# THE BRAIN ABSTRACTED

## SIMPLIFICATION IN THE HISTORY AND PHILOSOPHY OF NEUROSCIENCE

M. CHIRIMUUTA

# The Brain Abstracted

# The Brain Abstracted

## Simplification in the History and Philosophy of Neuroscience

M. Chirimuuta

The MIT Press
Cambridge, Massachusetts
London, England


The MIT Press would like to thank the anonymous peer reviewers who provided comments on drafts of this book. The generous work of academic experts is essential for establishing the authority and quality of our publications. We acknowledge with gratitude the contributions of these otherwise uncredited readers.

This book was set in Stone Serif and Stone Sans by Westchester Publishing Services.

Library of Congress Cataloging-in-Publication Data

Names: Chirimuuta, M. (Mazviita), author.
Title: The brain abstracted : simplification in the history and philosophy of neuroscience / M. Chirimuuta.
Description: Cambridge, Massachusetts : The MIT Press, [2024] | Includes bibliographical references and index.
Identifiers: LCCN 2023024648 (print) | LCCN 2023024649 (ebook) | ISBN 9780262548045 | ISBN 9780262378635 (epub) | ISBN 9780262378628 (pdf)
Subjects: LCSH: Neurology. | Neurons. | Neurosciences—Philosophy. | Cognitive science.
Classification: LCC QP355.2 .C475 2024 (print) | LCC QP355.2 (ebook) | DDC 612.8/233—dc23/eng/20231010
LC record available at https://lccn.loc.gov/2023024648
LC ebook record available at https://lccn.loc.gov/2023024649

O Natura omnium Mater Dea, artificiosa admodum Dea,
Suscitatrix honorabilis, multa creans, Divina Regina,
Omnidomans, indomita gubernatrix, ubique splendens.
—Hymn of Orpheus in Robert Boyle (1686/1996, 52)

# Contents

# Preface

Looking over my manuscript in its entirety, I found it more sprawling than I would have liked and less neatly argued than you might expect. Then it struck me that this fault is almost an inevitable one since the overarching argument of the book is that living nature, in its excess of complications, cannot be known as such. Scientific knowledge is necessarily produced by simplification, by pruning, taming, suppressing, dominating complexity—choose the metaphor at whichever degree of Baconian intensity you like. The brain, since it is a hypercomplex object (or process) makes the presence of this struggle most obvious to the observer of scientific practice. To present my argument in a neat and tidy package that masks as best as possible my own efforts to grapple with the complexity of the subject matter, to conform to the intellectual/aesthetic standard conditioned by a multigenerational conceit that nature is simple and truths clearly expressible, to write as if nothing were residual or refractory to my single-minded line of thought would, I submit, be self-undermining. "Sorry, not sorry" for the imperfect state of this book: had I more time, I could have made it shorter (and more cogent), but that would have been dishonest.

In the course of this project, I came to realize that science, as we know it, is founded on an essentially theological belief in the rational intelligibility of nature, which entitles seekers after the truths of nature to employ simplicity as their guide. When Nietzsche heralded the Death of God, he had the monotheism of Being in mind—the theology that denies ultimate reality, and hence value, to the ever-changing, ever-complicating appearances of the natural world. As irreligious and self-avowedly naturalistic science and philosophy of science were in the century after Nietzsche, the basics of the belief system were not updated. Beauty, truth, and parsimony were left high on their pedestals. Engineers are now on the scene to smash these idols. This book is the product of a historical rupture that has become visible in the twenty-first century between a classical scientific approach, which seeks

simple, intelligible principles underlying the manifest complexity of nature, and a data-driven engineering approach, which is rather too happy to give up on the search for elegant, explanatory laws and models.[1]

These engineers have taken to the mindset of unrestrained Will to Power. My wish, instead, has been to draw out another lesson from the recognition of the failures and limitations of the classical approach—a lesson of theoretical and practical humility. Although the stated aim of the book is not to reform neuroscience or offer advice to neuroscientists, but rather to interpret their work, I have not always kept myself within the bounds of description of scientific practice. The prescriptive voice—on the folly of overestimating our comprehension of nature in itself, in its full complexity—does frequently come out. What greater hubris, then, than headlong pursuit of interventions without even the attempt at understanding that which is being altered? Hubris or humility is precisely the juncture that is faced in the twilight of the old theology—too important a decision to be left to the scientists and technologists.

Even when writing an academic book, one likes to think that it will be read by as many people as possible. I have tried to accommodate at least three audiences: philosophers, neuroscientists, and any other interested parties. This ambition has led me sometimes to neglect philosophical and technical intricacies. I beg the reader's indulgence. When describing invasive and in many cases painful experiments on animals, I have followed the convention of a detached style of writing because it seemed that to do otherwise would cause an unnecessary distraction. This is not a reflection of my actual feelings, but I feel that it deserves some apology, to whom I know not.

Edinburgh, September 2022

Books, even books about history, are solid things, susceptible to the gnawing ants of time. I have added just a few references to work published since 2022, when the manuscript was essentially completed. My belief is that we can extrapolate gracefully from the content as it stands, even though the last decade seems an age away from neuroscience and technology today. For those seeking the up to date and the ephemeral, there is always the internet.

Bretby, July 2023

1. See Breiman (2001). For this division in a nutshell, consider the clash between Noam Chomsky and Peter Norvig over statistical language models (Katz 2012; Norvig 2012).

# Acknowledgments

My first conception of this project was that I would expand upon my previous publications on computational explanation in neuroscience. Thanks to a sabbatical supported by the University of Pittsburgh for the duration of 2018, I had the time for deeper reading into early twentieth-century philosophy of science and theoretical neurology, and this led to a substantial turnaround in my own views on the meaning and status of computational models of the brain. I am grateful to students and colleagues in the History and Philosophy of Science and Philosophy departments at the University of Pittsburgh, and Carnegie Mellon University, for queries and conversations that helped me find my way in this new endeavor: in particular, those who joined the History of Neuroscience seminar in 2017, including my coteacher Paolo Palmieri; members of the 2018 Perspectivism reading group, organized by Sandra Mitchell; and students in my 2019 seminar on scientific realism. This project benefited greatly from my attendance of Stephen Engstrom's 2018 Kant seminar, though, naturally, I am fully responsible for the many deviations from the true path that my book contains.

Early versions of this material were exposed to generous criticism on many occasions. My gratitude extends to audiences at the following events (I hope I have not forgotten any): the 2016 workshop on Grounding Sensible Qualities at Berkeley; 2017 workshop on Analogical Reasoning at LMU Munich; the 2017 Rutgers Center for Cognitive Science (RuCCS) Colloquia Series; the 2017 Philosophy colloquium at the University of Birmingham; PPN (Philosophy, Psychology and Neuroscience) seminars at the University of Glasgow in 2017 and 2020; the 2017 Philosophy colloquium at the University of Cincinnati; the 2017 Graduate Conference at the University of Waterloo; the 2018 lunchtime speaker series at the Institute of Philosophy,

London; the 2018 Neural Mechanisms webinar, organized by Marco Viola and Fabrizio Calzavarini; the 2018 PPIG (Philosophy, Psychology, Informatics Group) at Edinburgh University; the 2019 Annual Lecture Series at the Pittsburgh Center for Philosophy of Science; the 2019 Philosophy colloquium at the University of Hannover; the 2019 inaugural SURe (Scientific Understanding and Representation) workshop; the 2020 Summer of Consciousness webinar, organized by Uriah Kriegel; the 2020 meeting of the Scots Philosophical Association; the 2020 Learning Salon, organized by John Krakauer, Ida Mommenejad and Joshua Vogelstein; the 2020 webinar on Evidence, Models, and Explanation at the India Institute of Technology; the 2021 Stanford Center for Mind, Brain, Computation, and Technology symposium; the 2021 Cognition, Values, Behaviour (CVBE) research group at LMU Munich; the 2021 meeting of the European Philosophy of Science Association; the 2021 Dutch Distinguished Lecture Series in Philosophy of Neuroscience; the 2021 Philosophy colloquium at the University of Bristol; the 2021 Edinburgh Speaker Series in Philosophy; and the 142nd session of the Aristotelian Society, 2022 in London.

Very warm thanks go to everyone who took part in the reading group on the manuscript during 2021–2022 at the University of Edinburgh—Dimitri Coelho Mollo, Lachlan Devine, Carrie Figdor, Alistair Isaac, Michela Massimi, Camden McKenna, Kate Nave, Mark Sprevak, Dave Ward, and Jo Wolff—your enthusiasm was wonderful, and I can only hope that my other readers will be as perceptive and charitable. I must also thank various people who were kind enough to read individual chapters and share their thoughts: Colin Allen, JP Gamboa, Karen Kastenhofer, John McDowell, Nedah Nemati, Mark Paterson, and Paul Teller.

I thank my mother for believing in me, and my husband for being just incredulous enough so that I do not get complacent. One more thank-you goes to my children, Claire and Tendai, who remind me that the future is open-ended: you did not conspire to prevent me from writing this book, and for that I feel very lucky.

# Part I

# 1 Introduction

> The truth is, there's not anything that has, and doth still delude most men's understandings more, than that they do not enough consider the variety of nature's actions . . . ; preferring art and experiments, before reason; which makes them stick so close to some particular opinions, and particular sorts of motions or parts; as if there were no more motions, parts, or creatures in nature, than what they see and find out by their artificial experiments. Thus the variety of nature, is a stumbling block to most men . . . : and how should it be otherwise, since nature's actions are infinite, and man's understanding finite?
>
> —Margaret Cavendish (1668/2001, 99)

## 1.1 The Most Complicated Thing

Allow me to overwhelm you with details. Around 85 billion electrically excitable cells (*neurons*), and the same number again of additional cells (*glia*), are housed together in the human skull. The brain is only one part of the nervous system, and hundreds of millions more neurons lie within the spinal cord and surround the intestines—this is known as the *enteric nervous system*, which will be ignored in this book, despite claims that it forms part of the basis of mental life. Most of our attention will be on the cortex, the wrinkled outer bark of the human brain which, though critical for memory, perception, and deliberate action, contains only about 19 percent of the neurons in the brain and is estimated to parcel out into 180 anatomically and functionally differentiated regions per hemisphere. The variety of shapes and sizes of neurons, distinguished by the branching fibers of their axons and dendrites, presents the neuroanatomist with a specimen chamber of classifiable kinds. The retina—which is an outgrowth of the brain—alone is thought to contain

100–150 neuronal types. Glial cells, often regarded only as connective "brain-glue," make up their own family of types and subtypes, now thought to have distinct functional roles. The population of astrocytes alone includes sixteen different kinds: protoplasmic astrocytes, fibrous astrocytes, surface-associated astrocytes, velate astrocytes, radial glia, radial astrocytes, pituicytes, gomori astrocytes, perivascular and marginal astrocytes, ependymocytes, choroid plexus cells and retinal pigment epithelial cells, interlaminar astrocytes, polarized astrocytes, and varicose projection astrocytes.[1]

Yet it is when we look at the finer details, right down to the cell level, that things really start to get complicated. It used to be said that each neuron had input structures (*dendrites*) and an output fiber (*axon*) for sending on electrical messages. The axon of neuron 1 would connect to the dendrites of neuron 2 via a nanoscale gap called a *synapse*, across which neuron 1 would send little packets of neurotransmitter to either excite or inhibit neuron 2, making it either more or less likely to send its own electrical signals (*action potentials* or *spikes*) down its axon. Many observations upset this textbook picture. For one thing, dendrites of cortical neurons themselves produce action potentials and have been modeled as miniature neural network computers (Gidon et al. 2020). Details of dendritic structure are thought to have far more importance for neuronal function than has long been assumed (Larkum 2022). The standard account of learning in the brain supposes that assemblies of neurons adjust their connectivity patterns by enhancing or weakening the strength of synaptic connections through the plasticity mechanisms of *long-term potentiation* (LTP) and *long-term depression* (LTD), respectively. While it is anticipated that neurotransmission and plasticity could be achieved through a handful of proteins, molecular investigation of the synapse to recover the "proteome"—the number of kinds of proteins expressed—far outstrips any parsimonious picture, settling at 2,000–3,000 per synapse (Grant 2018). Within the cerebellum, the "little brain" that makes up 10 percent of the brain's mass but contains 80 percent of its neurons, learning was attributed to LTD, but experiments have revealed a surprising heterogeneity of plasticity rules (Suvrathan and Raymond 2018). The same story could be told about the heterogeneity of ion channels, the proteins spanning the cell membrane, gating the passage of specific ions in and out of the neuron (hence governing its

1. The sources for this paragraph are Furness (2006), Lent et al. (2012), Zeng and Sanes (2017), Van Essen and Glasser (2018), and Verkhratsky et al. (2019).

electrical activity), and similarly, the beginnings of an account of the physiological role that this complexity might have. But this book is not about the complications of the brain, it is about the seeking and making of simplicity.

