On the role of theory and modeling in neuroscience

Daniel Levenstein[1], Veronica A. Alvarez[2], Asohan Amarasingham[14], Habiba Azab[3], Richard C. Gerkin[4], Andrea Hasenstaub[5], Ramakrishnan Iyer[6], Renaud B. Jolivet[7], Sarah Marzen[12], Joseph D. Monaco[8], Astrid A. Prinz[13], Salma Quraishi, Fidel Santamaria[9], Sabyasachi Shivkumar[15],Matthew F. Singh[10], David B. Stockton, Roger Traub[11], Horacio G. Rotstein[16,*], Farzan Nadim[16,*], A. David Redish[17,*]

[1]NYU Center for Neural Science; NYU Neuroscience Institute. New York, NY

[2]Laboratory on Neurobiology of Compulsive Behaviors, National Institute on Alcohol Abuse and Alcoholism, NIH. Bethesda, MD

[3]Department of Neuroscience, Center for Magnetic Resonance Research, University of Minnesota, Minneapolis, MN

[4]School of Life Sciences, Arizona State University, Tempe, AZ

[5]Department of Otolaryngology - Head and Neck Surgery, University of California San Francisco, San Francisco, CA

[6]Allen Institute for Brain Science, Seattle, WA

[7]Department of Nuclear and Corpuscular Physics, University of Geneva, Geneva, Switzerland

[8]Department of Biomedical Engineering, Johns Hopkins University School of Medicine, Baltimore, MD

[9]Department of Biology, University of Texas at San Antonio, San Antonio, TX

[10]Department of Psychological & Brain Sciences, Department of Electrical & Systems Engineering, Washington University in St. Louis, St. Louis, MO

[11]IBM T.J. Watson Research Center, AI Foundations, Yorktown Heights, NY

[12]W. M. Keck Science Department, Pitzer, Scripps, and Claremont McKenna Colleges, Claremont, CA

[13]Department of Biology, Emory University, Atlanta, GA

[14]Department of Mathematics, City College of New York; Departments of Biology, Computer Science, and Psychology, The Graduate Center; City University of New York

[15]Brain and Cognitive Sciences, University of Rochester, Rochester, New York, United States of America

[16]Federated Department of Biological Sciences, New Jersey Institute of Technology and Rutgers University & Institute for Brain and Neuroscience Research, New Jersey Institute of Technology, Newark, NJ

[17]Department of Neuroscience, University of Minnesota, Minneapolis MN

*Co-senior author

Abstract

In recent years, the field of neuroscience has gone through rapid experimental advances and extensive use of quantitative and computational methods. This accelerating growth has created a need for methodological analysis of the role of theory and the modeling approaches currently used in this field. Toward that end, we start from the general view that the primary role of science is to solve empirical problems, and that it does so by developing theories that can account for phenomena within their domain of application. We propose a commonly-used set of terms — descriptive, mechanistic, and normative — as methodological designations that refer to the kind of problem a theory is intended to solve. Further, we find that models of each kind play distinct roles in defining and bridging the multiple levels of abstraction necessary to account for any neuroscientific phenomenon. We then discuss how models play an important role to connect theory and experiment, and note the importance of well-defined translation functions between them. Furthermore, we describe how models themselves can be used as a form of experiment to test and develop theories. This report is the summary of a discussion initiated at the conference *Present and Future Theoretical Frameworks in Neuroscience*, which we hope will contribute to a much-needed discussion in the neuroscientific community.

Correspondence: dl2820@nyu.edu, horacio@njit.edu, farzan@njit.edu, redish@umn.edu

Introduction

Theories are the primary tools by which scientists make sense of observations and make predictions. Given this central role, it is surprising how little methodological attention is given within the sciences to the general nature of theories: what they are, how they are used, and the processes by which they are developed. In part, this may be due to different uses of theory across different scientific fields, but it may also be due to the historical accident of how different scientific fields have grown.

Neuroscience is among the fastest growing areas in biology, due in part to strong funding boosts from the United States and European Union, through standard budgetary funding and special measures such as the *BRAIN Initiative*. While neuroscience has a strong history of experimental, theoretical, and computational interactions, often predicting biological mechanisms decades before their experimental confirmation (1–3), interactions within the field make it clear that there is still confusion as to the nature of theory, its role in neuroscience, and how it should be developed, used, and evaluated within the community (4, 5). Additionally, the surge of new technologies to scrutinize the nervous system at unprecedented levels have made clear the need to develop new theoretical frameworks to assimilate the growing quantities of resulting data (6) and establish relationships between their underlying processes.

Under the auspices of the National Science Foundation and the *BRAIN Initiative*, dozens of theoretical and experimental neuroscience researchers came together for a workshop, entitled Present and Future Theoretical Frameworks in Neuroscience (4-8 Feb 2019, San Antonio, TX) (Rotstein HG, Santamaría F. *In preparation*), to discuss future directions in theoretical neuroscience and to explore theoretical frameworks that would allow the neuroscience community to benefit from novel neurotechnologies. In this document, we report some of the conversations and insights from the San Antonio meeting on classifications of scientific theory, modeling, and simulations with a specific focus in their use in neuroscience. This document originated from the discussions of one out of the several discussion workgroups, and is not meant to capture the outcome of the full meeting. Furthermore, we do not aim to provide definite answers, but rather open a much needed discussion among the neuroscience community about scientific methodology.

Towards that end, we begin by outlining an idealized view of scientific progress we think best captures neuroscientific practice. By this view, science is a problem-solving endeavor in which we use models to connect theory and phenomena. We propose that a set of commonly-used categorizations - **descriptive**, **mechanistic**, and **normative** models - are best seen as corresponding to the type of problem a theory is being used to solve, and can thus act as methodological guides for scientific practice. We then show how this categorization has implications for how each of these approaches relates to the multiple levels of abstraction needed to account for neuroscientific phenomena, namely that descriptive models define a representation of a phenomenon at one level of abstraction, while mechanistic and normative models bridge levels of abstraction. These operations constitute one of the key roles of theoretical neuroscience, as they unify scientific theories across disparate experimental approaches and fields. Finally, we discuss the relationship between theory and experiment, and how models can themselves be used in a form of experiment in the ongoing process of theory development. This discussion leads to recommendations for how to formulate theory projects towards the goal of developing the necessary future theoretical frameworks of neuroscience.

What is a theory and what is it good for?

To inform the development of theoretical frameworks, we first consider the nature of theory and how it is used in scientific practice. Traditional descriptions of science tend to be based on the processes of

theory identification and falsification, with theories as proposed universal truths about the world to be tested and, eventually, rejected (7). In this formulation, when current theories can no longer explain observed data, paradigm shifts occur and new theoretical frameworks arise that can better account for the data (8). However, historical and sociological analyses show that these views do not describe scientific practice (9–12). New discoveries engender new questions not just answers (13), and scientists can have a variety of attitudes towards a theory rather than to simply accept or reject it (9, 14). Some have argued that theories need only be empirically adequate for a task at hand, without requiring belief in its underlying theoretical entities (15–17).

More complex views on the general role of theory in science have been proposed, such as arguments for the importance of mechanistic explanations and for experiments contributing to model confirmation and rejection (14); arguments for the importance of applied science in providing solutions to problems of controlling the natural world (18, 19); and arguments for the importance of theories in the creation of new questions, rather than simply answering old ones (13). Additional complexity lies in the theory-laden nature of observation, in which the decision of what to measure, how to measure it, and what those measures mean all depend on theoretical assumptions (whether stated explicitly or not), and in the fact that experiments are by their nature fraught with potential errors (some of which may be unrecognized at the time (20, 21)). While these issues have been widely discussed in the philosophy community, they do not seem to be playing much of a role in current discussions of scientific practice.

A pragmatic view: science as problem-solving

We take a pragmatic view of the scientific endeavor, which we think most adequately describes and informs scientific practice. By this view, scientific practice is a problem-solving endeavor (10, 18, 19, 22): a process by which we solve empirical problems[1]. Empirical problems are questions about observed phenomena, which can range from matters of purely scientific interest such as "How does the brain process visual signals?" or "How does an animal select between alternative choices?" to those with more obvious applications such as "Which brain functions are disrupted in schizophrenia?". Like any other problem, solving a scientific problem can be seen as a search to achieve a goal, which is specified by the statement of the problem along with criteria on what counts as an acceptable solution (23). However, scientific problems are often ill-defined (24), because the relevant questions and solution criteria are not always explicitly stated (e.g., during exploratory stages of research), and both evolve with additional discoveries (13). For example, the understanding that multiple memory systems (25–28) interact to produce multiple decision-making systems (3, 29–31) leads to the question of what happens when those systems are in conflict, while the question "How does the pineal gland generate consciousness?" (32) is now considered outdated and obsolete, if not nonsensical. Further, what may be seen as an adequate solution to a problem in one sociohistorical context may not be in another - as new data become available, standards change, or alternative solutions are presented. Despite the evolving landscape of problems and their proposed solutions, scientific theories have been used to make progressively more accurate predictions about more phenomena over history (18, 19). We maintain that this progress results from community-maintained standards under the drive to better predict and control environmental factors of potential relevance to society (19).

A problem-based view of scientific progress shifts theories from "proposals of truth to be falsified" to "proposed problem-solving tools," and prompts us to assess their utility: what empirical problems they can solve, how readily they can be used to do so, and how good their solutions are. This shift raises the following question: what constitutes a solution to a scientific problem? Or, equivalently, what are the

---

[1] Note: other kinds of scientific problems have also been discussed, such as conceptual problems: conceptual inconsistencies within or between theories and their associated worldviews/frameworks (8, 18). For conciseness we focus only on empirical problems.

criteria by which a theory is deemed to have "solved" a problem? Assessing solutions to scientific problems requires general standards to evaluate the quality of the solution and constraints on the *form* solutions can take. The first criterion (standards) may include features like accuracy of predictions, falsifiability, simplicity, and reproducibility - a set of epistemic virtues (15, 33, 34) which are the measure of a theory's utility, as well as the foundation of any inductive relationship between scientific practice and truth (9). The second criterion (constraints on form) includes an overarching set of concepts which can be used to solve problems and thus constitute a *theoretical framework* (table 1): a language of terminology within which theories are proposed. For example, the solutions proposed under the framework of Freudian psychology (35) are fundamentally different from those proposed under the modern medicalized frameworks of psychiatry (36–39), and thus the set of theories under the two frameworks would be composed of fundamentally different objects and language. While Freudian psychology suggests subconscious consequences of parental interactions, modern medicalized frameworks suggest physical dysfunctions in neural structure caused by pharmacological imbalances or genetic differences. In effect, the theoretical framework is a set of foundational theories that make up the "core" of a research program, and serve as foundational assumptions for other theories and experiments within that program (8, 9, 18)[2]. In addition to constraints on the content of theories, any theoretical framework includes a set of general problem-solving methods used within that framework.

