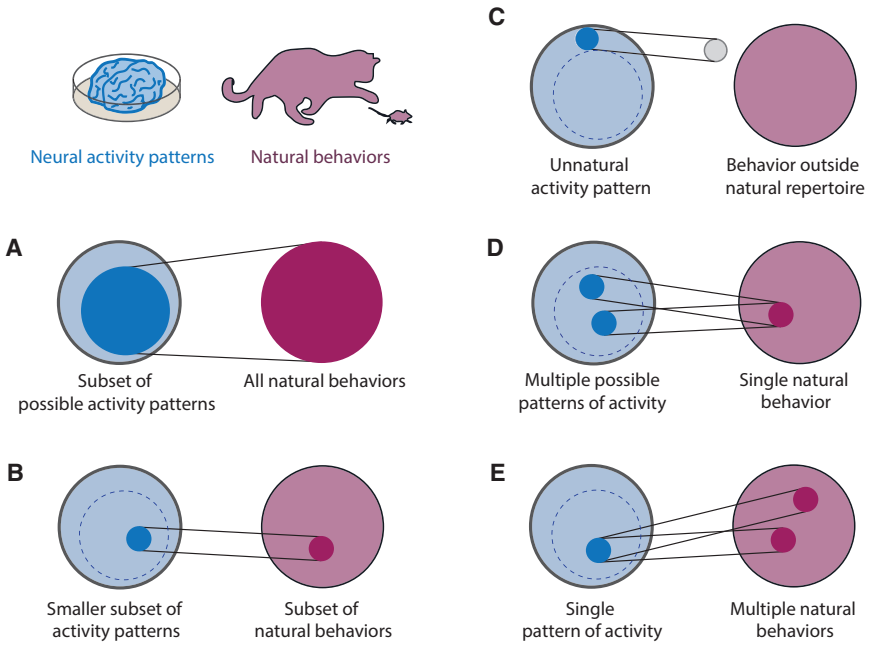


Figure 1. The Multiple Potential Mappings between Neural Activity Patterns and Natural Behaviors

(A) Of all the possible activity patterns of a brain in a dish (big pale blue circle), only a subset of these (medium dark blue circle) will be relevant in behaving animals in their natural environment (big magenta circle).

(B) Designing behavioral tasks that are ecologically valid (small magenta circle) ensures discovery of neural circuits relevant to the naturalistic behavior (small blue circle). Tasks that elicit species-typical behaviors with species-typical signals are examples (see Box 1).

(C) In order to study animal behavior in the lab, the task studied (small white circle) might be so non-ecological it elicits neural responses (small blue circle) that are never used in natural behaviors.

(D) Multiple Realizability: different patterns of activity or circuit configurations (small blue circles) can lead to the same behavior (small magenta circle).

(E) The same neural activity pattern (small blue circle) can be used in two different behaviors (two magenta circles). The circle with dashed perimeter in (B)–(E) is the subset of all possible neural activity patterns that map onto natural behaviors (from A).

readily apparent that it is very hard to infer the mapping between the behavior of a system and its lower-level properties by only looking at the lower-level properties (Badre et al., 2015; Carandini, 2012; Cooper and Peebles, 2015; Gomez-Marin et al., 2014). When we ask, "How does the brain generate behavior," we are primarily asking about how putative processing modules are organized so that they combine to generate behavior in a particular task environment. Relying solely on the collection of neural data, with behavior incorporated as an after-thought (and typically over-constrained; Box 1; Figure 1C), will not lead to meaningful answers. This is a question best answered through precise hypotheses articulated in an a priori conceptual framework, careful task design, and the collection of behavioral data.

How Did We Get Here?

New technologies have enabled the acquisition of massive and intricate datasets, and the means to analyze them have become concomitantly more complex. This in turn has led to a need for experts in computation and data analysis, with a reduced emphasis on organismal-level thinkers who develop detailed functional analyses of behavior, its developmental trajectory, and its evolutionary basis. Deep and thorny questions like "what would even count as an *explanation* in this context," "what is a *mechanism* for the behavior we are trying to understand," and "what does it mean to *understand* the brain" get sidelined. The emphasis in neuroscience has transitioned from these larger scope questions to the development of technologies, model systems, and the approaches needed to analyze the deluge of data they produce. Technique-driven neuroscience could be considered an example of what is known as the substitution bias: "[…] when faced with a difficult question, we often answer an easier one instead, usually without noticing the substitution" (Kahneman, 2011, p. 12).

In an interesting historical parallel to the argument we make here, the historian of science Lily Kay described how the discipline of molecular biology also arose from placing a premium on technology and its application to simple model systems (Kay, 1996). She quotes with concern Monod's line, "What is true for the bacterium is true for the elephant" (Kay, 1996, p. 5). Here we caution similarly against the idea that what is true for the circuit is true for the behavior. Monod's line has echoed through to the present day with the argument that molecular biology and its techniques should serve as the model for understanding in neuroscience (Bickle, 2016). We disagree with this totalizing reductionist view but take it as evidence that excessive faith in molecular and cellular biology may be partially to blame for the current dominance of interventionist explanations in neuroscience. We fully acknowledge the crucial role that technology plays in advancing biological knowledge and the value of interventionist approaches, but this tool-driven trend is not sufficient for understanding the brain-behavior relationship. Neural data obtained from new methods cannot substitute for developing new conceptual frameworks that provide the mapping between such neural data and behavior in an algorithmic sense (and not just a correlative or even causal way). Accomplishing this task requires hypotheses and theories based on careful dissection of behavior into its component parts or subroutines (Cooper and Peebles, 2015). The behavioral work needs to be as fine-grained as work at the neural level. Otherwise one is imperiled by a granularity mismatch between levels that prevents substantive alignment between different levels of description (Poeppel and Embick, 2005).

The first step for developing conceptual frameworks that meaningfully relate neural circuits to behavioral predictions is to design hypothesis-based behavioral experiments. Despite this pure behavioral step being of critical importance and highly informative in its own right, it has increasingly been marginalized

Box 1. Behavior

Tinbergen defined behavior as "The total movements made by the intact animal" (Tinbergen, 1955). Authors of a recent survey designed to investigate how working scientists define behavior came up with the following attempt at an updated definition, "Behavior is the internally coordinated responses (actions or inactions) of whole living organisms (individuals or groups) to internal and/or external stimuli, excluding responses more easily understood as developmental changes" (Levitis et al., 2009). The core of this definition is that behavior is the "internally coordinated responses … to internal and/or external stimuli." Clearly, however, some stimuli are more important than others in furthering our understanding of animals and their nervous systems.