When, in the spirit of the compilers of wonders of the world, people make lists of the most complex things in the known universe, the human brain wins the prize (e.g., Ladyman and Wiesner 2020, 61). There may be a touch of vanity in this, for even the nervous system of *C. elegans* (the roundworm) is not so simple that it is obvious how the combined operation of its 302 neurons gives rise to each of the little worm's movements and other behaviors (Sarma et al. 2018). Every nervous system is complex, and each in its own way. The task of the scientist, more than anything, is to make it seem as if this were not so. One of the most scandalous things about the ill-fated Human Brain Project—a billion-euro moonshot to build a detailed, realistic simulation of a mammalian cortex in a supercomputer—was that it did not try hard enough to simplify, to find a division between the "relevant" and "irrelevant" details out of all the troves of data that went into its compilation (Koch and Buice 2015). In contemporary neuroscience, computational theories are seen as the most promising way to achieve this division (Ballard 2015, 3). In the course of the book, I will discuss computation and other strategies. There will be theories and models that are elegant and beguiling in their simplicity. Yet we should never forget this first brush with complexity, which is the uncatalogued reality that each of us carries around in our cranium.

## 1.2 Thought Is a Bandpass Filter

Every time you conceptualize, categorize, and put a name on something that is not a proper name, you abstract away from its particularities. Picture daisies and clover flowers in a lawn. Those four ordinary nouns elide their differences. "Flower" co-categorizes the white and yellow types with the beige ones, and all the many other sorts to be found elsewhere. "Lawn" neglects the varieties of grass and all the nongrassy plants that are there. Zoom in, and you will find individuality and uniqueness everywhere. No two daisies, no two petals, are exactly alike, and yet they present to a quick glance a carpet patterned uniformly enough. For most practical purposes, the differences can be ignored—making a daisy chain, sunbathing, and the like. Not so, however, for the groundskeeper of a sports stadium, where the constituent grasses and their stages of growth really do matter. And to an infinite mind, with

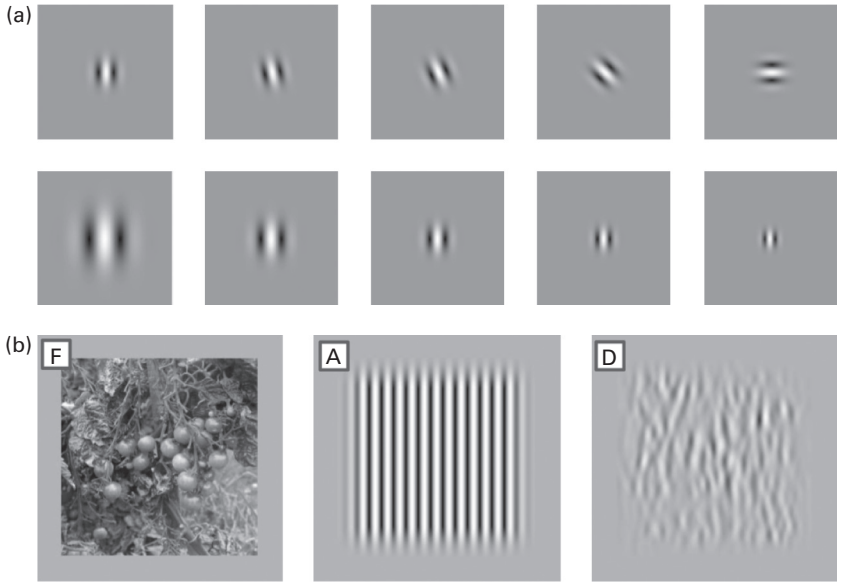


**Figure 1.1**
(a) The concept of spatial frequency illustrated with Gabor patch stimuli. The black-and-white pixels in each patch are modulated according to a sine wave, at a particular frequency and orientation. Compare the highest spatial frequency (farthest right) and lowest spatial frequency (farthest left). (b) Left is the original photograph, a "natural image." Right is the band-pass-filtered version of the image, which only shows structure at the orientation and spatial frequency range indicated by the central sinusoidal grating.

infinite memory, each blade of grass, with its own distinct life history, need not be co-categorized with all its fellows. Each could have its own name, as you yourself do.

We human beings each carry a separate name for ourself, marking us as individuals. Yet in so many practical, administrative tasks, humans are treated as just another kind of biomass. If estimating foot traffic through a supermarket or railway station, for planning purposes such as food distribution or maintenance of social distancing, the individuality of each person counts for nothing. A person becomes a number so often in this society, but to ourselves, we see that the details and differences matter. And so with hens, which are social animals. They are individuals to one another, having favorites, allies, and enemies. Packed together in a barn, they are just instances

of categories: good, average, or bad layers; twelve, eighteen, or six months old. For the kind of thought known as *discursive*, which employs linguistic and mathematical concepts, there is necessarily an abstraction away from the particularities of the things referred to, and the shape of the abstraction itself depends on the practical tasks within which the thought is operative. When talk is of "relevant" and "irrelevant" details, we must always ask, "to whom?" and "for what purposes?"

The band-pass filter was my first encounter with simplification of the brain. As a student, I was taught to do experiments in which filtered images like the ones in figure 1.1b were flashed up on a computer screen, and subjects were instructed to report which of a near identical pair had a slightly higher contrast. The body of theory motivating these experiments, in development since the 1960s, proposed that the early visual system, running from the retina through a precortical structure called the *thalamus* to the primary visual cortex (V1), contains an array of "spatial frequency channels" that selectively respond to elongated patterns at particular widths, like the ones in figure 1.1a. Evidence for this theory came both from psychophysical findings of selective adaptation to particular spatial frequencies and from physiological recordings of early visual neurons with a preferred sensitivity to a narrow range of spatial frequencies and orientations (De Valois and De Valois 1988). It is comparable to the theory of audition first proposed by Helmholtz in the nineteenth century, which says that the auditory system analyzes each complex sound wave arriving at the ear into the set of distinct frequency components that make up the stimulus, beginning with the hair cells in the cochlea, each sensitive to a limited frequency band. A task of my PhD was to find out if by positing that V1 is a spatial frequency analyzer and modeling the responses of individual neurons as essentially linear filters that detect the presence of edges in the image, we could predict the data collected from experimental subjects (including members of the lab) looking at ordinary black-and-white photographs. The results were mixed (Tolhurst et al. 2010), but it puzzled me that such drastically simplified models of neurons could even have a hope of summarizing the responses of the entire visual system. Indeed, chapter 5 of this book will be about the strengths and weaknesses of these kinds of models.

A curious thing is that the function ascribed to neurons in the visual system by the channel theory is the task of discursive thought, which is to selectively take up a certain pattern or regularity by letting it through

the filter—subsuming it under a concept—and purposefully ignoring everything else. To conceive of flowers is to create a filter in one's mind for all the regularities shared in common among flowers, ignoring the particularities that distinguish kinds of flowers and individual flowers from one another, and disregarding all else that is not flowery in the world. To look at a lawn and call all the white and yellow forms there "daisy" is to impose a low-pass filter on the scene, which blurs out the fine-grained, distinguishing features of each individual so all come out the same in one's conception. And the simplifications that neuroscience must undertake are like this too. No two brains, no two neurons, are really identical, but for the purpose of building theories and models that predict and explain responses, individuality must be blurred out and a uniformity of categorizations must be imposed.

Much more will be said in this book about the simplifying strategies of neuroscience. My final point here is an admission that this study is just as much a band-pass filter over an inexhaustible variety of methods and ideas that make up neuroscience, past and present. The thesis of the book is that the dominant ideas that have shaped neuroscience are best understood as attempts to simplify the brain. In presenting this thesis, I am putting to one side numerous other explanatory grids that could be imposed on the discipline and might each yield insights. The decision to concentrate on simplification in neuroscience is itself a concession to following the dictates of a simplifying strategy. What I can hope is that the results of this filtering process will prove instructive. To borrow a different visual metaphor, used often by Kurt Goldstein (1934/1939), I have chosen to put simplification in the foreground, but the reader should not forget that the background remains, visible but unattended.

## 1.3 Simple and Complex

I have not yet spelled out what I mean by "simple" and "complex." As many readers will know, there exists a science of complex systems, and one of its challenges has been to define complexity such that it can be consistently measured across entities as different as algorithms, cells, and economies. Since there is no consensus definition (Mitchell 2009a, 94, Ladyman, Lambert, and Wiesner 2013), I will not pin myself to one; I prefer instead to highlight some features of complex systems that best help to demonstrate why the brain presents such a challenge. The chapter began with the first

two of these criteria: the number and the heterogeneity of components. The billions of cells in the brain and the rest of the human nervous system are too many to enumerate. Complex systems will have a very large number of parts. But the number of atoms in a grain of salt is also, to all intents and purposes, uncountably big. The difference is that those sodium and chloride ions are homogenous and regularly arranged, whereas the neurons are highly diverse and only to some extent form a regular array, as with the stereotyped Purkinje cell circuits of the cerebellum. And it must be emphasized that this "crystalline" depiction of the cerebellum, as comprising a uniform elementary structure repeated over and over, can be only a distorting approximation since the fine-tuned differences and modulations of the cerebellar synapses are essential to its functions.[2]

This brings us to the third feature, which is the heterogeneity of the parts of the system across time. Brain tissues, like all living materials, are always going through processes of reconstitution like a Ship of Theseus. Neurons are especially sensitive and plastic. The reactivity and changeability of the brain, in response to experience in the short and long term, are essential to the role of this organ, which is to support appropriate behaviors in a changing environment. The brain is sculpted by time, which is fundamental to its being the seat of learning, and hence intelligence. But this same tendency to change makes the brain particularly challenging as an object of science, for the brain will never be in exactly the same state twice. Experimental neuroscience is required to gather neural response data from presentation of identical stimuli over multiple trials. The responses show trial-to-trial variability, which is normally averaged away and classified as noise. Chapter 7 will be about the "Heraclitean" brain and how this kind of complexity presents a limit to any claim of scientific models to convey a perspective-independent truth about neural function and activity.