We can thus consider the solution to an empirical problem to be a scientific explanation (40, 41) in which a theory is proposed to solve the problem[3] and that explanation is deemed adequate (or not) based on solution criteria of the problem (24), under the constraints imposed by the current theoretical framework. The scientific community has a continuing responsibility to evaluate their problem-solving criteria and methodologies -- to assess whether modifications are required, and if those modifications should involve the development of new theories, or if the development of entirely new frameworks is needed.

Model-based scientific explanation

So what does a scientific explanation look like, particularly in the current framework of theoretical neuroscience? We will, broadly speaking, consider a *theory* to be an idea or set of ideas that can be used as part of an explanation for observed phenomena (i.e., to solve an empirical problem). We take the (pragmatic (42)) view that theories are generally amorphous and defined by their function and practice, namely their use for scientific explanation. The main explanatory tool of the theoretical neuroscientist is a *model*: a structure that is interpreted to represent observed or theoretical phenomena (43, 44) (Table 1). Unlike theories, models are well-defined to the extent that they include a description of their structure and its intended interpretation. For example, the equation $\tau dV/dt = -V + E_L$ describes a mathematical structure that can be interpreted to represent the membrane potential, V, of a passive cell with time constant, τ, and resting potential, E_L (45). Mathematical or computational models force us to confront hidden assumptions in our theories (46) and are amenable to simulation and analytical treatment. However, even if we do not explicitly express a model in mathematical terms, we can still make an implicit (heuristic, intuitive, or mental) model of the phenomena at hand (47). Further, models can include other kinds of interpreted structures, such as structural representations of the double helix of DNA (48) or diagrammatic representations of the protein interactions involved in signaling cascades (49). For example, in experimental neuroscience physical structures are used in the form of "animal models" such as the 6-OHDA rat or the MPTP monkey, which are interpreted to represent the pathology of Parkinson's disease (50, 51). These different kinds of models are examples of the same overarching concept: namely

---

[2] "To accept one theory rather than another one involves a commitment to a research programme, to continuing the dialogue with nature in the framework of one conceptual scheme rather than another." (15).

[3] I.e., the theory corresponds to the explanans, and the problem corresponds to the explananda (40).

a structure, representing some phenomenon of interest, which can be used to solve empirical problems about that phenomenon.

In creating a model, a researcher has to make foundational assumptions in the terms they use, the form those terms take, and the relationships between them. We propose that these assumptions are the instantiation of a theory: they are explicit expressions of aspects of the theory in a well-defined form. The voltage equation above instantiates a theory that a neuron's electrical properties arise from a semipermeable membrane (2, 52, 53), while the 6-OHDA model instantiates the theory that Parkinson's disease arises due to dopaminergic dysfunction in the substantia nigra (54). Thus, models can simultaneously act as an instantiation of some aspects of a theory and as an abstraction for some aspects of a phenomenon (55, 56). As we will show, this dual role of models forms the foundation of model-based scientific explanation (57) -- one of the primary problem-solving strategies in theoretical neuroscience (53, 58, 59).

Frameworks, theories, and models interact throughout neuroscientific practice. Frameworks provide the form in which solutions are proposed and compared. Theories are the set of ideas used in those explanations. Models instantiate the theories in a structure that can be directly explored. Table 1 shows three examples from cellular, systems, and clinical neuroscience.

	Examples		
	Cellular	*Systems*	*Disease*
Framework A general description about the structure of the world, providing a language and a conceptual basis for developing theories.	Explanations for differences in neural functional properties can be appropriately described in terms of differences in the electrochemical properties of membranes and proteins.	Explanations of the production of movement by skeletal muscle contractions can be appropriately described in terms of patterns of action potentials in the central nervous system.	Explanations of neurodegenerative diseases can be appropriately described in terms of dysfunction in cellular processes.
Theory A set of ideas that can be used to explain a set of phenomena (the domain of the theory).	Specific voltage gated ion channels enable excitable properties of neurons such as the action potential.	Many movements are generated by central pattern generators that are primarily driven by internal oscillatory dynamics.	Parkinson's disease is due to loss of dopaminergic function in the substantia nigra.
Model An instantiation of a theory in an (often mathematical) structure, which is interpreted to represent a phenomenon.	The Hodgkin-Huxley equations capture the essential qualitative and quantitative properties of the action potential.	Half-cycle oscillators represent swimming processes in the lamprey, driven by alternating waves down the notochord.	The 6-OHDA rat and MPTP monkey have dopaminergic loss in the substantia nigra and show Parkinsonian behaviors such as bradykinesia and tremors.

Table 1: *Terminology used in this manuscript. Three neuroscience examples.*

Three kinds of theory; three kinds of models

Theories can have multiple aspects that can be instantiated in a variety of model formulations, and numerous classifications have been proposed (44). A problem-solving view of scientific practice prompts us to consider a classification of theories and models based on the type of problem they are being used to solve. We describe a taxonomy of common approaches in neuroscientific practice, which are used to solve three different types of problems: "what" problems, "how" problems, and "why" problems (60). As the same model can serve different purposes, models cannot be assigned to one class or another without stating the problem they are being used to solve. This apparent ambiguity can result in disagreement as to which class a given model should be considered. However, we argue that a focus on problem-context can help resolve this ambiguity, and serve as a guide for researchers to identify common ways of solving similar problems.

Descriptive Explanations

The first order question encountered in scientific research is: *what is the phenomenon?* This problem is solved with a **descriptive** theory, which is used to provide a concise summary of the phenomenon (61). Descriptive models are founded on basic assumptions of which variables to observe and how to relate them. For example, the theory that hippocampal cells are "place cells" (3) describes a set of properties that could be instantiated in a model specifying when a hippocampal cell will fire within an environment (3, 62, 63). At its heart, a descriptive model is simply a compressed representation of phenomenological data -- descriptive models are often called phenomenological models (64, 65), or, when they are well-established, phenomenological laws (57).

Much of the classical work in biology is in the form of descriptive models – diagrams and classification of observed structures in nature (66, 67). Such descriptive models were primarily qualitative, yet provided the basis for most biological discoveries until the mid twentieth century. Quantitative descriptive models specify the relationship between observable variables in a functional form for which modern statistical methods (68, 69) can be used to directly fit parameters to data and estimate their variability or goodness of fit. These methods specify relationships between variables with well-defined probability models and explicit statistical assumptions that can reveal ambiguities inherent in their qualitative counterparts (70, 71)

In delineating the attributes that define a phenomenon, descriptive theories delineate the attributes that are expected to be repeatable in future experiments and the necessary conditions for repeatability. This is extremely important for the current replication controversy (22, 72–74). The recent National Academy report (75) differentiates between reproducibility and replicability - reproducibility is obtaining the same results from the same data, while replicability is obtaining consistent results across multiple studies. Several authors have pointed out that claims should be replicable, not data, and that the replication crisis is in fact a crisis of theory development (22, 76, 77).

Mechanistic Explanations

After addressing the "what" question, one might ask: *how does the phenomenon arise?* This problem is solved with a **mechanistic theory**, which is used to explain a phenomenon in terms of its component parts, their actions, and their organization (24, 64, 78). A mechanistic model is based on an assumption of which parts and processes are relevant to the phenomenon, and illustrates how their interaction can produce a phenomenon or, equivalently, how phenomena can *emerge* from these parts. For example, traffic is an emergent phenomenon that can be represented by a mechanistic model that includes

interactions between the nuts and bolts of cars, the timing of traffic lights, the reaction times of drivers, equations of non-compressible flow, etc. (79–81).

Quantitative mechanistic models often take the form of a dynamical system (53, 58, 82–85), in which a set of variables (or their equilibrium conditions) represent the temporal evolution of component processes. For example, the classic Hodgkin-Huxley model uses a set of four coupled differential equations to represent the dynamics of membrane potential and voltage-dependent conductances, and shows how an action potential can emerge from their interaction, by producing a precise prediction of the progression of the membrane potential in time (2). However, qualitative mechanistic models are also commonly used in biology, psychology, and neuroscience, in which mechanistic theories of complex processes are summarized in schematic or conceptual structures that represent general properties of components and their interactions. For example, Hebb considered a conceptualization of neural processing in which coincident firing of synaptically-connected neurons strengthened the coupling between them. From this model, Hebb was able to intuit how memories could be retrieved by the completion of partial patterns (content-addressable memory) and how these processes could emerge from synaptic plasticity, as cells that were coactive during a particular stimulus or event would form assemblies with pattern-completion properties (86).