Unfortunately, animals do not raise a checkered flag to indicate when they are about to perform a behavior, or signal when it ends. They are constantly in motion and responding to whatever is around them, which invites the following easy mistake: since an animal is responding to stimuli, and physiological correlates are measurable, one is therefore studying an animal's behavior (see the "overly constrained behavior" of Figure 1C). If, following the lead of 20th century ethologists, we treat behavior as no less an evolved entity than is, say, the shape of the humerus (Tinbergen, 1963), then correctly labeling something as behavior is contingent on the outcome of an investigation into what the animal does to ensure its survival in its native habitat. In this way it would be discovered, for example, that bats navigate through dense forests in total darkness while hunting insects and that rodents beat a hasty retreat when they see a hawk diving towards them, while not responding to similarly sized birds flying straight over. It is therefore a significant confusion to label a coordinated response to a stimulus a "behavior" without first determining the relevance of the response to the animal's natural life (Tinbergen's second question, Figure 4). Not doing so is to conflate a "stimulus response" with a spatio-temporal pattern that is the product of selection over time. While any perturbation applied to an animal can lead to productive lines of inquiry, whether or not it is founded in anything ethological, the history of many of the most productive moments in neuroscience is a history of having ingeniously abstracted an animal's *Umwelt* (Von Uexküll, 1992) in such a way as to admit of controlled, repeatable experiments.

In light of the preceding, placing a behaving animal in a situation where it perceives sensory events that are behaviorally relevant, and can act on them in approximately the same way as if they were embedded in the world, can be enormously useful (Figure 1B). For example, engaging a subject with the real or simulated presence of another can capture behavioral principles that are common across species. Doing so with marmoset monkeys revealed vocal turn-taking behavior with similar patterns of phase-locking and entrainment as in human communication (Takahashi et al., 2013). Eschewing the purely big data approach where behavioral data are acquired blindly from large numbers of animals through automation and without regard for the individual (Anderson and Perona, 2014), this organismal-level study led to insights into both developmental (Takahashi et al., 2015) and evolutionary (Borjon and Ghazanfar, 2014) processes, and, subsequently, to computational principles shared across species (Takahashi et al., 2012). Similar ethological approaches in other species have led to a number of behaviorally driven investigations of neural level mechanisms: fish (Bass and Chagnaud, 2012), frogs (Leininger and Kelley, 2015), and birds (Benichov et al., 2016).

Ultimately, the most effective approach may be to simulate the entire natural task environment in order to elicit the full range of adaptive behavioral possibilities. Virtual reality (VR) systems developed for a host of model systems, including rodents, flies, and fish, offer such an approach (Dombeck and Reiser, 2012). VR systems were originally an answer to the problem of how to mediate the uneasy marriage of the tightly controlled but non-ethological world of laboratory physiology, and the poorly controlled but ethologically relevant world of the behaviors studied by ethologists (MacIver, 2009). It is critically important to realize, however, that effective use of VR requires a fine-grained quantitative understanding of the behavior under study as it occurs in the unimpeded animal. Only then can the investigator assess, for example, whether the VR system is in fact able to trick the animal into believing it is in the world as it would normally operate within. For example, a VR system has been developed by which we can reliably elicit prey capture behavior in larval zebrafish (which hunt *Paramecia*) (Bianco et al., 2011). Careful behavioral work on the free-swimming animal showed that prey detection is marked by the fish's eyes verging together to point to the prey (Bianco et al., 2011; Patterson et al., 2013). This eye movement is not seen during any other behaviors. With that knowledge in hand, we can assess the success of a VR system by how often we can see the eyes move in this manner when we display artificial *Paramecia* on a small screen in front of the animal, which is otherwise fixed in place by being embedded in a block of agar.

Understanding behavior and its component processes at the level of detail necessary to generate meaningful neural level insights will require an emphasis on natural behaviors performed by individuals. Although there are new technologies enabling the blind acquisition of massive behavioral datasets and the application of machine learning algorithms (Anderson and Perona, 2014), they will not lead to the detailed functional analyses of ethological behavior, its developmental trajectory, and its evolutionary basis that are necessary for appropriately constrained implementation-level theories.

or at best postponed (Anderson and Perona, 2014). It is disturbingly common for studies to include behavior as simply a hasty add-on in papers that are otherwise crammed full of multiple techniques, types of results, and even species. It is as if every paper needs to be a methodological decathlon in order to be considered important. Behavior must be seen as something that can stand alone as a foundational phenomenon in its own right (Box 1).

Why We Still Need Behavior

Perhaps the best-known example of a framework devised to formalize what it means to understand the link between brain

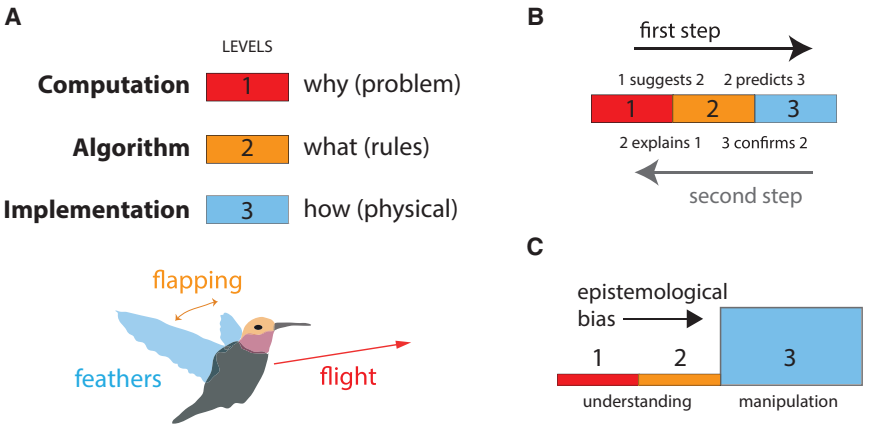


Figure 2. Marr's Three Levels of Analysis
(A) A bird attempts to fly (goal) by flapping its wings (algorithmic realization) whose aerodynamics depend on the features of its feathers (physical implementation). Feathers "have something to do" with flight and flapping, but what level of understanding do we achieve if we dissect the properties of the feathers alone? Bats fly but don't have feathers, and birds can fly without flapping.
(B) The relationship between the three levels is not arbitrary; step 1 comes before step 2: the algorithmic level of understanding is essential to interpret its mechanistic implementation. Step 2: implementation level work feeds back to inform the algorithmic level.
(C) An epistemological bias toward manipulation-based view of understanding induced by technology (black filled arrow).

and behavior is David Marr's levels of understanding of complex systems, originally fashioned as a critique of reductionist work in neurophysiology (Marr, 1982/2010). We do not intend to rehash or dissect Marr's particular formulation, as this has been done many times before, and sometimes arguments over the definitions of his three levels—or whether there should more than three—obscure the central and deep point made by Marr (and many others before and after [Anderson, 1972; Bechtel, 2008; Carandini, 2012; Cooper and Peebles, 2015; Oatley, 1978; Tinbergen, 1963; Badre et al., 2015; Fetsch, 2016; Frank and Badre, 2015; Schall, 2004]): understanding something is not the same as just describing it or knowing how to intervene to change it. To most it is not news that description is not understanding, but too often in neuroscience causal efficacy is taken as equal to understanding.