The fourth way in which the brain is extremely complex is in the high number of interactions among its components. The textbook account is that the synaptic connection between an axon and a dendrite is *the* way that the elements of the nervous system interact with one another. Just on those terms, the system is excessively interactive. For instance, each cortical neuron is estimated to have a synaptic connection with around 10,000

2. See, for instance, Zeng and Sanes (2017) and Cembrowski and Spruston (2019) on neuronal heterogeneity and the problem of cell type classification.

other cells (Kunkel et al. 2012). In addition to the basic account, there are other modes of interaction, such as chemical neuromodulation, signaling among glia, and between glia and neurons. The result of all this interactivity is that the behavior of each component at any one time is very context dependent, which is a fifth form of complexity. As will be discussed in chapter 3, reductionist strategies in science assume the opposite—that the behavior of a part of the system is approximately context *independent*, such that it is feasible to study parts in isolation from the whole in order to gain knowledge of their normal operation within the complete system. Reduction runs aground when the accumulation of knowledge concerning the parts in isolation is not helping to explain the behavior of the collective.[3] It is actually an important open question, whether this is the situation confronting neuroscience today. So many facts have been discovered about individual neurons, as well as subneuronal structures, by examining them under a microscope in slice preparations and in petri dishes, in isolation from the rest of the brain. Much of this research is what underpins systematic approaches to drug discovery for neuropathologies. But if the behavior of the parts in isolation does not reveal their role in the pathology as a whole, the strategy is flawed.

The sixth sort is the kind of complexity that I call *organizational depth*, which is akin to the notion of "degree of hierarchy" (Mitchell 2009a,109).[4] The idea here is that gross divisions of the nervous system can be made to demarcate it into working parts—an early discovered one being the sensory versus motor nerves at the roots of the spinal cord (Berkowitz 2015); furthermore, these parts also afford subdivisions into elements that themselves show intricate organization, and in turn into further sublevels of complex operations. This is in contrast to something like a clockwork toy in which the whole divides into metal parts with interesting shapes and an intricate system of organization, but those parts themselves are just homogeneous pieces of metal and do not reveal any other kind of clockwork system in miniature. Leibniz held that organisms differed from mechanisms produced

3. This is often given as a hallmark of *emergent* behaviors. I will not go into the relationship between emergence and complexity here, not least because working out a definition of emergence would be a diversion. Anderson (1972) is a classic reference on this question. Also see Gillett (2016) on concepts of emergence.

4. The source here is Herbert Simon (1962) on the "architecture of complexity." This topic will be revisited in chapter 10.

by the human hand, being "divine machines." That is, living beings were said to be machines whose parts were themselves machines, and the parts of the submachines were themselves machines, ad infinitum (Smith 2011). As I will argue in chapter 4, the brain has much more the character of a "divine machine," but the popular comparison of brain and computer—an artifact lacking so many layers of organizational depth—serves precisely as a means to abstract away from the complexity that is housed within the subsystems of the living factory that is the brain.

An idea from complexity science that has gained traction in neuroscience is that when you have a complex system with hierarchical organization into distinct "levels," the behavior of the upper-level structures (i.e., the ones closest to the whole system) are pretty much indifferent to variation in the structures and processes at the lowest levels. For example, Dehghani (2018, 2) writes:

> The macroscopic behavior of the system (such as network balance of excitation/inhibition) is insensitive to the computational state of individual neurons. . . . This insensitivity is not because the functional symmetry of individual elements transcends to the total state (Anderson, 1972), but because interconnectedness renders many details (at fine scale) to be irrelevant at the large-scale behavior of the system (Goldenfeld and Kadanoff, 1999). Thus, attempts for precise control of the system at its fine scale is precisely where it will fail.

The point here, that it is misguided to attempt to control the gross behavioral "outputs" of the brain by tweaking a neuron here and there (as those following reductionist precepts have done), is well taken. However, I advise caution, now and in the course of the book, not to be too quick to assume that all those details at the finest levels of organization do not matter to system behavior, and to trust that highly abstract models are preserving all the details essential to explaining it. A leading direction of complexity science has come from physicists eager to show how you can get simple lessons from complex things (see Goldenfeld and Kadanoff 1999). Physical systems quickly bottom out into quite homogeneous, elementary components, but living systems are not like that, and this cannot be unrelated to the fact that organisms can do countless things that can never be expected of inanimate objects. Moreover, because of mechanistic depth complexity and the complicated patterns of connectivity within the cortex, it turns out that there is no unambiguous hierarchy of levels to be found, such that low-level and high-level processes can be neatly separated from one another (Hilgetag and

Goulas 2020, Chirimuuta 2022a). An aim of this book is to shift people away from the appealing assumption that in neuroscience, there is something simple, and hence readily comprehensible, at the bottom of everything. That the changeable complexity of appearances reduces to elementary particles, or can be shown to be the product of compact laws, has been a working assumption of physics, but we must interrogate its suitability for employment elsewhere.

There will be little complexity science under discussion in the chapters to follow. But before leaving the subject, I would like to bring up the idea articulated by the physicist Murray Gell-Mann: his definition of *effective complexity*—namely, that systems that are complex are ones that seem to lie between order and randomness (Mitchell 2009a, 98–100). This is very much true of the complexity on display in the living world: there are patterns and regularity everywhere, and at the same time, on closer examination, there is always variation that looks random, although it also could be part of an indiscernible pattern. The regularity of kinds of organisms allows us to classify and utilize them, but we are also left with an intractable particularity, which at the very least in our own case—as individual human beings—we would not want to discount. A theme of this book is that the simplification of the brain has proceeded in neuroscience by seeking out orderliness and creating versions and representations of neural systems in which regularity is exaggerated while apparent disorder and particularity—which have their own claims to be essential features—are discounted. The science marches on by creating *ideal patterns*, and in that way, a complex system (lying between order and apparent randomness) takes on the guise of a simpler, more regular one.

From Cassirer, I borrow the thought that this kind of idealization, the projection of the observed, jumbled, real system onto a plane of ideal order, involves a reconstitution of the concrete objects of science in alien terms. He writes that "it is upon a peculiar interweaving of 'real' and 'not-real' elements, that every scientific theory rests. . . . That form of knowledge, whose task is to describe the real and lay bare its finest threads, begins by turning aside from this very reality and substituting for it the symbols of number and magnitude" (Cassirer 1910/1923, 117).

However, I depart from Cassirer in my insistence that the task of science is not description of reality per se; rather, it is description for the purposes of manipulation and control of its objects. As will be argued in chapter 8,

instrumental ends are not detachable from the most theoretical activities of neuroscience, and technological results are not the straightforward consequences of discoveries of pure science, as is too often assumed. More background on this broadly Kantian philosophy of science will be provided in chapter 2. I will now say more about three key forms of simplification that are operational in my case studies in part II.

## 1.4 Making Things Simple

There are three important simplifying strategies that have shaped the neuroscience under discussion in this book. They interact and reinforce one another in significant ways. The first strategy to consider is *mathematization*, the construction of a mathematical model, which is always an abstraction and normally an idealization—a purposeful distortion of the target of representation. The second is *reducing*, working materially with simplified preparations and making the reductionist assumption that the facts gleaned from those reduced experiments are building blocks for understanding the wider system. The third is the formation of *analogies* between the complicated, unfamiliar neural system and a simpler and more familiar artifact.

### 1.4.1 Mathematization

The kind of simplification afforded by mathematical models has received the most attention in recent philosophy of science, centered on the practices of abstraction and idealization. *Abstraction* is usually defined as the model's omission of details that exist in the target system, and *idealization* as the modeler's decision to represent the target in ways known to be false (e.g., Levy 2018 and references therein). For example, a model of a falling body that omits air resistance is said to be abstract because it leaves out a detail, whereas a model in genetics that posits an infinite breeding population is said to be idealized because no population could in fact be infinite. But since the omission of details is also a way to produce a false representation, it can be hard to see the difference between these two practices. To individuate them, it is instructive to consider their antonyms. The opposite of "abstract" is "concrete," whereas "ideal" stands in opposition to "real." Every mathematical rendering of a concrete entity or occurrence in nature is, properly speaking, an abstraction. A maximally detailed mathematical model, in which no forces are omitted, is still an abstraction in the sense of being a departure from the concrete. And

we see that it involves simplification because mathematization requires the projection of a homogeneity onto the items counted or measured, as well as neglect of their particularity. This is a point made by Henri Bergson:

> It is not enough to say that number is a collection of units; we must add that these units are identical with one another, or at least that they are assumed to be identical when they are counted. No doubt we can count the sheep in a flock and say that there are fifty, although they are all different from one another and are easily recognized by the shepherd: but the reason is that we agree in that case to neglect their individual differences and to take into account only what they have in common. . . . The idea of number implies the simple intuition of a multiplicity of parts or units, which are absolutely alike. (1889/2001, 76)

Any representation of a concrete system in the abstract language of mathematics is ipso facto a simplification. Against the view that mathematization does not always bring about simplification, because a system might be represented with a ridiculously complex mathematical structure, I assert that even in such a case, the mathematical model must abstract away from the particularities and qualitative properties of the target of representation, and the restriction to purely quantitative terms always brings about a simplification.

The opposition of "real" with "idealized" has the connotation of the superiority of an ideal version of things over a base reality. Although this connotation is not active in the current scientific terminology, it is worth pausing to examine it. The prototypical instance is the contrast between the actual shape of an object and the geometric form that it best matches. No real spherical ball could be a perfect sphere, and yet the perfect sphere, the one defined geometrically, is somehow the ideal shape that the real ball aspires to. With these geometrical idealizations, it is always the case that the ideal form is more regular and symmetric—simpler—than the real one. There is metaphysical and epistemological baggage behind this, to be discussed in section 1.5.1. My notion of the ideal pattern will extend this treatment to cases where patterns in nature are represented by a model, such as an equation whose form is likewise more regular, symmetrical, and in various other ways simpler than the reality depicted. Moreover, we will see in chapter 5 that experimental practices work to generate phenomena, which I also call ideal patterns, that are more regular and in various ways simpler than occurrences outside the controlled conditions of the laboratory. This brings us to the next strategy for simplification: reduction.

### 1.4.2 Reducing

*Scientists make things simple by making simple things*. The laboratory is a microcosm shielded away from the complexifying forces of uncontrolled variables, and within which the processes giving rise to observable phenomena can be kept on a tight reign of measurability and manipulability. It is only here that the ideal forms stated by the explanatory laws in physics actually apply—precisely the point argued in Nancy Cartwright's classic work on experimentation and modeling (Cartwright 1983, 1999). Neuroscientists refer to these simplifying settings as "reduced experiments." As we will see in chapter 5, the simplification is often brought about in cognitive neuroscience by choice of stimulus conditions to elicit less complex, more readable responses from the brain under investigation. The hope—now called into question by neuroscientists such as Hasson, Nastase, and Goldstein (2020)—was that the simpler stimuli and responses could form the basis for generalization to understanding neural activity in less controlled conditions. A reason to question the validity of the generalization stems from the brain's adaptiveness: neuronal responses quite readily conform to the statistics of the stimuli they are presented with, so by bombarding the brain with simple stimuli, one can create simplified patterns of response. This is the view argued by neuroscientists Gao and Ganguli (2015).[5] The mistake would be to think that by extrapolating from the results of those experiments, one can predict how the brain will respond in more complex, real-life situations. The advantage that the experimental physicist has over the neuroscientist is that the object of investigation is not inherently plastic and sensitive to the context of its surroundings, which is what the brain is to an extreme degree.

Experimental neuroscience has also employed reduction in the standard sense used elsewhere in biology, which means separation of a part of the organism from the rest, performance of precise examinations and interventions feasible only under such conditions of isolation—and with this detailed characterization of the parts in hand, the hope that an explanation of the behavior of the whole system can be assembled from those building blocks. Again, for reductionism to be viable, the assumption that the behavior of the parts is approximately the same in and out of context must hold well enough. Chapter 3 will present a case from the history of

5. Also see Gao et al. (2017), discussed in chapter 5.

neuroscience—the reflex theory of the brain—that took the reductionist track and failed, not least because of the unsuitability of the assumption of context independence of the part processes.