The utility of mechanistic explanation lies in the fact that mechanistic models represent (assumed) underlying processes that produce the phenomenon (61, 64). As a result, a mechanistic model can be used to make predictions about any circumstances where the same processes are presumed to operate (82) - including the effects of manipulations to component parts, and even circumstances beyond the scope of data used to calibrate the model. When the target phenomena and its underlying processes correspond to simultaneously-observable quantities, causal inference methods have formalized statistically-rigorous ways to express and assess the causal relationships represented in mechanistic models (87–90). These methods can be a powerful tool to constrain mechanistic models by observable data, and can even be used to identify latent variables that can provide further explanatory power. The latent variables are identified by asking *what if* changes are made to a mechanistic model, which are done by manipulating them and asking what alternate situations would obtain (i.e. counterfactual simulations). Causal inference has been used to identify relationships among parameters in genetic, epidemiological, behavioral, and psychiatric phenomena (91–93) and has been applied to fMRI, EEG, MEG, local field potentials and other continuous signals in neuroscience (94, 95). However, there are still issues in applying these approaches to phenomena in complex, non-stationary, and uncontrollable background environments. Furthermore, many mechanistic models represent processes that are not simultaneously measureable in practice, or even in principle due to their degree of abstraction from observable quantities

Normative Explanations

In contrast to the mechanistic question of "how", we can also ask the question: *why does the phenomenon exist?* This kind of problem is solved with a **normative theory,** which is used to explain a phenomenon in terms of a function or goal (96–98). Implicit normative theories are frequently used in biological sciences when we talk about the function of a system - for example that the visual system is "for" processing visual information. This function serves as a guiding concept that can be a powerful heuristic to explain the system's behavior based on what it ought to do to perform its function. When quantified, normative models formalize the goal in terms of an objective function (also known as a utility or cost function), which can be optimized under some constraints to derive the best possible solution. These models are founded on an assumed statement of a goal to be optimized and the constraints

under which the system is achieving the goal. Further, such an approach depends on an assumption of optimization - that the system is attempting to optimize some well-defined cost function.

The assumption of optimality is often founded on evolutionary arguments (e.g. (96, 99)), which through competition might be expected to optimize systems (100). Indeed, in some cases, normative models provide evidence that the system is performing close to optimal given physical limitations. For example, retinal transduction of light can be shown to be close to optimal, particularly for certain structures (such as rods in the eyes of cats) (101–103). However, evolution has no guarantee of optimization. For example, mammalian eyes are suboptimal because, in reference to the path of light, the photoreceptors are in the deepest layer of the retina, requiring a path for the axons (which originate from ganglion cells in the most superficial layer) to leave the retina, thus producing a blind spot (104). Notably, lens eyes in octopus and other cephalopods are not inverted in this way and do not have a blind spot (104). These differences speak to the limitations of evolutionary systems to achieve optimality and underlying costs imposed by the limitations of genetic search (105). Moreover, there are other physical processes that may perform some operation akin to optimization (e.g. learning, economic markets, etc. (106)). Different optimization processes may themselves impose distinct constraints that give unique signatures to the systems they optimize.

In fact, the usefulness of normative models often lies in their ability to identify when an observed system is performing suboptimally (100), which can provide additional information about unexpected goals or constraints. As another example, in a series of papers, Redish and colleagues found that rodent foraging is suboptimal, under the assumption that the animals are maximizing reward intake. In particular, the animals tended to remain at the reward sites longer and to accept expensive (long-delay) offers more than what was required for an optimal reward intake, implying suboptimality in decision processes (107–110). However, optimality could be restored by assuming an additional hidden cost in the cost-function (111) . Subsequent neurophysiological, computational, and behavioral analyses have revealed that this hidden cost shows functionality akin to "regret" at leaving an option (107, 110, 112).

On the context-specific nature of theory/model classification

Neither theory nor model exist in isolation, but are embedded in scientific practice. As our categorization reflects the problem being solved, it can be applied to either theories or models, depending on the context in which they are being used and the problems they are proposed to solve. Further, models with the same structure can be used for different purposes, and can thus be assigned a different category in different contexts. For example, the integrate and fire model can be used as a descriptive model for membrane potential dynamics or as a mechanistic model for the neuronal input-output transformation; and while the Hodgkin-Huxley model was discussed above as a mechanistic model for the problem of spike generation, it was originally proposed to be "an empirical description of the time course of the changes in permeability to sodium and potassium" (2). We also note that this categorization is *independent of an explanation being accepted by the scientific community*. A mechanistic explanation does not cease to be mechanistic if it is not adopted (e.g., even if some of its experimental predictions don't bear fruit, it is still proposed to solve the problem of how a phenomenon emerges from its parts).

As with the context-dependence of model classification, a given theory might have descriptive, mechanistic, and normative aspects. In fact, a theory may start as an effort to solve one class of problem, but over time develop aspects to address other problems in its domain. As a case study, consider the theory of visual coding: a body of ideas concerning neural activity in the so-called visual areas of the brain and its relationship to visual stimuli. Foundational work of the theory took the form of a descriptive model to explain the tuning properties of single neurons in the primary visual cortex, namely that their responses to visual stimuli are well represented by Gabor patches with a given receptive field,

preferred orientation, and spatial frequency (113, 114). A mechanistic model was proposed to explain how the observed tuning properties arise, relying on the convergence of inputs from center/surround on/off cells of the visual thalamus, and further mechanistic models have since been developed and incorporated into the theory (115). Recent work has used generalized linear models (GLMs) to further refine and parameterize models that describe the relationship between sensory input and neuronal spiking in ways that can be directly fit to neural data (116), and these models can even be formulated in ways that are "biophysically-interpretable", or directly connect to accepted mechanisms of spike generation (53, 117). Finally, a series of normative models have been used to explain why Gabor patches are the optimal solution to accurately represent visual space, under constraints such as minimizing the number of spikes fired (118). This example further illustrates different roles models can play and complement each other in a theory: often the terms in mechanistic or normative models are themselves descriptive models that describe the behavior of components or properties of constraints, and mechanistic/normative models are used to explain how/why descriptive models take the form they do, either as emergent properties or in order to perform some function. As we will describe in the next section, this corresponds to the differential role of the model types in connecting "levels of abstraction" - which we see as one of the primary roles of theoretical research.

Levels of abstraction

Abstraction consists in replacing part of the universe by a model of similar but simpler structure (55). Given the complexity of any phenomenon, every model is an abstraction of its target phenomena (43). It could be argued that abstraction is detrimental to model accuracy, and is only necessary so models can be tractable in light of practical and cognitive limitations (i.e. that "The best material model for a cat is another, or preferably the same cat", (55)). However, we argue that the role of abstraction in scientific practice extends beyond addressing those limitations, and that its importance is often underestimated. Moreover, we argue that the appropriate abstractions to make when building a model depend on the problem to be solved, and that the use of similar abstractions for multiple problems of interest results in the development of commonly-used descriptions of phenomena that differ in their relative degree of abstraction (i.e. levels of abstraction).

Indeed, classic accounts of neuroscientific practice emphasize analysis at distinct levels of abstraction (64, 119–122). However, despite the ubiquity of level-based views of neuroscience and a number of proposed schemes, no consensus can be found on what the relevant levels of abstraction are, or even what defines a level (123). A particularly concrete illustrative example of levels of abstraction comes from computer science (124, 125), in which higher level languages abstract the details specified in lower level languages by concealing detailed code in a single function that provides the same relationship. Computational abstraction simplifies a process, such that it is independent of its component processes or even its physical substrate. For example, there are many algorithms that sort a list of numbers, but any computational sort command produces the same result regardless of the algorithm used. Computational abstraction is used in neuroscience, for example, when we simplify the molecular process of synaptic transmission in a more abstract model that represents its net effect as an increased firing rate of a postsynaptic neuron. This simplification is akin to conceptual abstraction (126), by which more abstract, or idealized, models aim to capture general properties of a process rather than the specific details of any one event or dataset. Distinct levels of abstraction also arise in neuroscience when considering problems at different spatiotemporal scales (119). For example, we might consider synaptic transmission in terms of the interactions of various proteins at nanometer to micrometer scales, or we might consider a model at a higher level of abstraction in which neural activity is propagated across the cortex at scales of millimeters or centimeters. When we model phenomena at a given spatiotemporal scale, we make an abstraction that prioritizes organizational details at that scale (e.g.,

cellular), while further simplifying details at others (e.g., subcellular and network) (127). Given the needs of different problems and the range of possible abstractions one might make to solve them, it would be unreasonable to expect a single linear hierarchy of levels. However, we find that different forms of abstraction (i.e. computational or spatiotemporal) are related when we consider their treatment by descriptive, mechanistic, and normative models.

Mechanistic and normative models connect levels of abstraction defined by descriptive models

One promising perspective on the emergence of spatiotemporal levels suggests that models at higher levels of abstraction arise from their lower level counterparts via a natural dimensionality reduction of the parameter space (128, 129). Such a reduction is possible because models of complex systems are "sloppy": they have a large number of dimensions in parameter space along which model parameters can vary without affecting relevant macroscopic observables (i.e. the microscopic parameters are degenerate with respect to macroscopic behavior (130); for examples in neuroscience, see e.g., (131, 132). Thus, abstraction from lower to higher spatio-temporal scales can be seen as a reduction of the lower level parameter space that removes sloppy dimensions, but preserves "stiff" dimensions that have strong influence on observable properties at the higher level. The appropriate dimensionality reduction could be as simple as taking the mean or asymptote of some parameter over a population (133–135), or the set of microscopic parameters needed to produce the same macroscopic behavior might be nonlinear and complex (129, 131, 136, 137). The relationship between phenomena at the two levels can often be expressed in terms of a mechanistic model of a process by which higher level properties emerge from complex interactions of parts described by lower-level parameters. For example, we might abstract single-neuron activity in terms of membrane currents, or by listing the spike times, a natural reduction in the dimensionality, which may result from many combinations of currents . A mechanistic model (e.g. Hodgkin/Huxley (2)) that explains how spike times emerge from currents creates a connection between the abstractions made at the two different levels, and, in addition, as it does not claim mechanisms for how currents emerge from spike times, creates a stratification of higher and lower levels.

In neuroscience, computational abstraction is often discussed in terms of David Marr's three levels of analysis (120, 138): the implementational level is a low-level, concrete statement of a phenomenon, the algorithmic level is an abstraction of the implementational level, explaining the process by which the phenomenon occurs, and the computational level is a high-level (normative) statement of the goal of the process. Thus, **normative models** naturally arise from computational abstraction in that they often begin from the goal of the system at the computational level and link down to implementational or algorithmic levels by showing the optimal solution to achieve that goal, given the constraints at those lower levels. However, this downward linking extends beyond computational abstraction, as we need not limit our statement of goal to computational levels of abstraction. For example, the goal of thermostatic neurons in the mammalian hypothalamus is to maintain a constant body temperature (139, 140). While a mechanistic model could describe the process through, for example, a negative feedback loop, the normative theory says that the goal is the constancy of the temperature. Similar to their role in connecting levels of spatiotemporal abstraction, mechanistic models can connect phenomena at lower levels to their counterparts at higher levels of computational abstraction when they explain the process by which a computation is performed.