Marr took a strong position on the inadequacy of a strictly neurophysiological approach to understanding: "…trying to understand perception by understanding neurons is like trying to understand a bird's flight by studying only feathers. It just cannot be done" (Marr, 1982/2010) (Figure 2). Marr's main intuition was that it is much more difficult to infer from the neural hardware (or *implementation*; level 3) what *algorithm* (level 2) the nervous system is employing as compared to getting to it via an analysis of the *computational* problem (level 1) it is trying to solve. Marr's main objection to trying to understand the brain by recording from neurons was that this only leads to descriptions rather than explanations. A description of neural activity and connections is not synonymous with knowing what they are doing to cause behavior. Even strong believers in the work done at the level of neurons and molecules (the implementation level) concede Marr's point (Bickle, 2015). An analogy that helps get this point across is understanding of the game of chess. Understanding the game does not depend on knowing anything about the material out of which the board or chess pieces are made. Indeed, Marr suggested that the details of the nervous system may not even matter. While it is true that the physical properties of the chess pieces can impinge on application of the rules—for example, if one inadvertently gives a child a chess set for which all the pieces are too heavy for her to pick up. The "therapeutic" solution is lighter chess pieces, but this in no way has changed the analysis or understanding of chess. This chess analogy serves to make an important point: well-designed behavioral experiments in the absence of work at the neural level can be highly informative on their own. Behavioral experiments often are a necessary first step *before* a subsequent mutually beneficial knowledge loop is set up between implementation and behavioral level work.

A more concrete example of the problems that arise if neural data are used to infer a psychological process comes from the debates regarding the behavioral relevance of "mirror neurons." Mirror neurons, first discovered in the premotor cortex of monkeys, fire whether the monkey itself performs a particular motor goal or observes another individual doing so (di Pellegrino et al., 1992). A huge number of variants of these experiments have been done in both humans and monkeys, but they all have the same general approach: show a common neuronal firing (or fMRI or EEG/MEG activation pattern) when a goal is achieved either in the first person or observed in the third person. Interpretation then has the following logic: as neurons can be decoded for intention in the first person, and these same neurons decoded for the same intention in the third person, then activation of the mirror neurons can be interpreted as meaning that the primate has understood the intention of the primate it is watching. The problem with this attempt to posit an algorithm for "understanding" based on neuronal responses is that no independent behavioral experiment is done to show evidence that any kind of *understanding* is actually occurring, understanding that could then be correlated with the mirror neurons. This is a key error in our view: behavior is used to drive neuronal activity but no either/or behavioral hypothesis is being tested per se. Thus, an interpretation is being mistaken for a result; namely, that the mirror neurons understand the other individual. Additional behavioral evidence that the participant understands the other individual is lacking. This tendency to ascribe psychological properties to single neuron activity that can only be sensibly ascribed to a whole behaving organism is known as the mereological fallacy—a fallacy that we neuroscientists continue to fall for even though we've known about it since Aristotle's *De Anima* (Smit and Hacker, 2014). Thus, what is needed is a better a priori testable framework for behavioral-level understanding that can lead to more thoughtfully designed neurophysiological experiments.

Indeed, to the degree that action understanding has been examined in patients, the evidence does not support a role for the putative mirror neuron mechanism (Hickok, 2009).

A recent review exemplifies the neuroscience zeitgeist by stating that the time has come to go from considering individual neurons as the functional units of the nervous system to ensembles of neurons (circuits, networks) (Yuste, 2015). The main argument is that ensembles generate states that would never be appreciated by recording one neuron at a time. The claim is then made that, with the application of new technologies (e.g., two-photon imaging, multielectrode recordings, etc.), identification of these neural states will help us better understand the link between the brain and behavior (although again behavior is at best given backseat status) (Yuste, 2015). No overt new theory, however, is offered that will bridge ensemble activity and behavior. It is therefore unclear how fundamentally different it really is, conceptually, to move from "neuron" to "neurons." Modeling and studying the responses of the neural substrate on any scale—large or small—will not *by itself* lead to insights about how behavior is generated. One reason for this is that the properties of neural tissue may be more diverse than the subset actually exploited for natural behaviors (Figures 1A and 1D).

Without well-characterized behavior and theories that can act as a constraint on circuit-level inferences, brains and behavior will be like two ships passing in the night. The field has been here before. Concerns with the complete-circuit-description approach were already recognized almost 40 years ago with the publication—along with extensive accompanying commentary—of an article titled "*Are Central Pattern Generators Understandable?*" (Selverston, 1980). A plea, similar to those we hear today, was made for ever more detailed characterization of each element in the circuit and specification of the synaptic connectivity between these elements. Many of the commentaries pointed out, however, that increasingly complete descriptions at one level do not serve as a bridge to the next level. For example, Sten Grillner wrote that a central pattern generator may be understood better using the *intentional stance*, borrowing from the philosopher Daniel Dennett (Dennett, 1989), which is the view that when an entity is designed for a purpose it is therefore subject to rational rules that can be determined by studying its behavior without necessarily having to analyze all its physical parts (comment by Grillner in Selverston, 1980).

The phenomenon at issue here, when making a case for recording from populations of neurons or characterizing whole networks, is *emergence*—neurons in their aggregate organization cause effects that are not apparent in any single neuron. Following this logic, however, leads to the conclusion that behavior itself is emergent from aggregated neural circuits and therefore should also be studied in its own right. An example of an emergent behavior that can only be understood at the algorithmic level, which in turn can only be determined by studying the emergent behavior itself, is flocking in birds. First one has to observe the behavior and then one can begin to test simple rules that will lead to reproduction of the behavior, in this case best done through simulation. The rules are simple—for example, one of them is "steer to average heading of neighbors" (Reynolds, 1987). Clearly, observing or dissecting an individual bird, or even several birds could never derive such a rule. Substitute flocking with a behavior like reaching, and birds for neurons, and it becomes clear how adopting an overly reductionist approach can hinder understanding.