### 1.4.3 Analogy

The strategy of analogy is also one of simplification through making, although in this case, we are speaking of artifacts, technologies, and other cultural objects drawn in from outside neuroscience. The most saliant instance is the computer analogy for the brain, which will be the topic of chapter 4. And in chapter 6, I will argue that the positing of neural representations—a core idea in mainstream theoretical neuroscience—is driven by an analogy between neurons and artificial signaling systems. The deployment of analogies between the organ and the artifact works as a route toward simple models because while very complicated, the technologies are always relatively less complex than the brain. Thus they serve as a simplifying lens, a filter, through which to view the manifold processes within the brain. Moreover, artifacts have been designed by humans with particular functions in mind. By imposing a functional blueprint on the biological system (for which no explicit design exists), thus highlighting certain forms of interaction and not others, the scientist is permitted to abstract away from numerous details not deemed relevant to the function in question. The most famous instance of this procedure is David Marr's specification of the computational and algorithmic/representation levels of description for the visual system (Marr 1982).

While analogies with designed systems are salient in the history of biology (Canguilhem 1965/2008c), it is worth noting that modern physics likewise begins with the examination of machines. As will be explored in chapter 8, the thesis of historian of science Boris Hessen was that technology serves as a source of systems for science to theorize, and those theories and models are then used to explain the workings of things not humanly made (Freudenthal and McLaughlin 2009, 33). The salient example here is the discipline of mechanics, which started as the theory of machines and was then applied to celestial rotations and falling bodies of any sort. Again, the value of such analogical transfers is that work can begin with a relatively simple system, like a pulley or catapult, with the resulting model serving as a scaffolding for the examination of natural objects where the relationships between parts and forces is harder to grasp. At the same time, these analogical methods

themselves rest on the assumption that nature is simple, in the sense of being uniform in its operation from one kind of situation to another.

A kind of analogical strategy common in the history of physics is the transfer of models of familiar, observable, macroscopic systems to unobservable microscopic ones (Hesse 1966). Isaac Newton himself justifies this procedure via the assumption of the world's simplicity, writing that "nature is exceedingly simple and comformable to herself. Whatever reasoning holds for greater motions should hold for lesser ones as well" (quoted in Westfall 1981, 389). A key assumption of uniformity, discussed in chapter 8, is the one advocated by René Descartes—namely, that there is no essential difference between the operation of artifacts and the workings of nature, which underlies his account of living bodies as mechanisms and the role of machine analogies in his natural philosophy.[6]

A thing to observe, in conclusion of this section, is that these three strands together constitute the dominant way that simplicity has been sought within modern science. It is not inaccurate to say that the analogy of the universe as a machine is the centering thought that separates the main stream of research since the seventeenth century from older traditions of natural philosophy and their strands of continuity into the modern era, those being shaped by the analogy of the universe as an organism (Westfall 1981,14; more on this in section 1.5.2). Such root analogies form a *Weltanschauung*—the scientist's intuition, normally implicit, of how the world just is. The seeking of

6. The opposite view was expressed by Margaret Cavendish in the *Philosophical Letters* of 1664: "Art is not able to demonstrate nature" (quoted in Peterman 2021, 222).

The Cartesian view is certainly the dominant one nowadays, reinforced by the fashionable thought that it is impossible to draw a division between natural and cultural objects. However, I will argue at various points in this book that failure to attend to the differences, to disanalogies between living things and artificial devices, has been a mistake. In such cases, it is easy to see that the relevant distinction is between objects that are put together piecemeal by humans and living organisms that have come into being through some process of reproduction and are self-making (autopoietic). While the many organisms that have been selectively bred or genetically modified are both natural and artificial at the same time (just as human beings are both natural and cultural beings), they still count as different from artifacts in the sense relevant here. The brain of a genetically modified lab rat is no more like a computer than the brain of a wild rat, regardless of the part played by humans in the animal's creation. See Newman (2004) on historical treatments of the division between natural and technological objects.

simplicity through mathematization is related to this. As objects, machines are eminently mathematizable. Nothing is lost with the homogenizing gaze of mathematical abstraction since each machine, and each of its parts, is an interchangeable member of a class, not an individual unique unto itself. This is because machines are built for uniformity and for ready exchange and repair of parts (since they are not self-repairing). In turn, this relates to the suitability of machines for reductive explanation. The parts are characterizable independent of one another, and of the whole, since that is how they first came into being, at the hand of their manufacturer. The ontology of machine parts is not a relational one. The being of cogs and wheels and microchips (as opposed to these functionally derived labels) is not inherently dependent on their context within a device, as can be argued, in contrast, of the ontology of components of living systems (Dupré 2012). Reductive explanation reverses, through decomposition, the process through which a person would put together a machine. Thus, with the root analogy in place (namely, that the world and its inhabitants are mechanisms), it makes sense to seek simplicity by representing them mathematically and breaking them down into constituent parts.

## 1.5 On Simplicity and Truth

Why do scientists seek simplicity? I offer here two divergent answers to this question.[7] The first, the metaphysical-epistemological answer, is that the universe *is* fundamentally simple, such that a theory that is simpler is more likely to be true. The second, a psychological answer, is that human intelligence is limited since we are finite beings, so we are forced to look for and/or create simplicity because otherwise we cannot think and act. Two perspectives from physicists Galileo Galilei and Pierre Duhem, separated by centuries of time, express the contrast. It was observed prevously that mathematical descriptions are necessarily simpler than the concrete phenomena they depict. Thus, one way to express the view that the universe

7. Of course, there are other possible answers. One line would be to regard simplicity as a widely endorsed but negotiable scientific value, to be contrasted with alternative value schemes, which place higher value on heterogeneity, particularity, and complexity of interaction (Longino 1995). Another answer would be to cite the connection made by various scientists between simplicity, beauty, and truth (Ivanova 2020).

is fundamentally simple is to assert that mathematical entities have some sort of ontological priority. As is often quoted, Galileo's opinion was that the universe itself was a book "written in the language of mathematics,"[8] meaning that it was incumbent upon the investigator to learn the symbols of maths and geometry in order to acquire knowledge of it. The contrasting view about the significance of simplicity is expressed by Duhem:

> He [the scientist] will choose a certain formula because it is simpler than the others; the weakness of our minds constrains us to attach great importance to considerations of this sort. There was a time when physicists supposed the intelligence of the Creator to be tainted with the same debility. (1906/1954, 171)

Here, the seeking of simplicity is a requirement that stems from human limitations, not the nature of the universe beyond us. It is interesting that Duhem connects this issue to theological opinions—more on this in a moment. My sympathies are with the second answer. To the best of our knowledge, the world, especially the living world, is extremely complex, and this complexity has been refractory to various attempts to strip down the world to some simple order and structure. If truths about nature are to be had, they are likely to be "unsimple" ones (Mitchell 2009b). This section will explore how belief in the inherent simplicity of nature has shaped the past of science and how it figures in neuroscience today. In addition, it will show that the notion of simplicity in science is itself various, and complicated, figuring differently in different traditions of research, with a noticeable split between physics and biology in the way that simplicity has been conceptualized and sought.

### 1.5.1 Ockham and Desert Landscapes

There is no suggestion that Galileo's answer has been superseded by Duhem's one. In fact, both views are well represented in recent times. For example, Einstein tells us of a "deep faith that the principle of the universe will be beautiful and simple,"[9] whereas Ernst Mach, a physicist who was in other

8. The quotation is from Galileo's *The Assayer* of 1623, see the translation in Drake (1957, 237–238), and Remmert (2005) for discussion.

9. I have not yet been able to find the source for these often-quoted words. In the Herbert Spenser Lecture, which is Einstein's considered statement on the role of mathematical simplicity in physics, he tells us, "Our experience up to date justifies us in feeling sure that in Nature is actualized the ideal of mathematical simplicity" (Einstein 1934, 167). Note here that it is not an article of faith, but rather the *experience* of the

ways quite an influence on Einstein, argued that the scientist's seeking of compact laws is all about the achievement of "economy of thought," necessitated by the inherent limitation on how many empirical facts a human being can observe, memorize, and digest (Mach 1882/1895). Among computational and theoretical neuroscientists, many of whom began their scientific lives in physics, there seems to be a prejudice in favor of treating simple models as revealing of basic truths. We find embarrassment in how much of what is known about neurobiology is ignored or represented inaccurately by quantitative models, countered by the claim that they inform the scientist about some "essence." As Grace Lindsay writes, "All models *are* wrong, because all models ignore some details. All models are also wrong because they represent only a biased view of the processes they claim to capture. And all models are wrong because they favour simplicity over absolute accuracy. All models are wrong the same way all poems are wrong; they capture an essence, if not a perfect literal truth" (2021, 15).

The next passage is telling because it touches upon the psychological reason for the need for abstraction—that a simple model "allows us to think about a phenomenon more clearly"—but then veers, via another curious comparison with artistic production, into a conclusion that seems unmotivated by what has gone before, asserting the revelation of truth through the falsehood of abstraction; it is from an appendix to a major neuroscience textbook:

> What makes a model good? Clearly it must be based on biological reality, but modeling necessarily involves an abstraction of that reality. It is important to appreciate that a more detailed model is not necessarily a better model. A simple model that allows us to think about a phenomenon more clearly is more powerful than a model with underlying assumptions and mechanisms that are obscured by complexity. The purpose of modelling is to illuminate, and the ultimate test of a model is not simply that it makes predictions that can be tested experimentally, but whether it leads to better understanding. No matter how detailed, no model can capture all aspects of the phenomenon being studied. As theoretical neuroscientist Idan Segev has said, borrowing from Picasso's description of art, modeling is the lie that reveals the truth. (Abbott, Fusi, and Miller 2013)[10]

---

physicist that justifies belief in simplicity. Norton (2000) describes how Einstein was indifferent to the guiding power of mathematical simplicity until late in his career, with the development of the general theory of relativity.

10. Compare Eve Marder, interviewed in *Nautilus* magazine, who said: "The purpose of building a model should never be to attempt to replicate the fullness of biological complexity, but to provide a simplified version that reveals general principles."

It is not surprising to find these almost reflexive, not quite coherent expressions of the metaphysical belief that reality is simpler than it first appears, along with the epistemological principle that the search for simplicity is the path to some truth—whose discovery requires a little artistic flair. These ideas are entrenched in the history of philosophy, science, and theology. We need only think of the long reach of Platonism in the intellectual culture of Western Christendom. It is plausible to characterize the contrast between medieval natural philosophy and the mechanical philosophy of the seventeenth century (the latter being the world picture usually credited with paternity of modern science) as amounting to an intensified push for parsimony in the newer way of looking at things. The Aristotelian metaphysics of the Scholastics was routinely charged, by those seeking to replace it, with ontological excess—with the multiplication of beings beyond necessity. As reported by Pasnau (2011, 10–11), Ockham's procedure of reducing metaphysical categories into a limited set of more basic ones laid the groundwork for the early modern corpuscularian theory of matter,[11] and two central arguments in favor of the "mechanistic-corpuscularian framework" were its greater parsimony and intelligibility. These two theoretical virtues obviously pair well together since a simpler theory is easier to understand. But the premise that the simplicity and intelligibility of a theory are hallmarks of truth is not self-evidently true. "Even if we aim at intelligibility," Pasnau writes, "there is no guarantee that the world will cooperate. . . . Sometimes, from our vantage point, the world itself is just paradoxical" (2011, 48–49).[12]

---

But Marder then goes on to hint at a psychological reason for avoiding a very detailed model which represents neurobiological processes in a more realistic way: "[It] is destined to fail to produce new understanding because it will be as complex as the biological system" (Requarth 2015).

11. Ockham himself did not state Ockham's razor, and as noted by Spade (1999, 102) the maxim itself is not radical since versions of the thought can be found in Aristotle. The relevant breach is in how minimal an ontology is thought to be sufficient. The moderns set out to do more with far less than the norm among scholastics, and Ockham was a forerunner of the push toward minimalism.