Mechanistic and normative models cannot work with pure phenomena; they must work with a description of the parts, goals, and constraints. Thus, if we admit that descriptive models might describe theoretical, as well as observed, phenomena, we can consider the model terms at each level of abstraction to be a **descriptive model**. That is, the components of mechanistic models and the constraints of normative models each correspond to descriptive models at a lower level of abstraction,

while the emergent properties of mechanistic models and the goals of normative models each correspond to descriptive models at a higher level of abstraction.

Thus, we find that our categorization of theories falls naturally into different roles a theory can play in terms of levels of abstraction. **Descriptive theories** define abstractions at different levels. **Mechanistic** and **normative theories** cross levels of abstraction by connecting from description at a "source" level to description at a higher or lower "target" level (Figure). While many philosophies of neuroscience have emphasized the role of multi-level mechanistic explanation (61, 64, 141), little treatment has been given to that of normative explanation, even though it is common in recent neuroscientific practice (e.g., (96, 102, 103, 110, 142)). Phenomena in neuroscience are not simply mechanisms, they are mechanisms that perform functions. Normative explanations view neuroscience phenomena from the viewpoint of those functions, which may be at a range of levels, from cellular goals to behavioral or computational goals. Furthermore, descriptive models, rather than being a "mere" description of phenomena, are the foundation of normative and mechanistic explanations. Given their multi-level nature, a dialogue between descriptive, normative, and mechanistic models linking across levels of analysis is needed for a theoretical account of any neuroscientific phenomenon.

At what level of abstraction should a model be built?

As all models are abstractions, developing a model-based solution to a problem requires selecting a level of abstraction. Such a selection is rarely straightforward, but depends on the needs of the task at hand. Why might we pick one level over another for a given problem? To answer this question we further consider what is meant by level of abstraction, and from a pragmatic view, what levels of abstraction are good for.

First, by restricting our consideration to a given level, abstraction allows us to ignore aspects at other levels. Due to practical and cognitive limitations, this information reduction is imperative to explain any phenomenon in a useful manner. In current neuroscientific practices, there are two general approaches for selecting the appropriate level of abstraction, which serve different purposes. The first approach is to choose the lowest possible level that includes experimentally-supported details while still accounting for the phenomenon. This approach provides many details that can be matched to observable features of a phenomenon. However, it requires extensive calibration from data to ensure the model is accurate, and can be very sensitive to missing, degenerate, or improperly tuned parameters. The second approach is to choose the highest level of abstraction that can account for the phenomenon. Models built at higher levels of abstraction provide conceptual benefits in that they reduce a complicated system to a small number of effective parameters, which can then be used for powerful analysis on the influences to the systems properties, and to build intuition for how the system works. Highly abstract models are especially useful for generalization when similar abstractions can be made for different systems. Generalization can be useful to explain many different phenomena beyond the specifics of any one event and provides the basis for our understanding of broad classes of phenomena (143, 144).

Second, different levels may match with different experimental modalities. Every measurement is in fact an abstraction, in that it is a reduced description of the part of the universe corresponding to the measurement (33). The abstraction made by one measurement device might lend itself to explanations at a given level, but not others. Imagine, for example, trying to solve a problem (such as identifying the sources of social preference) that involves collective properties of neurons across the brain only with data from single-neuron recordings. The problem with such an approach is not just that the sheer amount of data to be analyzed is daunting, but that going from single-neuron data to cognitive functions requires crossing multiple levels of abstraction.

Third, distinct levels promote the development of distinct scientific communities or fields. While we shouldn't equate levels with scientific fields (because fields can organize around a range of shared explanatory goals, concepts, vocabularies, or techniques and methods, and can span multiple levels), different levels are often studied by distinct scientific communities. While a given level may not seem ideal to the problem at hand, the existence of a literature with a rich body of relevant work may influence the use of a model at that level, rather than reinventing the wheel at a new level of abstraction. Recent work using modeling approaches to study scientific organization indicates that having distributed communication networks might actually be beneficial for scientific progress (145, 146), and can allow for simultaneous development of diverse approaches and problem-solving tools. However, selecting or crossing levels can be a sociological problem as well as a methodological one because different fields of study often use different languages and operate under different theoretical frameworks that need to be navigated between.

Finally, the structure of nature itself might group entities into distinct levels (127). We emphasize that the levels we describe are levels of description, and need not correspond to any discretization in nature. However, an interesting scientific question is when, why, and how distinct levels might appear in natural phenomena (e.g., with the emergence of patterns with characteristic spatiotemporal scales, the emergence of computational systems, or the emergence of systems with goals). In these cases, discrete levels may reflect local maxima in the degree of regularity of entities at specific spatiotemporal scales (119, 147, 148) or hierarchical structures in causal organization (64), in which case identifying abstractions that correspond to those levels is akin to "carving nature at its joints".

In conclusion, it is very important that researchers spell out the abstraction being made, including both its purpose and what its limitations are. Given that every model is an abstraction, it is important for all models to do this. While these descriptions are often provided for highly abstract models, researchers working at models of lower spatio-temporal scales (such as detailed compartmental models of neurons) often claim to be building biologically realistic models. Our assertion is that these models are also abstractions, albeit at a different level, and a proper description of the abstractions made will help clarify both the uses and the limitations of the model.

Theory and experiment

We might consider that the overarching goal of science is to produce theories that are **precise** (i.e. make specific predictions), **general** (for a wide domain of phenomena), and **accurate** (that align closely to data). Once established, such theories can then be used to solve empirical problems as they arise and to direct scientific and engineering efforts towards the discovery, prediction, and eventual control of further phenomena. Achieving these epistemic virtues cannot be done with theoretical work alone, but relies on an interaction between theory and experiment. Traditional views emphasize the role of experiments to test theories (7), and even consider an interplay in which theories suggest new experiments and experiments require new theories to explain their results (10, 18). However, theories are not born fully-formed, but are developed over time. Such development generally does not happen independently of experiment, but often goes hand-in-hand. Unlike the Popperian propose-and-reject philosophy (7), in practice, theories change over time through their interaction with experiment, and we cannot understand the nature of theories without understanding this process of theory development (10, 18, 19, 149). Models play a key role in this process, and can be used as experiments even in the absence of data (43, 150). Together, this reveals a picture in which theoretical research is not relegated to simply proposing theories-to-be-tested, but instead plays an active role in the simultaneous development and assessment of theories.

Linking Theory and Phenomena

We define the **domain** of a theory as the set of phenomena that a theory sets out to explain. By being explicit about the intended domain of a theory, we provide an experimental range for the theory to be applied to. For example, the theory that action potentials arise from changes in ion-flow due to voltage-dependent changes in permeability (2, 45, 151, 152) should apply to the domain of all action potentials in all neurons. But we should not expect action potentials in all neurons to be driven by the same ion channels. Similarly, the theory of weakly coupled oscillators (153) can be applied to neural oscillators that interact through weak interactions, but not those that strongly reset when they interact. We can think of the domain as a set of data-imposed constraints on the theory, and that a good theory should provide the (minimal) set of explanations that satisfy the constraints. One can, of course, attempt to apply a theory outside of its original intended domain, and successfully doing so may reflect a serious development of the theory, but we argue that theoretical papers should be explicit about what phenomena do and do not lie in the intended domain. Further, we should not strictly judge nascent theories by their ability to explain all of the data in their proposed domain (12, 18) but should take into account an assessment of their ability to do so with further development (9). A good theory should be able to, for example, become more specific as more data become available.

How do we assess a theory's ability to explain phenomena in its domain? The strength of model-based explanation lies in our ability to directly compare parts of a model to experimental data, which lets us connect the theory that the model instantiates with phenomena they are proposed to explain. However, no model is directly comparable to experimental data by virtue of its structure alone. Such a comparison requires a **translation function**: a statement of how the model maps onto its target phenomena. By specifying the intended correspondence between model terms and phenomena, the translation function operationalizes the concepts associated with those terms in the theory (154, 155). The translation function can itself be a separate testable component of the model, similar to the "linking hypothesis" used to link parts of cognitive models to experimentally observed quantities (156). Translation functions are necessary to make experimental predictions from theories, and should be provided as part of the model definition. The closer experimentally-observable phenomena are to identifiable model components, the simpler the translation function can be. In some cases the translation function might be as simple as "variable V represents the membrane potential in mV", but it can also be that "variable V qualitatively corresponds to the slow changes in the membrane potential" and ignores, for example, all spiking activity. In other cases the translation function can be more complex, as parts of the model can have a loose correspondence to general features of large classes of data, and can represent highly abstract effective parameters or qualitative behaviors. For example, the units in Hopfield's attractor network models (157–159) are not meant to directly correspond to measurable properties of biological neurons, but are instead intended to reflect qualitative features of neural activity -- namely that neural populations are "active" or not.

Given a translation function, we can imagine a number of strategies for connecting theory to experiment. The simplest way to connect to data is via descriptive models - their variables correspond to observables and parameters can be fit to approximate the relationships between those observables. Mechanistic and normative models can connect to data by virtue of descriptive models at either the target or source level of abstraction. An ultimate test of a mechanistic or normative explanation is to produce a descriptive model at the target level of abstraction that is comparable to measurements. We can then say that the assumptions in the model (i.e. the theory) can account for the phenomenon. In the case of mechanistic models, their strength lies in the ability to combine many descriptive models at the source level (each of which may be fit to data) into a single unified mechanism. However, the vast majority of mechanistic models in neuroscience are too complicated to derive an analytical relationship between lower level parameters and a descriptive model at the target level of abstraction (i.e. in the

resulting dynamics). For example, although there are numerous models of rhythm generation in the medulla that behave akin to the inspiratory phase of mammalian respiration (160, 161), there is still no consensus model that parameterizes properties of the respiration rhythm (e.g., frequency) in terms that correspond to specific properties of medullary neurons (e.g., ion channel composition; (162, 163). To overcome such limitations, researchers often skip creating a descriptive model and match features of the simulated dynamics of mechanistic models to desired features of the data directly (164–166).