How has neuroscience dealt with this persistent gap between explanation and description? It has opted to favor interventionist causal versions of explanation. Unfortunately, there are no shortcuts in the trajectory from psychology, cognition, perception, and behavior to neurons and circuits. One might argue that techniques now exist that make it possible to manipulate neural circuits directly, for example, with optogenetics or transcranial magnetic stimulation, so that causal relations—and not just correlations—can be discovered (Bickle, 2015). The critical point, however, is that causal-mechanistic explanations are qualitatively different from understanding how component modules perform the computations that then combine to produce behavior.

The distinction between causal claims and understanding via algorithmic or computational processes should be apparent from argument alone. That said, the recent study by Jonas and Kording (Jonas and Kording, 2017) we referred to earlier provides an empirical demonstration of the fundamental difference between intervening and recording versus understanding how information flows through processing steps. The study poses the question of whether a neuroscientist could understand a microprocessor. They applied numerous neuroscience techniques to a high-fidelity simulation of a classic video game microprocessor (the "brain") in an attempt to understand how it controls the initiation of three well-known videogames (which they dubbed as "behaviors") originally programmed to run on that microprocessor. Crucial to the experiment was the fact that it was performed on an object that is already fully understood: the fundamental fetch-decode-execute structure of a microprocessor can be drawn in a diagram. Understanding the chip using neuroscientific techniques would therefore mean being able to discover this diagram. In the study, (simulated) transistors were lesioned, their tuning determined, local field potentials recorded, and dimensionality reduction performed on activity across all the transistors. The result was that none of these techniques came close to reverse engineering the standard stored-program computer architecture (Jonas and Kording, 2017).

A number of noteworthy points emerge from this study that should be further highlighted. The treatment of "behavior" perfectly represents how the neuroscience field typically tends to work with this concept. The behavioral data analyzed consisted of 10 s of spontaneous activity with no player actually playing the game (Jonas and Kording, 2017). This is a fragment of activity, which we refer to as "stimulus-response" (Box 1) to distinguish it from behavior, an adaptive pattern of activity (i.e., one that enhances fitness). As such, this activity is a starting point that is unlikely to result in understanding no matter how advanced the subsequent analysis. But let's suppose that instead of a fragment, we have a complete activity map for an entire game played by a person. Let's further suppose that a much better job, using better analysis algorithms, could be done with the data. We suggest this would similarly lead to no meaningful insights into the processor's functional architecture, since again no behavioral-level hypothesis is being tested; there

is no conceptual structure in place. Which of an infinite set of potentially interesting patterns in the data should be selected for further investigation (see Figure 1C, overly constrained behavior)? The best way to answer this would be to examine the game-play—the behavior—itself. An engineer trying to make a copy of the machine would generate a high-level task analysis of what the microprocessor needs to do in order to play its part of the game. For example, she might study the on-screen positions, shapes, and colors of agents over time, the value of point scores, and the responses to joystick input generated by the player. Then she might ask how is the chip in the machine fulfilling these higher-order needs of game-play. From that starting point, more productive work on the role of specific portions of the chip would be possible.

Thus, two questions need to be asked. First, is causal-mechanistic explanation at the neural level better in principle than algorithmic or computational accounts of behavior? Second, is causal-mechanistic explanation sufficient to explain a behavioral phenomenon? Clearly knowing the sufficient and necessary conditions to evoke a behavior as found through an optogenetic or similar manipulation can fall far short of knowing the rules needed to instruct a robot or computer to perform the activity in question. Thus, we would argue that if the question is "how does the brain lead to behavior?" we are first asking why is the brain performing this behavior and then asking how is it doing it. So using the flying analogy (Figure 2), once we agree that bird flight is an adaptive behavior, we then determine that it flies by flapping it wings and not by wiggling its feet. Once we have worked this out, we can start studying the feathers that make up the wing. Seen this way, understanding that the flapping of wings is critical to flight aids the subsequent study of feathers. It is unlikely that, from the outset, studying an ostrich feather in isolation would lead to the conclusion that there is such a phenomenon as flight or even that feather-like structures would be useful for flight.

Why Higher-Level Concepts Are Needed to Understand Neuronal Results: The Nature of "Mechanism"

Why is it the case that explanations of experiments at the neural level are dependent on higher-level vocabulary and concepts? The answer is that this dependency is intrinsic to the very concept of "mechanism." A mechanism can be defined as "a structure performing a function in virtue of its component parts, component operations, and their organization. The orchestrated functioning of the mechanism is responsible for one or more phenomena" (Bechtel, 2008, p. 13). Crucially, the components of a mechanism do different things than the mechanism organized as a whole (i.e., emergence) (Bechtel, 2008). A reductionist treatment of the components must be combined with investigation of how the total mechanism is organized and how it behaves when embedded in an environment; an approach that unavoidably spans two levels (Bechtel, 2008) (Box 1). Even the reductionist idea of causality needs to be qualified. An idea related to emergence is that of "downward causation." Take, for example, the cardiac rhythm—a behavior that is the net consequence of the interplay between a cell's membrane and the ion channels in it (Noble, 2012). The conceptual point is that the ion channels do not *cause* the cardiac rhythm—instead the rhythm just *is* the combination of the higher level of the cell membrane and the lower level of ion channels. So even when causality claims are sought they often only make sense when all levels are considered together simultaneously rather than seeing the higher level as subordinate or collapsible to the lower level. Ion channels do not beat, heart cells do. Neural circuits do not feel pain, whole organisms do.

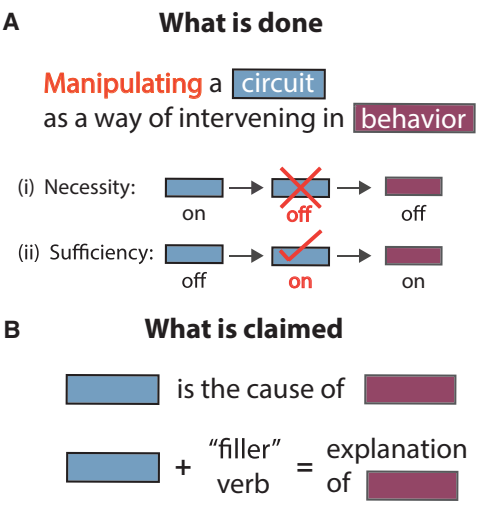
A potential objection to this might be to say, "Who cares what philosophers say about the differences between psychology and neuroscience, or reductionism in general? We are scientists, not philosophers!" The answer to this is simple: there is no escape from philosophy. Every scientist takes a philosophical position, either tacitly or explicitly, whenever they state that a result is "important," "fundamental," or "interesting." This is because such assertions are always a judgment from outside of science. There is no "interesting" variable inherent to the data that can be objectively plotted on a graph—abstract reasoning and normative claims cannot be substituted by, or obtained from, data. Tacit awareness that causal manipulative work, which tests necessity and sufficiency claims, is not the same as understanding is apparent with common sentences like, "The circuit X is *involved* in behavior Y." This is, however, just a restatement of the correlation or causal relation and does no extra explanatory work. The italicized word is known as a *filler term*, which indicates the lack of an explicit conceptual framework for the mapping between circuit and behavior (Craver, 2008) and so just fills in for it (Figure 3). Importantly, however, the use of filler terms signals a tacit awareness of the lack of, and a desire for, a different kind of understanding, which we argue can be obtained by doing empirical and theoretical behavioral work. This work would complement causal explanations at the neural and circuit level. A nice statement of this dual view of understanding in neuroscience was made by the cognitive scientist Longuet-Higgins, "In so far as the neurophysiologist is concerned to understand how the brain works, he must equip himself with a non-physiological account of the tasks which the brain and its peripheral organs are able to perform; only then can he form mature hypotheses as to how these tasks are carried out by the available 'hardware'—to borrow a phrase from computing science" (Longuet-Higgins, 1972, p. 256).