12. The world of quantum physics is certainly paradoxical in comparison with the world of classical physics, conceived in the early modern period. Yet the commitment to simplicity seems indispensable in physics. As Falkenburg (2007, 38) writes:

> Quantum theory has shaken the traditional belief in the rationality, uniformity and simplicity of Nature. But to completely dispense with these principles would mean dispensing with physics as a science. Weakened versions of the principles of unity and simplicity have survived the transition to quantum theory.

One path to justification of the premise that simplicity is an indicator of truth came from theology. There is a subterranean connection between monotheism and the high estimation of simplicity (etymologically, "one-foldness"). An instance of the theological justification is in Nicolas Malebranche's principle that the created world would not do honor to God unless governed by simple, general laws of nature (Jolley 1997, xxxiii). Leibniz goes further than Malebranche in making simplicity itself a criterion for the evaluation of possible worlds. The most perfect of the possible worlds will be one both "simplest in hypotheses and the richest in phenomena" (Jolley 1997, xxxiv). Indeed, as Vassányi (2011, 5) observes, "It is a principle of early modern philosophical theology that God always chooses the simplest means to achieve the greatest possible effect." Thus, the investigator into nature has a guarantee that a simpler law is more likely to be true because it is more likely to have been chosen by the Creator. Physics, even these days, has not put theology fully to one side. Stephen Hawking (1995, 193) described the prospect of a grand unified theory in physics as a chance to "know the mind of God." The language may well be figurative, not an expression of faith, but it nonetheless stakes a theological position for it erases the division between finite and infinite minds. If the laws governing all of physical reality are simple enough to be discoverable and intelligible to finite minds, then that means there is no unbridgeable gulf between the human intellect and infinite intelligence when it comes to knowledge of the natural world.[13]

The principle of parsimony animates the dominant ontology in Anglophone philosophy today, which is *physicalism*. Even though it is quite tricky to characterize what exactly should be meant by "physical" (Wilson 2006), the idea is that the kinds of entities, processes, and properties discovered by the physical sciences exhaust all that can be said, fundamentally, to

13. See Martin (1951/1955, 6) on the long Platonic tradition in which "thinking the truth means becoming like God," and the theological grounds for the possibility of physics that are given by Leibniz:

> If all possible worlds and among them this world, are continuously thought by God, the being of the world is primarily a being thought and the world is therefore in its original being intelligible, transparent to reason. It may still be the case that real insight into the world is only possible for an infinite understanding, and that human thinking is finite and limited; but all this does not prevent the existence of the world from being in principle rational and hence conceivable and understandable by human beings, at least through an infinitely extended approximation.

exist. Indeed, we speak of all kinds of objects that are left uncharacterized in fundamental physics (radiators, senses of humors, cows, and daisies), but these are all said to be supervenient on—have a relationship of dependency on—the physical. In this, contemporary analytic metaphysics performs the role assigned to it by Hegel, of being an expression of its era in thought. The fewer kinds of things there are, fundamentally, the more inherently simple the world is, the more likely it is to be fully theorized, and therefore controlled. And so there is a veiled Will to Power in Quine's (1948) aesthetic preference for desert landscapes,[14] as well as in Jackson's (1998) jibes against the kind of ontology that resists the physicalist's forced choice of "reduce or eliminate" when it comes to problematic kinds such as colors, values, and persons. The charge is that unless you boil things down (a lot), ontology is no more than the drawing-up of big lists, a kind of mindless stamp collecting, or worse, the development of an overpopulated slum!

But unless the contemporary metaphysician is willing to bring theology once again to the argument, it is unclear what the justification for parsimony is.[15] There are gestures toward the reductions of other branches of science to physics, but these claims unravel in their historical details. One might point to the technological achievements that stem from formulation of theories and models that are simple. But a fatal weakness here—to anticipate the content of chapter 8—is reliance on the common but mistaken assumption that technology is the downstream consequence of discovery of truths about nature. Simple descriptions of natural occurrences are indeed better for technological control, not because they capture any essential truths but because they include only a distorted subset of details relevant for control.

### 1.5.2 Unity and Purpose

We have just considered a notion of simplicity in nature most strongly associated with physics. This notion prizes a stripped-down ontological minimalism and does not countenance the idea that anything important is lost

14. However, Quine's remarks on the value of simplicity in *The Web of Belief* are admittedly more complicated than this (McNulty 2021).

15. We saw in note 9 that Einstein felt that his own experience of confronting physical problems led him to the belief in the underlying mathematical simplicity of nature. But this could not work as a *general* argument for parsimony—other scientists have contrary experiences, especially ones working outside physics.

in the translation of natural occurrences into mathematical abstractions. It meshes well with the three simplifying strategies outlined in section 1.4, and it was noted that all these tendencies accompany the shift to a mechanistic *Weltanschauung*. This basic story needs to be enriched by turning now to the notions of simplicity that float around the supposedly superseded worldview of the universe as organism-like, rather than machine-like. These, as we will see, are still relevant to the study of simplification in neuroscience today.

I follow Peterman's (2021) use of the idea that explanation comes about via the unification of phenomena. While this is usually cashed out as the subsumption of phenomena under laws of nature, another path to unification is in the positing that various phenomena relate to one another through their belonging to one integral entity, the prototypical case being that of the organism. Peterman's argument is that the positing of souls, common in various natural philosophies often denigrated as animistic, seeks explanatory unification by claiming inherent relatedness (e.g., "sympathies" and "antipathies") among parts and processes of an entire, individual organism, and analogously among more disparate occurrences in the world, such as planetary motions and seasons.[16] It is to be noted that "unity" also has its etymological basis in "oneness" and is a version of simplicity. The relevant point for the purposes of this chapter is that the notion of an organism, as a unified, integrated being whose diverse parts and operations serve a common purpose, affords a way to conceive of simplicity in nature very different from the one I described earlier.[17]

16. The focus of Peterman's essay is the explanation of order of nature in early modern natural philosophy through positing of a "world soul," using an analogy between the organizing soul of one animal and the organizing principle of the whole universe. The most influential philosopher to attribute the unity and purpose of living beings to the presence of souls (the *psyche*) was Aristotle, in *De Anima*. Thomas Willis is an important figure in the history of neuroscience, and his accounts in the *Two Discourses* of 1683 posited various souls of animals. As with the mathematical notion of simplicity, world soul theories have links to Platonism since "ψυχὴ τοῦ κόσμου" appears in the *Timaeus*. This was taken up by the seventeenth-century Cambridge Platonists such as Ralph Cudworth: "The Universe in some sense [is], as the Stoicks and Platonists define it, one vast entire Animal" (quoted in Peterman 2021, 6).

17. We should note an important point of difference between unification via the laws of nature and the positing of souls as unifying principles. The search for laws of nature aims at discovery of an absolute regularity of phenomena. Paradigmatic natural laws are exceptionless. This high level of regularity is what makes laws invaluable guides for the prediction and control of the occurrences that they subsume.

A salient feature of this view is the idea of a whole as something that has ontological priority over its parts. Thus, there is a strongly antireductionist tendency here, in contrast with a reductionism that would assert the metaphysical and/or epistemological priority of the parts over the ensembles to which they belong. There is also an emphasis on the harmonious interplay of the parts, which has historically tied this organicist or holistic view to theories of beauty in aesthetics.[18] Goethe, the poet and natural philosopher, is representative of the approach, and an inspiration among later organicist biologists, such as Goldstein. In *The Organism*, Goldstein (1934/1939, 479) quotes the following words of Goethe: "In the human mind, just as in the universe, there is no top or bottom, All parts have an equal claim upon a common center which manifests its hidden [geheimnes] existence in the harmonious relationship of the parts to it."

Sentiments such as these make sense as an expression of the worldview that begins with organismic unity rather than mechanistic intelligibility. We should note also that the idea of unification as harmony is apt within biology because it presents unity as encompassing, and in fact requiring particularity and heterogeneity, in a way that geometrical ideals of simplicity cannot. Acoustic harmony is an accord that necessarily contains heterogeneity, a synthesis of differences, as does the harmonious interaction of

---

But natural phenomena, as Peterman (2021, 210) writes, exhibit "variety and complexity that cannot easily be captured by general laws but is clearly not chaotic." She argues that this particularistic kind of order is what world soul theories seek to account for. It is interesting to observe that this order-with-apparent randomness, or regularity-with-particularity, is what was identified above as a characteristic of complex systems.

18. The philosophy of Lord Shaftesbury nicely instantiates this connection:

> To be beautiful, according to Shaftesbury, is to possess 'Unity of Design'. A beautiful thing is beautiful because all its parts 'concur *in one*', because it has the 'Character of *Unity*', because it is '*a Single Piece*'. A beautiful thing 'constitutes a *real* Whole, by a mutual and necessary Relation of its Parts'. It is 'a *Whole*, coherent and proportion'd in it-self'. (Gill 2021, 13, references omitted)

In the third *Critique*, Kant sets such ideas to work in characterizations both of the living being and of the beautiful. Cassirer (1918/1981, chapter 6) emphasizes that the ideal of harmony is what joins the seemingly disconnected topics of that book and credits Leibniz as the primary source. As Beiser (2010, 32) relates, "There is a deep aesthetic strand to Leibniz's metaphysics. . . . Unity amid variety is order or harmony, which is the structure of beauty itself. Hence living force manifests itself as beauty, so that beauty is the measure of the power of a substance." See also Phemister and Strickland (2015).

creatures within the natural world.[19] Likewise, an organism is an "organised body" (Cheung 2006): it contains heterogeneous parts, whose diverse operations serve a mutual benefit.

The theology around this view is also quite different from the one encountered in the previous section, which was centered on divine legislation. Peterman (2021, 190) observes that Acts 17:28—"In Him we live, and move, and have our being"—is perhaps more often cited than any other piece of Scripture in early modern natural philosophy.[20] The positing of a "world soul" (*anima mundi*) is a way to express this. Robert Fludd went as far as to identify God with the *anima mundi*, which is to say that God is the ultimate whole and unifying principle. Others, such as Anne Conway, characterized the world soul as a mediator, a "Middle Nature" between God and creation. Again, there seems to be a connection with monotheism, the attraction of the idea of the world soul being that "it enforces the primordial, divine unity in nature's diversity" (Peterman 2021, 189).

Arguably, it is because the ideal of simplicity as unity is especially apt to the phenomena that biologists must deal with—ones that show purposeful order among particularity and heterogeneity—that the history of modern biology is not a straightforward story of the rise of the mechanistic framework. Instead, its path since the seventeenth century has been an oscillatory one, with alternating trends of reductionist and antireductionist movements, labeled by Canguilhem (1955/2015) as "mechanist" and "vitalist." Vitalism, it should be noted, is one way to characterize theories that posit unifying principles, such as souls, to explain purposeful processes within living organisms. Although many of the terms of "Romantic" biology, such as the *Naturphilosphie* of the late eighteenth and early nineteenth centuries, are not now deemed scientifically respectable, it is well recognized that those movements are part of a trajectory continuous with neuroscience as we know it today (Zammito 2018). For example, Goethe, while pitting himself up against Newton's theory of color (not respectable!), happened

19. Note that in music, the conception of harmony is mathematical due to certain ratios of string lengths and sound frequencies, but this feature does not transfer in the analogy with ecological organization, where the harmony is not conceived as having a mathematical basis.