Modeling Experiments

Like their physical counterparts, mathematical and computational models can themselves be used in a form of experiment to test the viability of a theory (53, 58, 59, 167). They instantiate the theory in foundational assumptions that act as the hypothesis of the experiment and test if those assumptions are sufficient to account for data from the phenomena. If so, the modeling experiment can be seen to support the theory. For example, based on a theory, a researcher may hypothesize that some observations are relevant to system function, build a mechanistic model corresponding to the proposed relevant parts and their interactions to test if they're sufficient to reproduce features of the data. Alternatively, a researcher can hypothesize that the system is performing some function, make a normative model, and see if data from the system behaves as if the system is optimizing that function. If not, this suggests the need to look for missing constraints on the system or other functions the system is trying to solve. In each case, if the assumptions are unable to account for the data, such a modeling experiment can bring into question the viability of a theory, including its model instantiation or the translation function. One can imagine treating different parameters or model instantiations as independent variables in the experiment, and testing their sufficiency to achieve different aspects of the phenomenon (the dependent variables) (168, 169).

Such modeling experiments can be carried out even in the absence of data, as phenomena at both the target and source levels of abstraction can be pure theoretical entities. One can test the feasibility of theoretical claims by studying models that instantiate those theories in tractable idealized systems. For example, Hopfield's attractor network models (157, 158) provided strong support for Hebb's theory (86) that co-active firing of neurons leading to increased connectivity would create associative memory, by showing that strong connections between simple neuron-like entities were sufficient to produce cell assemblies that could be accessed through a pattern-completion process (159).

Thus, modeling experiments can be used to apply existing theories to account for observed phenomena, compare possible predictions within a theory, or even to compare theories with overlapping domains to see which does better. Such uses are analogous to confirmatory (hypothesis-testing) experiments. However, the value of modeling experiments extends beyond confirmatory research. Like their physical analogues (e.g. the 6-OHDA rat or the MPTP monkey), models are analogous processes that can be used for exploratory (hypothesis-generating) research: building an interactive set of sub-phenomena to observe what sorts of phenomena *might* emerge from it (just the expected ones, or maybe also unpredicted phenomena?), and can thus be used to explore the implications of the theory or extend its scope.

Exploratory modeling experiments produce new observations that can provide new hypotheses to incorporate into theories and to design physical experiments, for example, instantiating theories in not-yet-observed systems can be used to predict novel phenomena. The Hopfield model and others like it (157, 158, 170, 171) led to new experimental predictions that could be tested, including psychological-level cognitive science categorization experiments (172, 173), neural-level long-term changes in tuning curve experiments (174–181), and direct observations of pattern completion processes (182–186). Furthermore, exploratory modeling experiments can instantiate idealized aspects

of a theory to help build intuition for the theory itself. Hopfield's model and its subsequent derivatives have provided researchers with a deeper understanding of important properties of how memories can be accessed by content through pattern-completion processes and concepts such as "basins of attraction" - the set of patterns that will resolve to the same final pattern (157, 159). These computational discoveries can help build understanding of the theory, and lead to predictions and ideas for new experiments, which will lead back to new observations that need to be incorporated into the theories.

The utility of modeling experiments extends beyond testing and exploring the implications of existing theories; modeling experiments are extremely useful to the theorist in the context of theory development (187). When an explanation for a phenomenon is not readily available in an existing theory, or if the available explanations are unclear or conflicting, assumptions can be hypothesized that form the basis of a modeling experiment, from which the behavior of the model can reveal the sufficiency (or insufficiency) of the assumptions to account for the phenomenon. Often these modeling experiments precede a well-formed theory, and a theorist will perform numerous experiments with different models in the process of developing a theory (188). Over time, specific model formulations can become closely associated with the theory (become canonical instantiations of the theory), making the theory itself more readily applicable to specific problems and more precise in its proposed solutions.

From this perspective, an established theory can be considered a body of assumptions, which have been tested through "modeling experiments" in which those assumptions were found to be sufficient to account for some aspect of a phenomenon. A theory in this sense is not a formal set of laws, but a continuously developing body of canonical models and model-phenomenon correspondences, bound together partly by history and partly by shared problem-solving methods and standards (189).

Towards Future Frameworks in Theoretical Neuroscience

A scientific theory is a thinking tool: a set of ideas used to solve specific problems. As suggested in this manuscript, we can think of theoretical neuroscience as a field which approaches problems in neuroscience with the following problem-solving methodology: **theories** are instantiated in **models** which, by virtue of a **translation function**, can be used to assess a theory's ability to account for phenomena in the theory's **domain** or explore its further implications.

We identified three kinds of theoretical constructs that play distinct roles in this process: **descriptive theories and models**, which define the abstractions by which we describe a phenomenon; **mechanistic theories and models** which explain phenomena at higher levels of abstraction in terms of lower level parts and their interactions; and **normative theories and models**, which explain phenomena at lower levels of abstraction in terms of a higher level function or goal.

A conceptual or theoretical **framework** provides a language within which specific theories abide. The stability of an overarching framework allows theories to develop and change without rebuilding their conceptual foundations. For example, early theories of action potential function identified voltage-gated sodium channels as the primary depolarizing component (2), but when it was found that some cells showed action potentials that were not related to sodium concentrations (such as Purkinje cell complex spikes), it was easy to add the effects of other voltage-gated channels within the same conceptual framework and theoretical language (45, 53, 58). Such change is inevitable in the life of a theory. As theories become more strongly corroborated and more precise, they become better for solving empirical problems, provide more reliable and more accurate predictions, and can be applied more generally for larger domains. Over time and through the development of canonical model formulations, theories

become more rigorous, such that researchers agree on how they should be implemented to explain specific domains.

Frameworks themselves change as well. For example, as noted earlier, under the Freudian framework in psychiatry in the early 20th century explanations of psychiatric disorder are found in theories involving subconscious desires due to developmental relationships with one's parents, and would need to be treated with talk therapy. The medicalized framework that emerged in the late 20th century framed explanations of psychiatric disorders as changes in brain function to be treated with physical (e.g., pharmacological or electrical) manipulations of brain function. Even the categorizations of psychiatric phenomena are different under these paradigms, making theoretical comparisons difficult. Thus, changes in a dominant framework are proposed to be more dramatic and may be akin to paradigm shifts (8) and directly comparing explanations for the same phenomena across frameworks may be difficult or even impossible (12)[4]. However, this does not mean that all frameworks are equivalent. If theories within existing frameworks repeatedly fail to produce adequate solutions to new problems or to predict new phenomena in their domain, it may indicate that the conceptual foundation provided by the framework is inadequate. However, such shifts rarely happen without alternative competing frameworks that show promise in explaining phenomena in overlapping domains (9). In the last decade, a new framework known as computational psychiatry has emerged in which psychiatric disorders are identified as computational "failure modes" in the systems architecture of the brain (190–194). Under this framework, explanations for such disorders are to be found in theories that involve changes in information processing, and would be potentially treatable by changing that information processing, e.g., by changing the physical substrate (e.g. through electrical stimulation), by encouraging compensation processes (e.g. through cognitive training), or changes in the environment (e.g. by giving a student with ADHD extra time on a test). Interestingly, computational psychiatry emerged by applying conceptual frameworks of reliability engineering to computational neuroscience, suggesting that framework shifts may often arise from the translation of existing theories applied to new domains.

So, are new theoretical frameworks needed for progress in neuroscience? A number of recent proposals have suggested framework development in light of recent progress, including, for example, progress in deep learning (195), behavior (196, 197), and neural coding (198) or the combination of the dynamic and statistical languages (199). As these works have already led to extensive discussion within the community (e.g. in commentaries, conferences, and online communities), we conclude that development of the content-related constraints of our current framework is needed, but that such development is already under way. We have instead chosen to focus on the general problem-solving methodology and explanatory strategies of the current framework. It is our view that closer attention should be paid to these strategies, and doing so will help guide development of the necessary future frameworks of neuroscience. One way to implement these efforts is to better specify the deliverables of theory projects, both to funding agencies in grant proposals and to our experimental colleagues, as we discuss in the following section.

Deliverables

A major purpose of the San Antonio meeting was to help NSF, NIH and other funding agencies determine what to expect from grant proposals that have a theoretical component. The typically constructed grant that proposes to perform a traditional experiment is well-designed for experimental research, but less well-designed for theoretical research. We hope that the formulation of theoretical

---

[4] As pointed out by (18) and (19), because science includes both theoretical abstractions and applied practical components, one can compare across conceptual frameworks by asking how well they allow us to control our environment, i.e. by comparing their epistemic virtues.

neuroscience put forward in this manuscript can provide guidance for what constitutes a good theory project. We summarize our assertions in the following points.

Be specific. A theory should be specific, not necessarily in terms of what the theory is, but rather what the theory is attempting to explain and the strategies for doing so. In particular the theory should define what problems it is trying to solve, and provide the criteria for an adequate solution. It is important to define the *descriptive*, *mechanistic*, and *normative* components of the theory as well as the levels of abstraction and the rationale behind their selection.

Identify the domain and translation function. The utility of a theory is not just in the equations, but in describing the relationship of its components to component parts of the phenomena within its domain. Be clear about how the theory relates to experimental observations, which data it explains and what experimental predictions it makes.

Define which aspects of the research are exploratory and which are confirmatory. The fact that models are a form of experiment creates a way forward for theoretical grant proposals. For example, a researcher can propose to build a model that crosses levels in order to address the question of theoretical viability. Such a proposal may have preliminary data to show that one can build models at each level, even if the researcher has not yet put those levels together. Similarly, a grant proposal can define the domain even if the literature review is incomplete. One can also identify that one is going to explore the parameter space of a set of models to determine how those parameters affect phenomena across levels.

By being explicit about the scientific question being addressed, about the assumptions of the theory, the domain the theory is purporting to address, and the process of building and testing models underlying that theory, grant proposals could be viable even if the theory itself remains incomplete. We call on funding agencies and reviewers to recognize that theory is the foundation of any science, and that construction of rigorous theory and systematic computational modeling are time-consuming processes that require dedicated personnel with extensive training. Our hope is that the framework and associated language outlined in this document can be used to specify deliverables that can be understood by both funders and investigators.