Behaviorally Driven Neuroscience Yields More Complete Insights

Here we describe four examples of what we take to be the essential interplay between computational theories and algorithmic formulations of behaviors, on the one hand, and their neural implementation, on the other.

Bradykinesia

Bradykinesia is one of the cardinal symptoms of Parkinson's disease, which is manifest as a lack of movement vigor. Bradykinesia is causally related to dopamine depletion in the substantia nigra and is partially and transiently reversed by increasing dopamine with medication. Do these facts truly help us understand why dopamine depletion leads to bradykinesia? The answer is clearly no—loss of a neurotransmitter may explain the necessary and sufficient *starting* conditions for bradykinesia, and indeed the investigation of these starting conditions—dopaminergic cell



C Most used "filler" verbs

reflects	encodes
reveals	induces
involves	enables
regulates	ensures
mediates	supports
generates	promotes
modulates	determines
shapes	plays a role in
underlies	contributes to
produces	is associated with

Figure 3. The Interventionist Type of Understanding Is Not Sufficient

The current dominant notion of what constitutes understanding in neuroscience is based on an interventionist approach to causality.

(A) Intervening at the neural level (blue) as a way to explain the behavior (magenta) through necessity and sufficiency claims.

(B) The result that "X is necessary and sufficient for Y to occur" allows a causal claim. An additional explanatory sentence is often added but is merely the same causal result rearticulated with a "filler" verb.

(C) Common filler verbs.

death—is a major area of ongoing work at the cellular level. Nevertheless, work restricted to this level cannot explain why dopamine depletion causes a loss of vigor. This is also true when neural correlates are found. In contrast, psychophysical studies in patients with Parkinson disease suggested that low vigor in these patients is caused by a loss of implicit motivation for moving fast because of a skewed cost function containing effort and accuracy terms (Mazzoni et al., 2007). These human psychophysical experiments then led to analogously designed experiments in mice that demonstrated that there are cells in the dorsal striatum whose activity correlates with movement vigor, and that suppression of these cells optogenetically reduces vigor (Panigrahi et al., 2015). Note again that these mouse experiments very much adhere to the causal interventionist approach to explanation. The complete interpretation of these experiments, however, *requires* the explanatory framework provided by the initial human behavioral work.

Sound Localization

The study of sound localization in avian and mammalian brains provides a nice example of the value of the behaviorally driven approach we are advocating. In the dark, both avian (e.g., barn owls) and mammalian (e.g., gerbils) brains must localize sound sources in the horizontal plane (that is, left or right with respect to their own body plan). The behavioral problem is thus well specified: namely, to localize a sound source based only on auditory cues.

One way to solve this problem is with inter-aural time difference cues. A way to calculate sound source location on the basis of such time-difference cues in a principled way was proposed by Jeffress (Jeffress, 1948). He showed that the computational goal could be achieved via an algorithm that combines delay lines (one waveform from each source, or ear) with coincidence detectors, enabling a temporal activation pattern to be translated into a place code. In subsequent decades, Jeffress' influential model stimulated a range of anatomical, physiological, psychophysical, and formal investigations to understand sound localization (for an excellent review, see Grothe, 2003). In several tour-de-force experiments, the predictions of this algorithmic model were in large part confirmed at the implementation level in barn owls. That is to say, the algorithmically inspired experiments sought to identify possible implementations of the procedures, such as delay lines and coincidence detectors.

Given that success, we now have a causal, mechanistic explanation of barn owl sound localization that is embedded in a computational theory. The specific algorithm follows from the details of the circuit (e.g., the relevant nuclei of the barn owl are organized in a manner that directly licenses that algorithm and not some other formal procedure). But, crucially, to identify the precise circuitry that underpins delay-line-plus-coincidence computation, the computational level (behavioral) characterization and the algorithmic hypothesis were necessary prerequisites to identify the neural substrates. It is hard to imagine how (even extensive) recordings, staining, and anatomical studies of the barn owl auditory system in the absence of such a behaviorally motivated computational theory would have yielded such a rich explanatory framework.

It was subsequently discovered that the mammalian auditory system, such as in gerbils, solves the same computational problem using a different algorithm with different implementation at the circuit level (Grothe, 2003). This finding is a compelling example of multiple realizability and deep degeneracy, since different implementational features yield different algorithms, but both solve the same computational problem (Katz, 2016; Marom et al., 2009). Implementation/circuit A yields algorithm A′, which solves computational goal X; and implementation B yields algorithm B′, which also solves computational goal X (Figure 1D). Once again, without formally specified analyses at the behavioral level as well as explicit algorithmic hypotheses, these fully developed mechanistic accounts would have been hard, indeed perhaps impossible, to identify.

Electrolocation

The neurobiology of weakly electric fish is grounded in comparative, behavioral, and evolutionary analyses. Through a series of studies, Walter Heiligenberg was able to provide perhaps the most complete understanding of a vertebrate circuit from sensory input through to motor output in his analysis of the jamming avoidance response (JAR). JAR is a natural behavior (although possibly initially discovered as a stimulus response [Watanabe and Takeda, 1963]) in which two weakly electric fish that are discharging an oscillating electric field at the same frequency will

modify their discharge frequencies so as to avoid "jamming" each other's electrolocation system (Heiligenberg, 1991). This is important because their ability to localize objects in the dark is mediated by sensing distortions to their self-generated field. Analysis of the behavior was first performed by investigators who built a simple virtual world (Box 1) for electric fish: a small tank in which a fish's own electric field was picked up, then slightly shifted in frequency and fed back, resulting in the reliable elicitation of what is now called the JAR (Watanabe and Takeda, 1963).