20. The verse plays an important role in motivating Malebranche's occasionalism (Jolley 1997, xi), and Berkeley's idealism in the *Dialogues*.

to make important observations concerning the psychology of color sensation. Johannes Müller, noted for his findings on the reflex arc and the theory of specific nerve energies, was posthumously criticized for his vitalism. He was the mentor of a generation of students, including Hermann von Helmholtz and Emil du Bois-Reymond, whose life's work in physiology was to abolish the idea that the living world had any sui generis principles of operation (Otis 2007).[21] When the pendulum of fashion pushes back against organicist programs, it may well be because their offerings are less clear and rigorous than mechanistic, reductionistic ones and hold less promise for making nature intelligible through mathematization.[22]

In Aristotle's philosophy, the paradigmatic beings (*ousias*) are organisms with a unifying purpose (*telos*). This is indicative of the nonreductive character of his philosophy, in that it does not place its first metaphysical footing in the smallest parts of matter (elements or atoms), but rather in living bodies with structural complexity. Teleology, the deployment of purpose as a basic explanatory principle, has been thought of as an embarrassment of modern biology, although of course the concept can be reengineered in terms consistent with Darwinism (Mayr 1988). What we find in neuroscience today is that some of its simplifying ideas are centered around function and purpose, but this is undergirded by the notion of design in artifacts. Indeed, the opening gambit of cybernetics, an interdisciplinary movement important to the history of theoretical neuroscience, was the attempt to account for teleology in terms of simple feedback mechanisms that could be instantiated equally well in a body and a machine (Rosenblueth, Wiener, and Bigelow 1943). More commonly now in neuroscience, arguments for the importance of functional and top-down approaches are accompanied by the computer analogy. The impossibility of understanding how computers work just by examination of individual microprocessors,

21. They were advocates of physicalism in its original sense of a unity of science view (Sebestik 2011). The assertion was that the explanatory principles of physical sciences are universal and sufficient for all the other natural and social sciences. This denies the autonomy of the other sciences, including biology and psychology.

22. An example of the oscillation of fashion is to be found in chapter 5. Barlow's (1972) neuron doctrine was written at the high point of reductionism in visual neuroscience. We find neuroscientists more recently espousing the importance of multineuron "emergent" effects, as well as the need to factor in the animal-environment relationship when designing studies.

without reference to design and software, is brought up in various antireductive arguments (e.g., Carandini 2012). What this indicates is that the organismic notion of simplicity as unity with a common purpose among heterogenous parts has been altered to fit into a basically mechanistic and physicalistic worldview.

### 1.5.3 *Docta Ignorantia*

For the purposes of this study, we can take the brain to be infinitely complex. With its billions of non-identical neurons and trillions of ever-changing synapses, to concede that it is infinitely complex must be more realistic than to hope that it is approximately simple. In short, the brain is immensely more complex than any of the models and theories that could be simple enough to be intelligible to human scientists. In the fifteenth century, Nicholas of Cusa employed the term *docta ignorantia* (knowing ignorance) to refer to the stance that we should take toward a target of knowledge that far outstrips the capacity of our finite minds, which was for him God.[23] The point of tutoring one's ignorance is to gain awareness and clarity about what can and cannot be known in order to achieve insight into the limits of one's knowledge. Cusa applied this apophatic stance of knowing ignorance to empirical knowledge of the natural world alongside his theology (Hoff 2013). And it is the attitude I suggest we take toward knowledge of hypercomplex objects, "divine machines," such as the brain. Neuroscience, like all science, depends on simplification, so we must retain skepticism about the adequacy of models, theories, and analogies to encompass the fullness of what they aim at. The brain, as with countless things in the living world around us, may well be as far from what is simple and intelligible to our minds as a star, receding light years away, is distant from our view. What is captured by the model, like the tiny glimmer in the night, is only a minuscule, fractional drop from the immense source to which our sight is directed. What is simple, regular, and comprehensible in our picture of the workings of living nature is to be recognized as the product of human efforts to hold the infinite in a finite cup. This may not be all to the story, but it seems good enough to me, as a first approximation.

---

23. See Cusanus (1954). Also, see Cranz (1953) for a discussion of this point, including Cusa's borrowing of the term *docta ignorantia* from St. Augustine.

## 1.6 Overview

To some, the recommendation of knowing ignorance will sound like a counsel of despair—an encouragement to investigators to pack up their labs and give up on the slow but monumental task of helping neuroscience to progress. But this is to misunderstand the aim of this book, which is not to give policy recommendations to neuroscientists but to give philosophers and other interested parties some of the understanding that they need to fairly interpret neuroscience. If, as will be argued, brain complexity presents an insurmountable obstacle to there ever being one unified and general theory explaining how the brain gives rise to cognition, this limitation needs to be as widely known as the claims made on behalf of the more ambitious unifying programs of neuroscientific research. I am not saying that neuroscience is doomed to fail in all of its less ambitious goals, but it is necessary to interpret its local successes in the light of their coming about through use of an array of simplifying tricks. This provides a counterbalance to common, naive readings of experimental and theoretical results, which take them to be telling a straightforward story of how the brain just works, applicable to everyday cognition as much as to the special circumstances engineered within a laboratory or hypothesized in a model.

I grant that as a matter of psychological fact, neuroscientists themselves may need to be more committed to general unifying programs, more monistic in their conceptual outlook, and more optimistic about their ability to find order in the face of all this complexity than I think is justified by past history and current states of affairs. Those of us whose life's work does not rest on the belief that the brain is a *problem* that the collective, intergeneration efforts of neuroscientists will *solve* can afford to take an unprejudiced view of the feasibility of the more grand challenges. The following chapters should be read with this point in mind: the intention is not to change neuroscience, but to interpret it.

The next chapter will conclude part I of this book by setting out the core philosophy of science ideas at work in the case studies of part II. Given the complexity of objects of neuroscientific investigation, a pluralist and *perspectivist* approach to this body of knowledge is to be preferred over a standard scientific realism, which takes it that the best-established scientific theories are approximately true representations of systems in nature. Standard realism is not possible, I argue, when those systems are so complex

that any scientific categories and concepts employed to describe them are the result of active construction of simple patterns and regularities to which the theoretical terms refer. The kinds of heterogeneity discussed in this chapter show how the brain by itself does not offer a determinate catalog of neuronal types and organizational hierarchy. There are multiple, justifiable ways of structuring this complexity, and the indeterminacy of typing motivates the shift from *formal realism*, the view that structures represented in the best models exist independently of the models, to *formal idealism*, the view that those structures depend on the schematizing efforts of the scientists as much as on the brain itself.

Cautions about the risks of simplification have appeared a few times in the history of the neurosciences. Ramón y Cajal (1937, 302–304) wished "to warn young men against the invincible attraction of theories which simplify and unify seductively." As a psychologist and philosopher, William James was also sensitive to this concern:

> The theorizing mind tends always to the oversimplification of its materials. This is the root of all that absolutism and one-sided dogmatism by which both philosophy and religion have been infested. (1902, 27)

Chapter 3 will be about an episode in early neuroscience, dating to around one hundred years ago, when a simple theory of the organization of brain and nervous system came to dominate the field, only to be discarded shortly afterward. The dramatic fall of the reflex theory perhaps serves as a cautionary tale—what simplification gains in immediate appeal is lost when achieved at the expense of adequacy to observable facts.

Chapter 4 is about the theoretical framework that succeeded the reflex theory, resting on the idea that the essential neural processes giving rise to cognition are computational. The argument will be that computational models of the brain serve as simplifying analogies, dependent on the scientist's selective perception of similarities between neural systems and devices doing somewhat comparable tasks. Although a computer is itself a very complex artifact, it is far less complicated than even a small invertebrate brain; and because it has been consciously designed and manufactured by people, there is awareness of its operating principles. By assuming that the brain is a fleshy computer, the neuroscientist gains some explanatory traction in the face of otherwise uninterpretable neural signals. The philosophical lesson of the study, though, is that we should not fall into the habit of thinking that

the brain is literally a computer, since the disanalogies between these two kinds of things can not be discounted.

Chapter 5 is about the relationship between experimental practice and modeling techniques in computational cognitive neuroscience. Scientists measure neural activity during specific behavioral tasks and produce models that interpret the activity as carrying out certain computations, offering explanations of the brain's involvement in the cognitive performances. The computational models are abstract and idealized representations of neural activity, but they also depend on simplifications introduced early in the experiment-to-model pipeline. I argue that the simplifications introduced in experiment and data processing play a crucial role in allowing the modeler to build a mathematical representation of the neural activity that is simple enough to be interpretable.

Chapter 6 takes up the contested topic of neural representations. Theories in cognitive neuroscience often posit representations in the brain, and their status is controversial, generating much debate over whether they meet the criteria for genuine representations or if use of this term is only a misleading metaphor. I offer a new interpretation of the theory and practice, arguing that the positing of neural representations helps to simplify brain research by licensing scientists' focus on the relationships between neural activity and events, while neglecting processes within the brain and peripheral nervous system.

Chapter 7 is an application of the perspectivist ideas presented in chapter 2. The motor cortex is an area of the brain that has been the site of quite wide disagreement about its basic function and operating principles. One important theory of the motor cortex is a computational and representational one, which posits that individual neurons in this region represent patterns of muscle activations or other movement parameters. Another approach based on *dynamical systems theory* has challenged the basic assumptions of the representational theory. These two perspectives on the motor cortex are in some respects complementary. They each employ a different set of abstractions and idealizations that have their own justification. One can be ecumenical, so long as neither framework is interpreted as offering the final word on the workings of the motor cortex. Given the changeableness of the brain, its Heraclitean nature, it cannot be theorized as it is "in itself," independent of quite drastic simplifying assumptions. That is the conclusion of part II of the book.

In part III, I go on to consider the implications of these studies. Chapter 8 takes up the issue of whether there are limits to what neuroscience can achieve through its simplifying procedures. The prompt for the question is the introduction of machine learning methods for modeling neural systems, in which greater predictive accuracy is achieved through less reliance on idealization (although these new models are still very simple, relative to the actual brain). The catch is that the models themselves are so complex that they cannot be readily comprehended by the scientist. If we think of science as the project of attempting to understand natural systems, while at the same time devising means to predict and control them, it does seem that it reaches a limit to its ambitions in neuroscience: the brain is not simple enough to be simultaneously understood and controlled.

The acknowledgment that computational models of the brain, including the most advanced artificial neural networks, are highly simplified and neglectful of countless neural facts has implications for how claims for AI should be interpreted. The idea that consciousness and general intelligence can potentially be replicated in inorganic machines depends on the assumption that there is a common computational structure shared between the brain and the silicon-based device. However, the analysis of neurocomputational models given in chapter 4 showed that the usefulness of the brain-computer analogy does not rest on there being such a structure in common. In chapter 9, I argue that this gives reason to be doubtful of the promise of truly intelligent AI.

Chapter 9 develops an argument for "biological naturalism," the view that consciousness and general intelligence are capacities that animals have that depend on the living, material constitution of their nervous systems as they operate within the rest of the body. The point is that an artificial device, lacking the material complexity of a living system, will not develop these capacities. In chapter 10, I further explore the topic of embodied cognition. Through a close reading of Herbert Simon's argument in favor of replication of cognitive function in digital computers, we will see that this idea of the multiple-realizability of cognition itself rests on a set of simplifying assumptions that treat mind, brain, body, and environment as systems that operate quasi-independently of one another, such that they can best be theorized in isolation. These *Cartesian idealizations*, as I call them, may find justification in the pragmatics of science, but they have been deeply misleading when taken up as theoretical commitments within philosophy of mind.

The conclusion of the book is that greater awareness of the simplifying schemas imposed by necessity within science must lead to greater commitment for philosophy of mind to pursue its inquiries in an autonomous manner. Philosophical views about the mind should not be read off from the science of the brain because the scientific demands for simplification, manipulation, and control slant the results of investigation in ways that are antithetical to the aims of philosophy. For this reason, naturalistic programs in philosophy of mind have a task, so far neglected, to explain when and how scientific results can safely inform specific philosophical opinions. My hope is that this book will prompt a few readers to make efforts in this direction.