Finally, it is interesting to consider that we might apply our taxonomy to our own metatheoretical framework. The concept that the ultimate goal of a theory is to provide tools that allow one to better explain and control one's environment is a normative theory of the goal of scientific theories; the concept that models instantiate theories and allow one to test their viability and their relationship to phenomena is a mechanistic theory of how those theories achieve that goal; and the concept that theories live within a framework that a community applies to them is a descriptive theory of theories. One could imagine a metascientific research program which studies the available phenomena - for example, the scientific literature - to test and further develop those theories, and even the use of models of the scientific process (e.g. (200)). The benefits of such a research program extend beyond satisfying an esoteric interest in scientific methodology, but could prove as impactful for scientific practice as other theories have proven for manipulation of phenomena in their domain.

Acknowledgments

References

1. D. Marr, *From the Retina to the Neocortex: Selected Papers of David Marr* (Edited by L. M. Vaina. Birkhäuser, 1991).

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

3. J. O'Keefe, L. Nadel, *The Hippocampus as a Cognitive Map* (Clarendon Press, 1978).

4. R. E. Goldstein, Are theoretical results "Results"? *Elife* **7** (2018).

5. W. Bialek, Perspectives on theory at the interface of physics and biology. *Rep. Prog. Phys.* **81**, 012601 (2018).

6. R. Phillips, Theory in Biology: Figure 1 or Figure 7? *Trends Cell Biol.* **25**, 723–729 (2015).

7. K. R. Popper, The logic of scientific discovery New York. *Science* (1959).

8. T. S. Kuhn, *The Structure of Scientific Revolutions: 50th Anniversary Edition* (University of Chicago Press, 2012).

9. I. Lakatos, Science and pseudoscience. *Philosophical papers* **1**, 1–7 (1978).

10. S. Firestein, *Failure: Why Science Is So Successful* (Oxford University Press, 2015).

11. P. Godfrey-Smith, An introduction to the philosophy of science: Theory and reality (2003).

12. P. Feyerabend, *Against Method* (Verso, 1993).

13. S. Firestein, *Ignorance: How It Drives Science* (Oxford University Press, USA, 2012).

14. M. Ben-Ari, *Just A Theory: Exploring The Nature Of Science* (Prometheus Books, 2011).

15. B. C. van Fraassen, *The Scientific Image* (Clarendon Press, 1980).

16. N. David Mermin, What's Wrong with this Pillow? *Phys. Today* **42**, 9–11 (1989).

17. D. Kaiser, History: Shut up and calculate! *Nature* **505**, 153–155 (2014).

18. L. Laudan, *Progress and Its Problems: Towards a Theory of Scientific Growth* (University of California Press, 1978).

19. H. Douglas, Pure science and the problem of progress. *Stud. Hist. Philos. Sci.* **46**, 55–63 (2014).

20. A. Franklin, Forging, cooking, trimming, and riding on the bandwagon. *Am. J. Phys.* **52**, 786–793 (1984).

21. M. Jeng, A selected history of expectation bias in physics. *Am. J. Phys.* **74**, 578–583 (2006).

22. A. D. Redish, E. Kummerfeld, R. L. Morris, A. C. Love, Opinion: Reproducibility failures are essential to scientific inquiry. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 5042–5046 (2018).

23. A. Newell, H. A. Simon, *Human problem solving* (Prentice-Hall, 1972).

24. W. Bechtel, R. C. Richardson, *Discovering Complexity: Decomposition and Localization as Strategies in Scientific Research* (MIT Press, 2010).

25. W. B. Scoville, B. Milner, Loss of recent memory after bilateral hippocampal lesions. *J. Neurol. Neurosurg. Psychiatry* **20**, 11–21 (1957).

26. L. R. Squire, *Memory and Brain* (Oxford University Press, 1987).

27. L. Nadel, "Multiple memory systems: What and Why, an update" in *Memory Systems 1994*, D. L. Schacter, E. Tulving, Eds. (MIT Press, 1994), pp. 39–64.

28. D. L. Schacter, *The Seven Sins of Memory* (Houghton Mifflin, 2001).

29. B. W. Balleine, A. Dickinson, Goal-directed instrumental action: contingency and incentive learning and their cortical substrates. *Neuropharmacology* **37**, 407–419 (1998).

30. N. D. Daw, Y. Niv, P. Dayan, Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nat. Neurosci.* **8**, 1704–1711 (2005).

31. A. D. Redish, *The Mind within the Brain: How we make decisions and how those decisions go wrong* (Oxford, 2013).

32. R. Descartes, *Discours de la Méthode Pour bien conduire sa raison, et chercher la vérité dans les sciences* (1637).

33. H. Chang, Scientific Progress: Beyond Foundationalism and Coherentism 1. *Royal Institute of Philosophy Supplements* **61**, 1–20 (2007).

34. H. Chang, The Persistence of Epistemic Objects Through Scientific Change. *Erkenntnis* **75**, 413–429 (2011).

35. P. Luyten, L. C. Mayes, P. Fonagy, M. Target, S. J. Blatt, *Handbook of Psychodynamic Approaches to Psychopathology* (Guilford Publications, 2015).

36. ,ICD-11 (January 16, 2020).

37. American Psychiatric Association, *Diagnostic and Statistical Manual of Mental Disorders (DSM-5®)* (American Psychiatric Pub, 2013).

38. B. N. Cuthbert, T. R. Insel, Toward the future of psychiatric diagnosis: the seven pillars of RDoC. *BMC Med.* **11**, 126 (2013).

39. T. R. Insel, B. N. Cuthbert, Medicine. Brain disorders? Precisely. *Science* **348**, 499–500 (2015).

40. C. G. Hempel, P. Oppenheim, Studies in the Logic of Explanation. *Philos. Sci.* **15**, 135–175 (1948).

41. J. Woodward, "Scientific Explanation" in *The Stanford Encyclopedia of Philosophy*, E. N. Zalta, Ed. (2019).

42. R. G. Winther, The Structure of Scientific Theories. *The Stanford Encyclopedia of Philosophy* (2016).

43. M. Weisberg, *Simulation and Similarity: Using Models to Understand the World* (OUP USA, 2013).

44. R. Frigg, S. Hartmann, Models in science (2006).

45. B. Hille, *Ion Channels of Excitable Membranes* (Sinauer, 2001).

46. E. Marder, Models identify hidden assumptions. *Nature Neuroscience* **3**, 1198–1198 (2000).

47. J. M. Epstein, Why model? *Journal of Artificial Societies and Social Simulation* **11**, 12 (2008).

48. J. D. Watson, F. H. Crick, Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* **171**, 737–738 (1953).

49. U. Alon, *An introduction to systems biology: design principles of biological circuits* (Chapman and Hall/CRC, 2006).

50. A. D. Dorval, W. M. Grill, Deep brain stimulation of the subthalamic nucleus reestablishes neuronal information transmission in the 6-OHDA rat model of parkinsonism. *J. Neurophysiol.* **111**, 1949–1959 (2014).

51. W. Schultz, *et al.*, Deficits in reaction times and movement times as correlates of hypokinesia in monkeys with MPTP-induced striatal dopamine depletion. *J. Neurophysiol.* **61**, 651–668 (1989).

52. W. Rall, "Cable theory for dendritic neurons" in *Methods in Neuronal Modeling*, C. Koch, I. Segev, Eds. (MIT Press, 1992), pp. 9–62.

53. W. Gerstner, W. Kistler, *Spiking Neuron Models* (Cambridge University Press, 2002).

54. J. W. Langston, J. Palfreman, *The Case of the Frozen Addicts: How the Solution of a Medical Mystery Revolutionized the Understanding of Parkinson's Disease* (IOS Press, 2013).

55. A. Rosenblueth, N. Wiener, The Role of Models in Science. *Philos. Sci.* **12**, 316–321 (1945).

56. T. Stafford, What use are computational models of cognitive processes? in *Connectionist Models of Behaviour and Cognition II: Proceedings of the 11th Neural Computation and Psychology Workshop*, Mayor, J., Ruh, N., Plunkett, K, Ed. (World Scientific., 2009).

57. N. Cartwright, Models: The Blueprints for Laws. *Philos. Sci.* **64**, S292–S303 (1997).

58. C. Koch, I. Segev, Eds., *Methods in Neuronal Modeling* (MIT Press, 1989).

59. W. Gerstner, H. Sprekeler, G. Deco, Theory and simulation in neuroscience. *Science* **338**, 60–65 (2012).

60. P. Dayan, L. F. Abbott, *Theoretical Neuroscience* (MIT Press, 2001).

61. D. M. Kaplan, W. Bechtel, Dynamical models: an alternative or complement to mechanistic explanations? *Top. Cogn. Sci.* **3**, 438–444 (2011).

62. A. D. Redish, *Beyond the Cognitive Map: From Place Cells to Episodic Memory* (MIT Press, 1999).

63. L. L. Colgin, Five Decades of Hippocampal Place Cells and EEG Rhythms in Behaving Rats. *The Journal of Neuroscience* **40**, 54–60 (2020).

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

65. D. M. Kaplan, Explanation and description in computational neuroscience. *Synthese* **183**, 339 (2011).

66. C. Linnaeus, *Systema Naturae* (1758).

67. S. R. y. Cajal, N. Swanson, L. W. Swanson, *Histology of the nervous system of man and vertebrates* (Oxford University Press, 1995).

68. D. Salsburg, *The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century* (Macmillan, 2001).

69. R. E. Kass, U. T. Eden, E. N. Brown, *Analysis of Neural Data* (Springer, 2014).

70. M. T. Harrison, A. Amarasingham, R. E. Kass, "Statistical Identification of Synchronous Spiking" in *Spike Timing: Mechanisms and Function*, P. M. DiLorenzo, J. D. Victor, Eds. (CRC Press, 2013), p. 77.

71. A. Amarasingham, S. Geman, M. T. Harrison, Ambiguity and nonidentifiability in the statistical analysis of neural codes. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 6455–6460 (2015).