As the electric field is controlled by the fish's motor system, it enabled the complete characterization of a behavior using precisely controlled signals and set the groundwork for Heiligenberg's subsequent discoveries of the circuits underlying the behavior (Heiligenberg, 1991). From that period forward, electric fish were typically studied by applying field distortions across the entire fish. As it turns out, these "wide-field" signals are interpreted by the fish as communication signals, since electrolocation targets such as prey result in focal "narrow-field" distortions. This was not fully understood until prey capture behavior was automatically tracked in three dimensions, and then an entire suite of empirically validated models was used to compute the neuronal signal going to the brain over the course of prey capture behavior (MacIver et al., 2001; Nelson and MacIver, 1999; Snyder et al., 2007). Once investigators started applying distortions in focal manner to mimic the effect of prey, they unmasked a filtering system within the hindbrain that processed these signals differently from the wide-field signals that had been formerly been examined (reviewed in Fortune, 2006). The point is that even within a behaviorally oriented model system, understanding key components of this system did not enable a full understanding of how the brain processes signals related to predation until careful behavioral and computational simulation work was done.

Motor Learning

A final example of the benefits of a well-motivated behavioral perspective comes from motor learning. Motor learning is often studied using adaptation paradigms in which participants are subjected to an external perturbation, which causes systematic errors that are then corrected (Shadmehr et al., 2010). A large body of empirical and theoretical work, originating in large part from Marr himself, has suggested a crucial role for cerebellar circuitry for this kind of learning (Therrien and Bastian, 2015). In essence, Marr initially believed—prior to elaborating his idea of levels of analysis—that the mechanism of error-based learning could be reverse-engineered from examination at the circuit level. Indeed, much work after Marr has made considerable progress in determining how cerebellar circuitry performs error-based learning. Crucially, however, it is now known from careful psychophysical work that many distinct learning algorithms operate together to counter the effects of a perturbation during adaptation, even though phenotypically their summed behavior can look like pure error-based cerebellar learning (Huang et al., 2011; Taylor et al., 2014). This is a further example of multiple realizability (Figure 1D), a long-standing basis for an objection to reductionism (Sober, 1999). The core argument is that if there are many ways to neurally generate the same behavior, then the properties of a single circuit at best are a particular instantiation and do not reveal a general design principle.

The distinct algorithms operating in adaptation experiments include error-based learning, reinforcement learning, and cognitive strategies (Huberdeau et al., 2015). The relative weighting of these processes depends on how the task is framed with respect to the kind of feedback and instructions given. For example, when subjects are given endpoint feedback and instruction they solve the same visuomotor rotation task differently to when no instruction is given and they are provided with continuous visual feedback. This higher-level organization operates at the level of the global task goal, which can only be identified top-down by observing the summed behavior and then decomposing it psychophysically, not by first knowing how each of these components is neurally implemented (Taylor et al., 2014); the adaptation task is perceived and solved by the whole brain in a body, not just by any given circuit component.

Two points are illustrated with these four examples. The first is that experiments at the level of neural substrate are best designed with hypotheses based on pre-existing behavioral work that has discovered or proposed candidate algorithmic and computational processes (for a range of arguments why experimental work based on algorithm-level hypotheses is foundational, see Cooper and Peebles, 2015). Second, the explanations of the results at the neural level are almost entirely dependent on the higher-level vocabulary and concepts derived from behavioral work. Lower levels of explanation do not "explain away" higher levels.

The Need for More Pluralistic Neuroscience

> *If we look far into the future of our science, what will it mean to say we 'understand' the mechanism of behaviour? The obvious answer is what may be called the neurophysiologist's nirvana: the complete wiring diagram of the nervous system of a species, every synapse labelled as excitatory or inhibitory; presumably, also a graph, for each axon, of nerve impulses as a function of time during the course of each behaviour pattern … Real understanding will only come from distillation of general principles at a higher level, to parallel for example the great principles of genetics—particulate inheritance, continuity of germ-line and non-inheritance of acquired characteristics, dominance, linkage, mutation, and so on … it seems possible that at higher levels some important principles may be anticipated from behavioural evidence alone. The major principles of genetics were all inferred from external evidence long before the internal molecular structure of the gene was even seriously thought about.*—Richard Dawkins (Dawkins, 1976, pp. 7–8)

Neuroscience has been focused of late on neural circuits. This is largely due to the recent development and incorporation of techniques that allow both causal manipulation and the rapid acquisition of large amounts of data. There seems to be an implicit assumption that implementation-level description will not only allow causal claims but also somehow lead to algorithmic and computational understanding ("naive" emergence). We

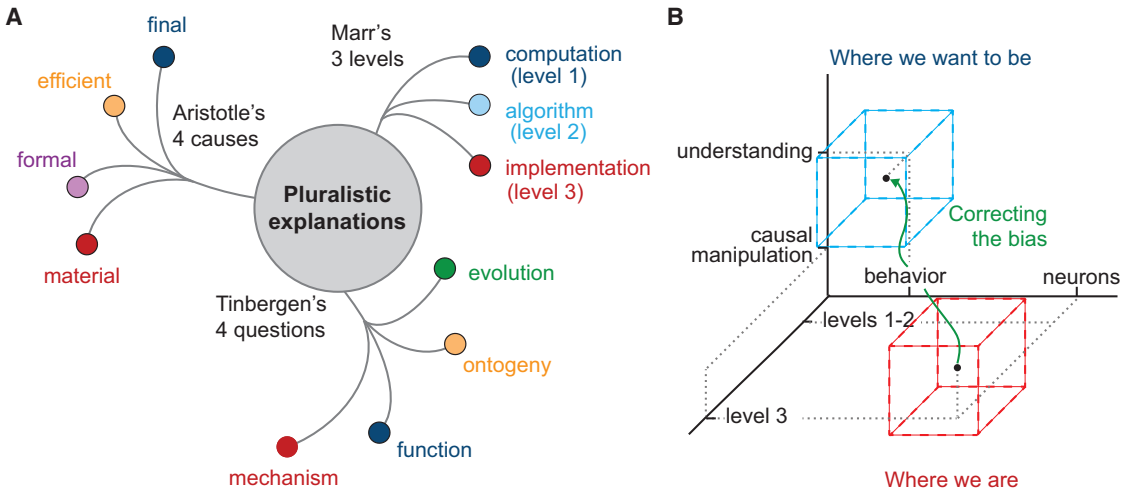


Figure 4. The Future History of Pluralistic Explanation

(A) That understanding of a phenomenon is multidimensional has long been appreciated. Aristotle posited four kinds of explanation: to explain "why" something changes, a polyhedric notion of causality is necessary; one that includes not only the material cause (what it is made out of), but also the other three "whys": formal (what it is to be), efficient (what produces it), and final (what it is for). Tinbergen also devised four questions about behavior: to go beyond its proximate causation (mechanism) to also considering its evolution, development, and real-world function. Marr's three levels are also shown.