# 2 Footholds

The task of this chapter is to articulate the epistemological principles that will be at work in the rest of the book. I am presenting my framework as if it were an established entity rather than arguing for it systematically by pointing out its dialectical advantages over rival frameworks, and so forth. This is to make the chapter shorter than it would be, and to avoid distractions, for really, the vindication of these principles is to come about through their application in the body of the text. Those results, I hope, will show that this is a framework worth having when trying to make sense of the achievements and limitations of neuroscience. Regarding the scope of the account, it is motivated by the issues that arise with neuroscience and needs some adjustment to be applied to sciences such as cosmology and paleobiology, which do not involve interventions on their objects. However, I think that the claims about the structures represented in scientific models and theories do generalize beyond neuroscience, especially the sciences of extremely complex systems.[1] Hence, for ease of exposition, I present this as a general philosophy of science, with the caveat that it will need modification to be applied to some of the sciences. This situation is no worse than with most of the classic general philosophies of science: they were tailor-made to physics and could not be generalized without some tweaking, if at all.

The reason for devoting this chapter to these foundational issues is to indicate to the reader how my starting assumptions—these initial footholds—are arranged quite differently from what it is usually found in the philosophy

1. Falkenburg (2007), Lenk (2017), and Chang (2022) present somewhat similar accounts that draw from studies of the physical sciences.

of neuroscience, especially with regard to scientific realism. It is good to call attention to what's different at the outset in order to offset some misunderstandings later. This chapter may be of less interest to readers who are not academic philosophers. The case studies and core arguments of parts II and III should be intelligible even if the reader is not familiar with this chapter. At relevant places in the subsequent chapters, where this framework is being employed, I will refer back to some of the sections here.

I am presenting an alternative to scientific realism known as *haptic realism*. While scientific realism asserts that the best-confirmed theories offer approximately true representation of how things stand in nature, haptic realism insists that the acquisition of scientific knowledge is an active process in which the scientist's schematization and the work that goes into shaping the material target of research leave an indelible imprint on scientific knowledge. This means that scientific representations, theories, and models of systems in nature should not be interpreted as approximately true accounts of those things as they are in themselves, independent of interaction with the scientist.

The idea is that there is an incoherence in the presumption of standard scientific realism that scientific knowledge, at its best, could deliver an account of nature in itself, purified of any input from human knowers. This is an appropriation—not a replication—of Kant's transcendental idealism, and it is motivated by the considerations of complexity that arose in chapter 1. The conclusion there was that whatever regularities and patterns exist in nature, most obviously in the living world, they are in themselves vastly more complicated than can be mirrored within the finite minds granted to human beings. Therefore, it is necessary to give up on the conception of the mind as the passive recipient of knowledge, a mirror of at least some of nature. The task of knowledge formation is to rig up some fit, some gearing, between what amounts to the infinite complexity of natural systems and the limited workings of human cognition. Such a fit can be achieved only by the ordering, structuring, and simplification of whatever patterns are suggested via empirical observation.

The difference between standard scientific realism and haptic realism turns on a point of metaphysics, which is ultimately theological. In Leibniz's philosophy, the universe runs on a rational plan. It is prepackaged by God such that it is (at least in part) cognizable to the finite human mind. Indeed, Leibniz followed a Platonic tradition in linking scientific and mathematical

knowledge to divine understanding—they are treated "as an approximation to divine thought" (Martin 1951/1955, 7). Gottfried Martin here argues that Leibniz is the point of departure for Kant's philosophy, beginning with a suspicion that Leibniz's account of being and knowing is rather too optimistic.[2] The result of Kant's inquiry is that the finite rational knower has the job of doing the packaging that makes nature intelligible. Simplicity in nature, which is to say systematicity, unity, and order, turn out not to be discoverables, but demands of human reason. For example, Kant proposes that the search for a more and more unified account of forces—we may note here that research on a "grand unified theory" in physics is an ongoing case of scientists' striving after simplicity in nature (see section 1.5.1 of chapter 1)—is mandated by "reason's logical principle," which, "calls upon us to bring about such unity as completely as possible" (Kant 1787/1929, A649/B677, quoted and discussed in Longuenesse 1993/1998, 151).

This chapter argues that science always presents its object through a schematizing medium—a set of formal concepts, a quantitative model, or an evocative analogy. Because nature is so complex, it affords multiple justifiable ways of representing each of its inhabitants. This motivates the approach known as *perspectivism* in philosophy of science: there is a plurality of possible, empirically confirmed scientific views on any given object of investigation. The existence of one perspective does not by itself conflict with or invalidate another, but neither can claim to deliver the absolute truth about their subject matter.

Furthermore, scientific perspectives are differentiated by the various aims of inquiry. Neurobiologists aiming to treat memory loss will develop an account of the hippocampus that is incongruent with a theory of the same brain region produced by researchers in a machine learning collaboration, working on artificial cognitive agents. Neuroscientific knowledge is not absolute—an approach to how things stand with the "brain-in-itself"—nor

2. Martin (1951/1955, 60–63) later discusses Kant's case against the "theological foundation of truth" presupposed by Plato, Augustine, Malebranche, Newton, and Leibniz. Kant's argument comes through in the antinomies of the Transcendental Dialectic in the first *Critique*:

> If it is true that the world is both finite and infinite, that it contains both atoms and matter which occupies space continuously, the never-ending conflict of these antinomic characters of the world could not come from God's thinking. They must in principle come from human thinking. (62)

is it disinterested. Haptic realism is committed to the view that scientific knowledge is as much about creating effects as it is about understanding, that representing nature scientifically, and intervening in it, are two sides of the same enterprise.

## 2.1 Haptic Realism/Formal Idealism

> Language can be literal, science certain, in an Aristotelian world, but in no other. For in no other world do things speak to us, of themselves, each in its own kind, without our first invoking and evoking them. We, as Kant showed us, must by our categorizing contribute to the making of the world we know.
>
> —Marjorie Grene (1963, 238)

Scientific realism asserts that the world has a determinate, mind-independent structure; that scientific theories are to be interpreted literally, as truth-apt representations of their target domain; and that well-established and predictively successful scientific theories are approximately true representations of their targets such that observable and unobservable entities posited by those theories actually exist.[3] For example, the well-established and predictively successful theories of nuclear physics posit that there are such entities as protons and neutrons comprising the nucleus of larger entities called *atoms*; these particles are part of the furniture of the world as it is, independently of scientific activity and conceptualization; this account is to be taken at face value, as a description of these unobservable particles, and given the maturity and accuracy of the theory in predicting observable data in physics experiments, it should also be credited as being approximately true. Scientific realism has been the dominant stance in postwar Anglophone philosophy of science, including philosophy of neuroscience.[4] It is an unreflected background

3. This sentence paraphrases the three theses, *metaphysical*, *semantic*, and *epistemic*, which jointly make up scientific realism, according to (Psillos 1999, xix). One qualification made by Psillos is that the entities said to "inhabit the world" need only be very similar to those posited, not exactly like them.

4. For overviews of what is now a wide collection of theories, see Psillos (1999) and Chakravartty (2007). The dialectical opponent of scientific realism is normally taken to be logical empiricism or instrumentalism,rather than the kind of Kantian view presented here. However, Kuhnian constructivism, which was itself very influential, can be placed in the Kantian tradition.

assumption of many other branches of philosophy today, not least naturalistic philosophy of mind. The root conception of scientific knowledge, at work in realism, is that of a mirror being held up to nature: the world is out there, and it is the task of science to receive an image of it. If it is a simplistic view of scientific knowledge, this stems more fundamentally from a belief in the simplicity of the world to be discovered by science. It is assumed that the wealth of things and processes and events in nature is well demarcated and regular, independently of people and their science, and these predefined structures are of a straightforward enough kind to be absorbed and comprehended, without distortion, within the necessarily limited scope of collective human cognition. While it is not assumed that any theory yet devised has in it an image of nature in its totality, the approximately true ones are thought to be piecemeal approaches to a completed picture. It is consistent with scientific realism that science *could* one day deliver a univocal, all-encompassing, and true representation of the natural world.[5]

The alternative to scientific realism is best initiated by rejecting the visual metaphor that grounds it. Nature is immeasurably complicated, and scientific knowledge is acquired not through passive absorption, but rather by actively grappling with things, cutting them down to manageable size.[6] The alternative conception is centered on the metaphor of touch. A key feature of touch is that the fact of contact or some kind of interaction between skin and the perceived object is undeniable. To explore the world through touch, we must move around, reach, and grasp things in particular ways, leaving our traces on those things. In addition, sensing by touch is more often than not linked to the performance of some deliberate action. You might tap and rotate a melon to gauge how hard it is, the best angle to cut into it, and how much force to use to do so. These are characteristics of touch that, according to Jonas (1954, 514), most distinguish its phenomenology from that of vision:

5. Nancy Cartwright (1999) would be an exception here: she is a scientific realist who does not accept the fundamental simplicity of nature. Her world is instead "dappled," and representations of it must also be.

6. See Teller (2018) for another argument to the effect that acknowledgment of the complexity of the world is incompatible with standard scientific realism. There is also overlap with the constructive nominalism of Elgin (2019).

> The very coming into play of this sense already changes the situation between me and the object. . . . We therefore do not have in touch that clear separation between the theoretical function of information and the practical conduct, freely based on it, that we have in vision.

The hand is both the primary sense organ for touch and our foremost means for affecting changes in the world—the root of the word "manipulation" is the Latin for "hand." I suggest we think of scientific practice—and the theories and models that spring from it—along the very same lines. In other words, we should reject the traditional realist's conception of knowledge attainment as the picturing of objective facts, with its ideal of disinterestedness. Scientists learn about the world through tinkering and interacting with it, and these learning practices are bound up with their practical intentions. They learn about their objects and systems because of their haphazard and human-centered engagements with things, not in spite of them.

I have called this picture "haptic realism" (Chirimuuta 2016, 2023c).[7] It is a kind of realism simply because it grants that knowledge has a basis in a world beyond the scientist, and the social networks that the scientist inhabits. This is the minimum commitment of realism. However, it rejects the usual claim of realism to there being, in the best cases, some epistemic relationship with a stratum of being that is entirely human- and mind-independent. I suggest now that this version of realism can also be taken as a kind of transcendental idealism. It grants that things exist beyond the human mind, but it also holds that we cannot know them as they are in themselves, as they are not in respect to relations with the human cognizers.[8] In short, the assertion is that

---

7. The 2016 paper discusses precedents for the account, including Helmholtz and pragmatist philosophers of science such as Ian Hacking (1983) and Hasok Chang (2012). A similar account is the *scheme*-interpretationist scientific realism of Hans Lenk, who also refers to transcendental idealism and employs the haptic metaphor quite frequently:

> Any "graspability" whatsoever is interpretation-laden. The world is real, but (any description and action of) 'grasping' the world is always interpretative, i.e. only conceived of and formed by scheme-interpretation. It is furthermore internally action-bound and deeply societal. (Lenk 2017, 274; and also see Lenk 2019)

His reference to "schemas" is comparable to my use of "forms.".

8. My uptake of transcendental idealism has been informed by the "modest metaphysical interpretation" of Allais (2015). In my first presentation of haptic realism (Chirimuuta 2016), I made heavy use of Giere's (2006b) analogy between color vision and perspectival scientific knowledge. When we come to identify an object by learning its color, our knowledge depends on a relational property of the object (i.e., how it affects

what the knower brings to the interaction from which scientific knowledge comes about is ineliminable; the goals, idiosyncrasies, and constraints that flow from the knower cannot be removed from this knowledge by a process of purification. Science is not an absolute knowledge, a knowledge that has no relation to the human condition, a representation of mind-independent nature as it exists regardless of the presence of the minds that represent it. To aspire to empirical knowledge of things in themselves is to subscribe to an incoherent account of knowledge, which is to say, one that leads to skepticism.[9] That, to me, is the point of transcendental idealism.