72. M. Baker, 1,500 scientists lift the lid on reproducibility. *Nature* **533**, 452–454 (2016).

73. S. N. Goodman, D. Fanelli, J. P. A. Ioannidis, What does research reproducibility mean? *Sci. Transl. Med.* **8**, 341ps12 (2016).

74. D. Fanelli, Opinion: Is science really facing a reproducibility crisis, and do we need it to? *Proc. Natl. Acad. Sci. U. S. A.* **115**, 2628–2631 (2018).

75. National Academies of Sciences Engineering, Medicine, *Reproducibility and Replicability in Science* (The National Academies Press, 2019).

76. drugmonkey, Generalization, not "reproducibility." *Drugmonkey* (2018) (January 5, 2020).

77. P. Smaldino, Better methods can't make up for mediocre theory. *Nature* **575**, 9 (2019).

78. P. Machamer, L. Darden, C. F. Craver, Thinking about Mechanisms. *Philos. Sci.* **67**, 1–25 (2000).

79. A. D. Redish, R. Kazinka, A. B. Herman, Taking an engineer's view: Implications of network analysis for computational psychiatry. *Behav. Brain Sci.* **42**, e24 (2019).

80. L. Elefteriadou, *An Introduction to Traffic Flow Theory* (Springer, New York, NY, 2014).

81. M. S. Gazzaniga, Whos in charge. *Free will and the science of the brain. New York: Ecco* (2011).

82. S. P. Ellner, J. Guckenheimer, *Dynamic Models in Biology* (Princeton University Press, 2006).

83. E. M. Izhikevich, *Dynamical Systems in Neuroscience* (MIT Press, 2007).

84. G. Bard Ermentrout, D. H. Terman, *Mathematical Foundations of Neuroscience* (Springer Science & Business Media, 2010).

85. F. Gabbiani, S. J. Cox, *Mathematics for Neuroscientists* (Academic Press, 2017).

86. D. O. Hebb, *The Organization of Behavior* (Wiley, 1949).

87. J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference* (Morgan Kaufmann, 1988).

88. J. Pearl, *Causality: Models, Reasoning and Inference* (Cambridge University Press, 2009).

89. G. W. Imbens, D. B. Rubin, *Causal Inference in Statistics, Social, and Biomedical Sciences* (Cambridge University Press, 2015).

90. J. Pearl, Causal inference in statistics: An overview. *Stat. Surv.* **3**, 96–146 (2009).

91. S. Ma, P. Kemmeren, C. F. Aliferis, A. Statnikov, An Evaluation of Active Learning Causal Discovery Methods for Reverse-Engineering Local Causal Pathways of Gene Regulation. *Sci. Rep.* **6**, 22558 (2016).

92. A. V. Alekseyenko, *et al.*, Causal graph-based analysis of genome-wide association data in rheumatoid arthritis. *Biol. Direct* **6**, 25 (2011).

93. S. Mani, C. Aliferis, S. Krishnaswami, T. Kotchen, Learning causal and predictive clinical practice guidelines from data. *Stud. Health Technol. Inform.* **129**, 850–854 (2007).

94. K. J. Friston, J. Kahan, B. Biswal, A. Razi, A DCM for resting state fMRI. *Neuroimage* **94**, 396–407 (2014).

95. K. E. Stephan, *et al.*, Nonlinear dynamic causal models for fMRI. *Neuroimage* **42**, 649–662 (2008).

96. H. B. Barlow, Others, Possible principles underlying the transformation of sensory messages. *Sensory communication* **1**, 217–234 (1961).

97. K. P. Kording, J. B. Tenenbaum, R. Shadmehr, The dynamics of memory as a consequence of optimal adaptation to a changing body. *Nat. Neurosci.* **10**, 779–786 (2007).

98. W. Bialek, *Biophysics: Searching for Principles* (Princeton University Press, 2012).

99. W. Bialek, S. Setayeshgar, Cooperativity, sensitivity, and noise in biochemical signaling. *Phys. Rev. Lett.* **100**, 258101 (2008).

100. G. A. Parker, J. M. Smith, Optimality theory in evolutionary biology. *Nature* **348**, 27–33 (1990).

101. F. Rieke, D. Warland, R. de Ruyter van Steveninck, W. Bialek, *Spikes* (MIT Press, 1997).

102. G. D. Field, F. Rieke, Nonlinear signal transfer from mouse rods to bipolar cells and implications for visual sensitivity. *Neuron* **34**, 773–785 (2002).

103. E. Doi, M. Lewicki, Optimal retinal population coding predicts inhomogeneous light adaptation and contrast sensitivity across the visual field. *Journal of Vision* **14**, 1188–1188 (2014).

104. R. Gregory, P. Cavanagh, The Blind Spot. *Scholarpedia J.* **6**, 9618 (2011).

105. S. J. Gould, *Hen's Teeth and Horse's Toes* (Norton, 1983).

106. L. Valiant, *Probably Approximately Correct: NatureÕs Algorithms for Learning and Prospering in a Complex World* (Basic Books, 2013).

107. A. Wikenheiser, D. W. Stephens, A. D. Redish, Subjective costs drive overly-patient foraging strategies in rats on an intertemporal foraging task. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 8308–8313 (2013).

108. B. M. Sweis, *et al.*, Sensitivity to "sunk costs" in mice, rats, and humans. *Science* **361**, 178–181 (2018).

109. B. Schmidt, A. A. Duin, A. D. Redish, Disrupting the medial prefrontal cortex alters hippocampal sequences during deliberative decision making. *J. Neurophysiol.* **121**, 1981–2000 (2019).

110. B. M. Sweis, M. J. Thomas, A. D. Redish, Mice learn to avoid regret. *PLoS Biol.* **16**, e2005853 (2018).

111. H. A. Simon, Theories of bounded rationality. *Decision and organization* **1**, 161–176 (1972).

112. A. P. Steiner, A. D. Redish, Behavioral and neurophysiological correlates of regret in rat decision-making on a neuroeconomic task. *Nat. Neurosci.* **17**, 995–1002 (2014).

113. D. H. Hubel, T. N. Wiesel, Brain mechanisms of vision. *Sci. Am.* **241**, 150–162 (1979).

114. D. H. Hubel, T. N. Wiesel, Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiol.* **160**, 106–154 (1962).

115. M. Carandini, Area V1. *Scholarpedia J.* **7**, 12105 (2012).

116. L. Paninski, Maximum likelihood estimation of cascade point-process neural encoding models. *Network* **15**, 243–262 (2004).

117. K. W. Latimer, F. Rieke, J. W. Pillow, Inferring synaptic inputs from spikes with a conductance-based neural encoding model. *Elife* **8** (2019).

118. B. A. Olshausen, D. J. Field, Sparse coding with an overcomplete basis set: a strategy employed by V1? *Vision Res.* **37**, 3311–3325 (1997).

119. P. Churchland, T. J. Sejnowski, *The computational Brain* (MIT Press, 1994).

120. D. Marr, *Vision* (W. H. Freeman and Co., 1982).

121. G. M. Shepherd, *Neurobiology* (Oxford University Press, 1994).

122. T. Sejnowski, C. Koch, P. Churchland, Computational neuroscience. *Science* **241**, 1299–1306 (1988).

123.   S. Guttinger, A. C. Love, Characterizing scientific failure: Putting the replication crisis in context. *EMBO Rep.* **20**, e48765 (2019).

124.   T. Colburn, G. Shute, Abstraction in Computer Science. *Minds Mach.* **17**, 169–184 (2007).

125.   J. M. Wing, Computational thinking and thinking about computing. *Philos. Trans. A Math. Phys. Eng. Sci.* **366**, 3717–3725 (2008).

126.   T. O'Leary, A. C. Sutton, E. Marder, Computational models in the age of large datasets. *Current Opinion in Neurobiology* **32**, 87–94 (2015).

127.   M. I. Eronen, D. S. Brooks, Levels of Organization in Biology. *The Stanford Encyclopedia of Philosophy* (2018).

128.   B. B. Machta, R. Chachra, M. K. Transtrum, J. P. Sethna, Parameter space compression underlies emergent theories and predictive models. *Science* **342**, 604–607 (2013).

129.   M. K. Transtrum, *et al.*, Perspective: Sloppiness and emergent theories in physics, biology, and beyond. *J. Chem. Phys.* **143**, 010901 (2015).

130.   R. N. Gutenkunst, *et al.*, Universally sloppy parameter sensitivities in systems biology models. *PLoS Comput. Biol.* **3**, 1871–1878 (2007).

131.   A. A. Prinz, D. Bucher, E. Marder, Similar network activity from disparate circuit parameters. *Nat. Neurosci.* **7**, 1345–1352 (2004).

132.   D. Panas, *et al.*, Sloppiness in spontaneously active neuronal networks. *J. Neurosci.* **35**, 8480–8492 (2015).

133.   H. R. Wilson, J. D. Cowan, A mathematical theory of the functional dynamics of cortical and thalamic tissue. *Kybernetik* **13**, 55–80 (1973).

134.   D. J. Pinto, J. C. Brumberg, D. J. Simons, G. B. Ermentrout, A Quantitative Population Model of Whisker Barrels: Re-examining the Wilson-Cowan Equations. *J. Comput. Neurosci.* **3**, 247–264 (1996).

135.   A. Destexhe, T. J. Sejnowski, The Wilson–Cowan model, 36 years later. *Biol. Cybern.* **101**, 1–2 (2009).

136.   J. Jalics, M. Krupa, H. G. Rotstein, A novel mechanism for mixed-mode oscillations in a neuronal model. *Dynamical Systems: An International Journal iFirst*, 1–38 (2010).

137.   H. G. Rotstein, T. Oppermann, J. A. White, N. Kopell, A reduced model for medial entorhinal cortex stellate cells: subthreshold oscillations, spiking and synchronization. *J. Comput. Neurosci.* **21**, 271–292 (2006).

138.   Z. W. Pylyshyn, *Computation and Cognition: Toward a Foundation for Cognitive Science* (MIT Press, 1984).

139.   C. L. Tan, Z. A. Knight, Regulation of Body Temperature by the Nervous System. *Neuron* **98**, 31–48 (2018).

140.   S. F. Morrison, K. Nakamura, Central neural pathways for thermoregulation. *Front. Biosci.* **16**, 74–104 (2011).