(B) Three-dimensional space with axes of understanding-manipulation, behavior-neurons, and Marr's levels. The red box is where we are and the blue is where we should be.

contend that such an approach is simply not going to yield the kind of insight and explanation that we ultimately demand from the neurosciences, at least those parts concerned with developing an understanding of the link between brain and behavior that goes beyond causality claims.

Since the causal-manipulation view by itself will not lead to understanding, a more pluralistic conception of mechanistic understanding can only help neuroscience. Pluralism in science can be defined as "the doctrine advocating the cultivation of multiple systems of practice in any given field of science" (Chang, 2012). Here we have argued that when scientists ask "how does the brain generate behavior," they are in fact asking a question best approached through behavioral work, specifically task analysis, aided by theory, that allows behavior to be decomposed into separable modules and processing operations.

As Woese has argued (Woese, 2004), science is driven by both technological advances and a guiding vision. The key is to balance their contributions, "…without the proper technological advances the road ahead is blocked. Without a guiding vision there is no road ahead" (Woese, 2004, p. 173). Insofar as the goal of a neuroscience research question is to explain some behavior, be it a phenomenon from vision, communication, motor control, navigation, language, memory, or decision making, the behavioral research must be considered, for the most part, epistemologically *prior* (Figure 4). The neural basis of behavior cannot be properly characterized without first allowing for independent detailed study of the behavior itself.

ACKNOWLEDGMENTS

We thank the following for invaluable critical discussion about and/or careful reading of the evolving manuscript: Björn Brembs, Philip Buttery, Rui Costa, William Dauer, Greg DeAngelis, Shimon Edelman, Stuart Firestein, Adrian Haith, David Huberdeau, Don Katz, Konrad Körding, David Linden, Zachary Mainen, Tony Movshon, Yael Niv, Bence Ölveczky, Bijan Pesaran, Scott Rennie, Nick Roy, Scott Small, Aaron Wong, and Jing Xu.

REFERENCES

Anderson, P.W. (1972). More is different. Science *177*, 393–396.

Anderson, D.J., and Perona, P. (2014). Toward a science of computational ethology. Neuron *84*, 18–31.

Badre, D., Frank, M.J., and Moore, C.I. (2015). Interactionist Neuroscience. Neuron *88*, 855–860.

Bargmann, C.I. (1998). Neurobiology of the *Caenorhabditis elegans* genome. Science *282*, 2028–2033.

Bass, A.H., and Chagnaud, B.P. (2012). Shared developmental and evolutionary origins for neural basis of vocal-acoustic and pectoral-gestural signaling. Proc. Natl. Acad. Sci. USA *109* (*Suppl 1*), 10677–10684.

Bechtel, W. (2008). Mental Mechanisms: Philosophical Perspectives on Cognitive Neuroscience (Routledge).

Benichov, J.I., Benezra, S.E., Vallentin, D., Globerson, E., Long, M.A., and Tchernichovski, O. (2016). The forebrain song system mediates predictive call timing in female and male zebra finches. Curr. Biol. *26*, 309–318.

Bianco, I.H., Kampff, A.R., and Engert, F. (2011). Prey capture behavior evoked by simple visual stimuli in larval zebrafish. Front. Syst. Neurosci. *5*, 101.

Bickle, J. (2015). Marr and reductionism. Top. Cogn. Sci. *7*, 299–311.

Bickle, J. (2016). Revolutions in neuroscience: Tool development. Front. Syst. Neurosci. *10*, 24.

Borjon, J.I., and Ghazanfar, A.A. (2014). Convergent evolution of vocal cooperation without convergent evolution of brain size. Brain Behav. Evol. *84*, 93–102.

Carandini, M. (2012). From circuits to behavior: a bridge too far? Nat. Neurosci. *15*, 507–509.

Chang, H. (2012). Is Water H2O?: Evidence, Realism and Pluralism *Volume 293* (Springer Science & Business Media).

Cooper, R.P., and Peebles, D. (2015). Beyond single-level accounts: the role of cognitive architectures in cognitive scientific explanation. Top. Cogn. Sci. *7*, 243–258.

Craver, C.F. (2008). Explaining the brain: mechanisms and the mosaic unity of neuroscience. Psychol. Med. *38*, 899–900.

Dawkins, R. (1976). Hierarchical organisation: A candidate principle for ethology. In Growing Points in Ethology, P. Bateson and R.A. Hinde, eds. (Cambridge University Press), pp. 7–54.

Dennett, D.C. (1989). The Intentional Stance (MIT Press).

di Pellegrino, G., Fadiga, L., Fogassi, L., Gallese, V., and Rizzolatti, G. (1992). Understanding motor events: a neurophysiological study. Exp. Brain Res. *91*, 176–180.

Dombeck, D.A., and Reiser, M.B. (2012). Real neuroscience in virtual worlds. Curr. Opin. Neurobiol. *22*, 3–10.

Fetsch, C.R. (2016). The importance of task design and behavioral control for understanding the neural basis of cognitive functions. Curr. Opin. Neurobiol. *37*, 16–22.

Fortune, E.S. (2006). The decoding of electrosensory systems. Curr. Opin. Neurobiol. *16*, 474–480.

Frank, M.J., and Badre, D. (2015). How cognitive theory guides neuroscience. Cognition *135*, 14–20.

Gomez-Marin, A., Paton, J.J., Kampff, A.R., Costa, R.M., and Mainen, Z.F. (2014). Big behavioral data: psychology, ethology and the foundations of neuroscience. Nat. Neurosci. *17*, 1455–1462.

Grothe, B. (2003). New roles for synaptic inhibition in sound localization. Nat. Rev. Neurosci. *4*, 540–550.

Heiligenberg, W. (1991). Neural Nets in Electric Fish (MIT Press).

Hickok, G. (2009). Eight problems for the mirror neuron theory of action understanding in monkeys and humans. J. Cogn. Neurosci. *21*, 1229–1243.

Huang, V.S., Haith, A., Mazzoni, P., and Krakauer, J.W. (2011). Rethinking motor learning and savings in adaptation paradigms: model-free memory for successful actions combines with internal models. Neuron *70*, 787–801.

Huberdeau, D.M., Krakauer, J.W., and Haith, A.M. (2015). Dual-process decomposition in human sensorimotor adaptation. Curr. Opin. Neurobiol. *33*, 71–77.

Jeffress, L.A. (1948). A place theory of sound localization. J. Comp. Physiol. Psychol. *41*, 35–39.

Jonas, E., and Kording, K. (2017). Could a neuroscientist understand a microprocessor? PLoS Comput. Biol. *13*, e1005268.