141.   W. Bechtel, Mechanisms in Cognitive Psychology: What Are the Operations? *Philos. Sci.* **75**, 983–994 (2008).

142.   E. Fehr, I. Krajbich, "Social preferences and the brain" in *Neuroeconomics (Second Edition)*, (Elsevier, 2014), pp. 193–218.

143.   M. Gilead, Y. Trope, N. Liberman, Above and Beyond the Concrete: The Diverse Representational Substrates of the Predictive Brain. *Behav. Brain Sci.*, 1–63 (2019).

144.   M. Gilead, N. Liberman, A. Maril, Construing counterfactual worlds: The role of abstraction. *European Journal of Social Psychology* **42**, 391–397 (2012).

145.   L. Wu, D. Wang, J. A. Evans, Large teams develop and small teams disrupt science and technology. *Nature* **566**, 378–382 (2019).

146.   P. Grim, *et al.*, Scientific Networks on Data Landscapes: Question Difficulty, Epistemic Success, and Convergence. *Episteme* **10**, 441–464 (2013).

147.   T. J. Sejnowski, P. S. Churchland, J. A. Movshon, Putting big data to good use in neuroscience. *Nat. Neurosci.* **17**, 1440–1441 (2014).

148.   W. C. Wimsatt, "Reductionism, Levels of Organization, and the Mind-Body Problem" in *Consciousness and the Brain: A Scientific and Philosophical Inquiry*, G. G. Globus, G. Maxwell, I. Savodnik, Eds. (Springer US, 1976), pp. 205–267.

149.   W. Bechtel, *Philosophy of science: An overview for cognitive science* (Psychology Press, 2013).

150.   Y. Dudai, K. Evers, To simulate or not to simulate: what are the questions? *Neuron* **84**, 254–261 (2014).

151.   Y. Goldman, M. Morad, Ionic membrane conductance during the time course of the cardiac action potential. *J. Physiol.*

268, 655–695 (1977).

152.   A. M. Katz, Cardiac ion channels. *N. Engl. J. Med.* **328**, 1244–1251 (1993).

153.   G. B. Ermentrout, N. Kopell, Frequency Plateaus in a Chain of Weakly Coupled Oscillators, I. *SIAM J. Math. Anal.* **15**, 215–237 (1984).

154.   P. W. Bridgman, The logic of modem physics. *New York* (1927).

155.   H. Chang, *Inventing Temperature: Measurement and Scientific Progress* (Oxford University Press, 2007).

156.   G. de Hollander, B. U. Forstmann, S. D. Brown, Different Ways of Linking Behavioral and Neural Data via Computational Cognitive Models. *Biol Psychiatry Cogn Neurosci Neuroimaging* **1**, 101–109 (2016).

157.   J. J. Hopfield, Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl. Acad. Sci. U. S. A.* **79**, 2554–2558 (1982).

158.   J. J. Hopfield, D. Tank, ``Neural'' computation of decisions in optimization problems. *Biol. Cybern.* **52**, 141–152 (1985).

159.   J. Hertz, A. Krogh, R. G. Palmer, *Introduction to the Theory of Neural Computation* (Addison-Wesley, 1991).

160.   C. A. Del Negro, G. D. Funk, J. L. Feldman, Breathing matters. *Nat. Rev. Neurosci.* **19**, 351–367 (2018).

161.   J.-M. Ramirez, N. Baertsch, Defining the Rhythmogenic Elements of Mammalian Breathing. *Physiology* **33**, 302–316 (2018).

162.   F. Peña, M. A. Parkis, A. K. Tryba, J.-M. Ramirez, Differential contribution of pacemaker properties to the generation of respiratory rhythms during normoxia and hypoxia. *Neuron* **43**, 105–117 (2004).

163.   C. A. Del Negro, *et al.*, Sodium and calcium current-mediated pacemaker neurons and respiratory rhythm generation. *J. Neurosci.* **25**, 446–453 (2005).

164.   D. Levenstein, G. Buzsáki, J. Rinzel, NREM sleep in the rodent neocortex and hippocampus reflects excitable dynamics. *Nat. Commun.* **10**, 2478 (2019).

165.   P. J. Gonçalves, *et al.*, Training deep neural density estimators to identify mechanistic models of neural dynamics. *bioRxiv*, 838383 (2019).

166.   S. R. Bittner, *et al.*, Interrogating theoretical models of neural computation with deep inference. *bioRxiv*, 837567 (2019).

167.   J. Gunawardena, Models in biology: "accurate descriptions of our pathetic thinking." *BMC Biol.* **12**, 29 (2014).

168.   C. Omar, J. Aldrich, R. C. Gerkin, Collaborative infrastructure for test-driven scientific model validation. *of the 36th International Conference on …* (2014).

169.   R. C. Gerkin, R. J. Jarvis, S. M. Crook, Towards systematic, data-driven validation of a collaborative, multi-scale model of Caenorhabditis elegans. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **373** (2018).

170.   T. Kohonen, *Self-Organization and Associative Memory* (Springer-Verlag, 1984).

171.   T. Kohonen, *Content-Addressable Memories* (Springer, 1980).

172.   G. Lakoff, *Women, Fire, and Dangerous Things* (University of Chicago Press, 1990).

173.   E. Rosch, Prototype classification and logical classification: The two systems. *New trends in conceptual representation: Challenges* (1983).

174.   K. Obermayer, G. G. Blasdel, K. Schulten, Statistical-mechanical analysis of self-organization and pattern formation during the development of visual maps. *Phys. Rev. A* **45**, 7568–7588 (1992).

175.   K. Obermayer, T. J. Sejnowski, Howard Hughes Medical Institute Computational Neurobiology Laboratory Terrence J Sejnowski, T. A. Poggio, *Self-organizing Map Formation: Foundations of Neural Computation* (MIT Press, 2001).

176.   N. V. Swindale, H.-U. Bauer, Application of Kohonen's self-organizing feature map algorithm to cortical maps of orientation and direction preference. *Proc. R. Soc. Lond. B Biol. Sci.* **265**, 827–838 (1998).

177.   N. V. Swindale, How different feature spaces may be represented in cortical maps. *Network* **15**, 217–242 (2004).

178.   E. de Villers-Sidani, M. M. Merzenich, Lifelong plasticity in the rat auditory cortex: basic mechanisms and role of sensory experience. *Prog. Brain Res.* **191**, 119–131 (2011).

179.   M. Nahum, H. Lee, M. M. Merzenich, Principles of neuroplasticity-based rehabilitation. *Prog. Brain Res.* **207**, 141–171 (2013).

180.   D. J. Freedman, M. Riesenhuber, T. Poggio, E. K. Miller, Categorical Representation of Visual Stimuli in the Primate

Prefrontal Cortex. *Science* **291**, 312–316 (2001).

181.   D. J. Freedman, M. Riesenhuber, T. Poggio, E. K. Miller, A comparison of primate prefrontal and inferior temporal cortices during visual categorization. *J. Neurosci.* **23**, 5235–5246 (2003).

182.   T. J. Wills, C. Lever, F. Cacucci, N. Burgess, J. O'Keefe, Attractor Dynamics in the Hippocampal Representation of the Local Environment. *Science* **308**, 873–876 (2005).

183.   L. L. Colgin, *et al.*, Attractor-Map Versus Autoassociation Based Attractor Dynamics in the Hippocampal Network. *J. Neurophysiol.* (2010).

184.   T. Yang, M. N. Shadlen, Probabilistic reasoning by neurons. *Nature* **447**, 1075–1080 (2007).

185.   K. Jezek, E. J. Henriksen, A. Treves, E. I. Moser, M. B. Moser, Theta-paced flickering between place-cell maps in the hippocampus. *Nature* **478**, 246–249 (2011).

186.   E. Kelemen, A. A. Fenton, Coordinating different representations in the hippocampus. *Neurobiol. Learn. Mem.* **129**, 50–59 (2016).

187.   O. Guest, A. E. Martin, How computational modeling can force theory building in psychological science (2020) https:/doi.org/10.31234/osf.io/rybh9.

188.   I. van Rooij, G. Baggio, Theory before the test: How to build high-verisimilitude explanatory theories in psychological science (2020).

189.   W. Bechtel, Integrating sciences by creating new disciplines: The case of cell biology. *Biology and Philosophy* **8**, 277–299 (1993).

190.   A. D. Redish, Addiction as a computational process gone awry. *Science* **306**, 1944–1947 (2004).

191.   A. D. Redish, S. Jensen, A. Johnson, A unified framework for addiction: vulnerabilities in the decision process. *Behav. Brain Sci.* **31**, 415–487 (2008).

192.   P. R. Montague, R. J. Dolan, K. J. Friston, P. Dayan, Computational psychiatry. *Trends Cogn. Sci.* **16**, 72–80 (2012).

193.   A. D. Redish, J. A. Gordon, Eds., *Computational Psychiatry: New Perspectives on Mental Illness* (MIT Press, 2016).

194.   Q. J. M. Huys, T. V. Maia, M. J. Frank, Computational psychiatry as a bridge from neuroscience to clinical applications. *Nat. Neurosci.* **19**, 404–413 (2016).

195.   B. A. Richards, *et al.*, A deep learning framework for neuroscience. *Nat. Neurosci.* **22**, 1761–1770 (2019).

196.   A. Gomez-Marin, A. A. Ghazanfar, The Life of Behavior. *Neuron* **104**, 25–36 (2019).

197.   P. Cisek, Resynthesizing behavior through phylogenetic refinement. *Atten. Percept. Psychophys.* **81**, 2265–2287 (2019).

198.   R. Brette, Is coding a relevant metaphor for the brain? *Behav. Brain Sci.*, 1–44 (2019).

199.   R. E. Kass, *et al.*, Computational Neuroscience: Mathematical and Statistical Perspectives. *Annu Rev Stat Appl* **5**, 183–214 (2018).

200.   B. Devezer, L. G. Nardin, B. Baumgaertner, E. O. Buzbas, Scientific discovery in a model-centric framework: Reproducibility, innovation, and epistemic diversity. *PLoS One* **14**, e0216125 (2019).