Kahneman, D. (2011). Thinking, Fast and Slow (Macmillan).

Katz, P.S. (2016). Evolution of central pattern generators and rhythmic behaviours. Philos. Trans. R. Soc. Lond. B Biol. Sci. *371*, 20150057.

Kay, L.E. (1996). The Molecular Vision of Life: Caltech, the Rockefeller Foundation, and the Rise of the New Biology (Oxford University Press).

Leininger, E.C., and Kelley, D.B. (2015). Evolution of Courtship Songs in Xenopus : Vocal Pattern Generation and Sound Production. Cytogenet. Genome Res. *145*, 302–314.

Levitis, D.A., Lidicker, W.Z., and Freund, G. (2009). Behavioural biologists don't agree on what constitutes behaviour. Anim. Behav. *78*, 103–110.

Longuet-Higgins, H.C. (1972). The algorithmic description of natural language. Proc. R. Soc. Lond. B Biol. Sci. *182*, 255–276.

MacIver, M.A. (2009). Neuroethology: From Morphological Computation to Planning. In The Cambridge Handbook of Situated Cognition, P. Robbins and M. Aydede, eds. (Cambridge University Press), pp. 480–504.

MacIver, M.A., Sharabash, N.M., and Nelson, M.E. (2001). Prey-capture behavior in gymnotid electric fish: motion analysis and effects of water conductivity. J. Exp. Biol. *204*, 543–557.

Marder, E., and Goaillard, J.M. (2006). Variability, compensation and homeostasis in neuron and network function. Nat. Rev. Neurosci. *7*, 563–574.

Marom, S., Meir, R., Braun, E., Gal, A., Kermany, E., and Eytan, D. (2009). On the precarious path of reverse neuro-engineering. Front. Comput. Neurosci. *3*, 5.

Marr, D. (1982/2010). Vision: A Computational Approach (MIT Press).

Mazzoni, P., Hristova, A., and Krakauer, J.W. (2007). Why don't we move faster? Parkinson's disease, movement vigor, and implicit motivation. J. Neurosci. *27*, 7105–7116.

Nelson, M.E., and MacIver, M.A. (1999). Prey capture in the weakly electric fish *Apteronotus albifrons*: sensory acquisition strategies and electrosensory consequences. J. Exp. Biol. *202*, 1195–1203.

Noble, D. (2012). A theory of biological relativity: no privileged level of causation. Interface Focus *2*, 55–64.

Oatley, K. (1978). Perceptions and Representations: The Theoretical Bases of Brain Research and Psychology (Methuen).

Panigrahi, B., Martin, K.A., Li, Y., Graves, A.R., Vollmer, A., Olson, L., Mensh, B.D., Karpova, A.Y., and Dudman, J.T. (2015). Dopamine is required for the neural representation and control of movement vigor. Cell *162*, 1418–1430.

Patterson, B.W., Abraham, A.O., MacIver, M.A., and McLean, D.L. (2013). Visually guided gradation of prey capture movements in larval zebrafish. J. Exp. Biol. *216*, 3071–3083.

Poeppel, D., and Embick, D. (2005). Defining the relation between linguistics and neuroscience. In Defining the Relation between Linguistics and Neuroscience Twenty-First Century Psycholinguistics: Four Cornerstones, A.E. Cutler, ed. (Lawrence Erlbaum), pp. 103–118.

Reynolds, C.W. (1987). Flocks, herds and schools: A distributed behavioral model. Comput. Graph. *21*, 25–34.

Schall, J.D. (2004). On building a bridge between brain and behavior. Annu. Rev. Psychol. *55*, 23–50.

Selverston, A.I. (1980). Are central pattern generators understandable? Behav. Brain Sci. *3*, 535–540.

Shadmehr, R., Smith, M.A., and Krakauer, J.W. (2010). Error correction, sensory prediction, and adaptation in motor control. Annu. Rev. Neurosci. *33*, 89–108.

Smit, H., and Hacker, P.M. (2014). Seven misconceptions about the mereological fallacy: A compilation for the perplexed. Erkenntnis *79*, 1077–1097.

Snyder, J.B., Nelson, M.E., Burdick, J.W., and MacIver, M.A. (2007). Omnidirectional sensory and motor volumes in electric fish. PLoS Biol. *5*, e301.

Sober, E. (1999). The multiple realizability argument against reductionism. Philos. Sci. *66*, 542–564.

Sternberg, S. (2011). Modular processes in mind and brain. Cogn. Neuropsychol. *28*, 156–208.

Takahashi, D.Y., Narayanan, D., and Ghazanfar, A.A. (2012). A computational model for vocal exchange dynamics and their development in marmoset monkeys. In 2012 IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL), http://dx.doi.org/10.1109/DevLrn.2012.6400844.

Takahashi, D.Y., Narayanan, D.Z., and Ghazanfar, A.A. (2013). Coupled oscillator dynamics of vocal turn-taking in monkeys. Curr. Biol. *23*, 2162–2168.

Takahashi, D.Y., Fenley, A.R., Teramoto, Y., Narayanan, D.Z., Borjon, J.I., Holmes, P., and Ghazanfar, A.A. (2015). LANGUAGE DEVELOPMENT. The developmental dynamics of marmoset monkey vocal production. Science *349*, 734–738.

Taylor, J.A., Krakauer, J.W., and Ivry, R.B. (2014). Explicit and implicit contributions to learning in a sensorimotor adaptation task. J. Neurosci. *34*, 3023–3032.

Therrien, A.S., and Bastian, A.J. (2015). Cerebellar damage impairs internal predictions for sensory and motor function. Curr. Opin. Neurobiol. *33*, 127–133.

Tinbergen, N. (1955). The Study of Instinct (Clarendon Press).

Tinbergen, N. (1963). On aims and methods of ethology. Zeitschr. Tierpsychol. *20*, 410–433.

Von Uexküll, J. (1992). A stroll through the worlds of animals and men: A picture book of invisible worlds. Semiotica *89*, 319–391.

Watanabe, A., and Takeda, K. (1963). The change of discharge frequency by A.C. stimulus in a weak electric fish. J. Exp. Biol. *40*, 57–66.

White, J.G., Southgate, E., Thomson, J.N., and Brenner, S. (1986). The structure of the nervous system of the nematode Caenorhabditis elegans. Philos. Trans. R. Soc. Lond. B Biol. Sci. *314*, 1–340.

Woese, C.R. (2004). A new biology for a new century. Microbiol. Mol. Biol. Rev. *68*, 173–186.

Yuste, R. (2015). From the neuron doctrine to neural networks. Nat. Rev. Neurosci. *16*, 487–497.