### 2.1.1 Kant's Hylomorphism

Kant glossed his transcendental idealism as a "formal idealism" (1781/1787/1998, B519n) to distinguish himself from the Berkeleian idealist for whom all that exists is sensation or spirit of some sort. By taking up this notion of formal idealism, it will be all the more clear how my proposed framework is to rest on the considerations of complexity and simplification that were introduced in chapter 1.[10] The relevant notion of "form" here is the one that pairs with "matter" in the hylomorphic theory made famous by Aristotle.

---

our spectrally sensitive visual system). Generalizing this, one can think of all knowledge as restricted to the properties that relate things to human beings and their instruments. Indeed, it is incoherent to claim empirical knowledge of something except by its affecting you or bearing a relation to you in some way. Hence, knowledge of objects as they are in themselves—as bearers only of properties unrelated to knowing subjects—is ruled out. The secondary-quality analogy for transcendental idealism, offered by Kant in the *Prolegomenon*, forms the basis of Allais's reading. A valid concern about transcendental idealism is that it demands a fundamental division between the human knower and the ultimate, nonhuman grounds of empirical knowledge. So while I take this Kantian approach to be the best way to account for the possibility and limitations of exact, scientific knowledge of nature, noting that science is deeply informed by the supposition of a division between the human knower and the known world, I do not take it to be conclusive on the matter of knowledge more generally.

9. Here, we can take Kant's epistemology to be a response to the Cartesian skeptical predicament. If one makes immediacy a condition of knowledge—lack of an interactive process by which the subject engages with the object of knowledge—then it turns out that what one may know without doubt of its existence is restricted to the content of one's own consciousness, which is presumed to be known immediately; hence, there is doubt about the existence of a world beyond the mind.

10. Indeed, my treatment of the notion should be understood as a taking-up, or appropriation, for new philosophical purposes. Hence, the exposition will be cursory and superficial, in comparison with what should be expected in a work aiming at historical

Aristotle was a formal realist in the sense that he took forms to be there in the world, not inherently related to anyone knowing them. Every natural substance, such as a particular organism (e.g., a pine tree growing in the park), is a composite of form and matter. Hylomorphism, though not confined to living beings, speaks to the problem of how to make sense of the stability and identity of organisms in spite of the flux of material that passes through them as they grow and ingest. Form, while inseparable from the matter of the entity and not a distinct part of it, is what orchestrates the matter into maintenance of the persistent organism. As Grene (1963, 120) relates, "being must control becoming." Movement and change of matter in the natural world are always directed toward the fixed points that are the forms.

It is because of the forms that the natural world is, for Aristotle, in a deep way intelligible, which is to say that one can come to know the essences of things. The scholastic maxim was that *forma dat esse rei*—the essence of the thing resides in its form.[11] A further way to consider the connection between form and intelligibility is to note that an Aristotelian substance is a self-standing entity, and this determinateness is due to its having a form. Something with a form is delimited, bounded; it is set out from the rest of all that is, and knowable independently. As readers of the critical philosophy will soon encounter, the form/matter distinction is employed frequently by Kant. But form is there attributed to our faculties working on the matter of sensation (Pippin 1982, 30–39). It is the determining, delimiting activity of the mind that gives form to the matter that comes from our sensory organs.[12] "Form," as Pippin (1982, 13) relates, "is not itself an object of knowledge but a 'condition' for knowledge." Thus we have the basic difference between formal realism and Kant's formal idealism.

---

scholarship. See Pippin (1982), Longenuesse (1998), and Boyle (unpublished) for proper discussions of Kant's hylomorphism.

11. Kant's commentary on this is discussed by Pippin (1982, 12) and Boyle (unpublished, 1).

12. A detail to be noted here is that the form/matter relationship does not occur just at one level, referring to the contribution of the *understanding* as opposed to sensations delivered from *intuition*, but rather at multiple levels, to invoke what it is that thought is doing (namely, determining its object): "All thinking is an activity of *determining* (giving *form* to) a *determinable* (*matter*)" (Longuenesse 1998, 148; emphasis in original).

An intuitive illustration of this idea comes from considering the perceptual constancies.[13] The proximal stimulation of our sensory organs—the response of our retina or hair cells to light and sound—is an indeterminate flux of heterogeneous, ever-changing activity. Stable objects are not just given to us at the point of sensory transduction. The instability of basic visual sensation is somewhat apparent if we force ourselves to take the "painterly eye" and notice all the variation in light and shadow and shades of color that the visual world contains. Sensory experience affords knowledge insofar as it can be given form, which means that similarities and regularities in the flow of sensation are appropriately categorized as recurrences of the same object. The visual system is endowed with light and color constancy, which achieves exactly this shaping of indeterminate proximal stimulation into determinate, stable, and separate objects that are then taken to be the distal causes of your raw sensory responses.

There are some parallels to be drawn between this perceptual example and the generation of scientific knowledge, in this Kantian account. But it is important to note a difference. With a case like the color-constant perception of a book on a table, bathed in dappled light, common sense takes it for granted that there is a stable, persistent object, the book, which is the distal cause of the unstable sensations and is recovered in constant perception. The view about science that I encourage the reader to entertain is that it should not be taken for granted that the regular, persisting patterns resulting from the process of scientific investigation are simply there in nature, conditioning the investigation and being recovered when the process is successful, as opposed to the stability and regularity being necessitated by the activity of scientific thought (individual and collective) when it works on observational data that call for further determination. As in the perceptual case, the determining work of cognition is required for there to be an experience of the world as stable and regular. But to subscribe to formal idealism about scientific knowledge, we must decline to think of the achievement of stable and regular objects of knowledge as amounting to the recovery of some preexisting patterns and objects.

13. Cf. Burge (2010), who argues, along Kant-influenced lines, that the capacity for constancy is what first enables any animal to perceive the world as containing discrete, stable objects, as opposed to experiencing a flux of sensations. Again, nothing turns here on whether this is true to Kant's original philosophy.

A few observations will show why the more radical stance is actually quite apt once complexity in nature is fully appreciated. In chapter 1, we saw that complex systems present an ever-changing array of seeming regularities amid apparent randomness. Among living things, no phenomenon ever exactly repeats, and no organism stays precisely the same. The brain offers an extreme example of this complexity and changeability. Scientific categorization can be analogized to putting a low-pass filter over a subtly varying pattern, highlighting the somewhat regular repetitions, masking differences and irregularities, thus yielding kinds and lawlike behavior among the phenomena. Under Aristotle's formal realism, it was assumed that the ordering presence of fixed forms was a feature in nature, waiting to be known. But as Marjorie Grene points out, we are not dealing with Aristotle's world, which was a simpler one, since beings were fixed. Contemporary scientific realism is in its own way committed to formal realism, the assumption of a recoverable simplicity in nature, exemplified by "natural kind structure," or "joint-cutting" types of beings, underlying the flux of particularities. Yet our world is heterogeneous, flowing, complex.[14] In other words, if our starting point is the acknowledgment that the natural world, which is the target of scientific knowledge, is vastly complex, it is better to conclude that the high regularity and simplicity found within scientific representations are the result of regularization and simplification, not the discovery of order hidden beneath disordered appearances. Formal realism supposes a preexisting simplicity in nature, and absent that assumption, formal idealism is the route to take.

I will finish here with some refinements and caveats. First, it is important to appreciate that the form-giving role of scientific thought is not just that of filtering, letting through some details of the data and leaving out the rest (cf. Danks 2020, 129). Rather, it is an active shaping of patterns that are only enchoate in the data (hence determinable) and could be formed

14. Cf. Grene (1963, 237): "Can we not . . . have Aristotelian predication in a flowing world? Of course not."

I am asserting here that changeability, the temporal heterogeneity of the living world, is manifest to observation, just from examining what organisms are like, and does not depend on inference from sophisticated theories. So it is not just that species are mutable, as is to be inferred from the theory of natural selection, but that changeability is a pervasive and readily apparent feature of living systems. This is the lesson of "process biology" (Dupré and Nicholson 2018), and also see chapter 7 on how neuroscience instantiates this lesson.

in various alternative ways. The path that the determination takes is itself shaped by the goals of inquiry. The various practical activities that a scientific investigation is related to will influence the concepts employed both in the production and interpretation of data, and therefore bring about a distinctive perspective on the target of investigation. More will be said on perspectives in a moment. Second, I do not mean to claim here that all the regularity presented in scientific models and theories is an imprint of these formal determinations. This notion of "imposition" is the wrong way to read haptic realism. I refer again to interaction, which is the guiding idea. With these scenarios of cyclical interaction, it is impossible to extract the contribution of the knower, on the one hand, and the target of knowledge, on the other. Thus I am not asserting that world is formless, absent determination by human cognizers.[15] What I am saying is that the forms and

15. An absolutely formless world would be one in which there were no systems or entities delineated against one another. This would be a world of homogeneous mutual dependency, in which no thing is differentiable from its background context. Merleau-Ponty (1942/1967, 43) remarks that it is a condition of the possibility of science that the world is not like this: "If everything really depended upon everything else, in the organism as well as in nature, there would be no laws and no science."

He follows with the interesting observation that different traditions of science have taken different positions on the spectrum between conceiving of nature and individual organisms, as integrated, undifferentiated unities, and as collections of sharply delineated structures and substructures. The former view is expressed in post-Kantian Romantic biology, and the latter in the mechanistic physiology of the reflex, which will be the topic of chapter 3. Gestalt psychology, he notes, is in an intermediate position between these two extremes.

It seems to me that the truth is indeed likely to be in the middle ground. The critique of mechanistic physiology, as we will see in chapter 3, rests on the point that it exaggerates the sharpness of delineations between subsystems of the animal through experimental interventions that physically separate one part of the body from the other, procedures such as lesioning parts of the nervous system. This is a way to bring about determination, so it is a simplifying strategy. Yet, says Merleau-Ponty, what is revealed is an organism in a pathological state. Belief in the simplicity of nature—which is manifest in the view that I call *formal realism*—engenders the mistake of thinking that the constructed, pathological state is a faithful model for how the organism operates when left to its own devices. To digress further, it is worth asking how many of today's ecological crises involve precisely this underestimation of the integrated operations of nature as a whole. To think that nature comes ready-made into isolatable modules for research and manipulation is to grant it a high degree of inherent simplicity and intelligibility. But if this is the wrong assumption to make, people will find themselves caught on a perpetual treadmill of performing what they think

regularities that are simple enough to be intelligible to the scientist, and represented in their models and theories, should not be taken to exist independently of this work. These regularities, the products of this interaction, are what I will refer to in chapters 4 and 5 as *ideal patterns*, as opposed to real patterns, the ones assumed to exist independently.[16] Third, given people's interests and goals, not any simplification will be helpful. The challenge of science is to find simplifications that will work as desired. As such, the ideal patterns that are accepted and established by the scientific community are constrained by inherent features of the target system in conjunction with the goals of research.

### 2.1.2 Perspectivism

Mary Hesse, with reference to Gottfried Martin (1951/1955), makes Kant the originator of the view that natural science can do no more than present the world through the medium of models and loose analogies:

> It is plausible to interpret his [Kant's] attitude to theoretical science. . . . in terms of possible models which are ways in which we think about the world, but which may not and sometimes cannot be literal descriptions of the world, and in any case can never be known to be such descriptions. This attitude leaves the scientist free, as Kant remarks in connection with theories of void or plenum, to adopt the theory which is convenient. (Hesse 1962, 172)

We see here a close connection between the rejection of the scientific realist's aspiration for theories to be literal, true descriptions of their targets, and a more liberal view, which grants that any one subject matter may afford representation with a range of suitable theories. The result is a *perspectival pluralism* that I characterize as the claim that there can be, within science, multiple research traditions and modeling approaches that all target the same systems, but rest on incompatible simplifying assumptions and result in unrecognizably different descriptions of their domains. The differences between perspectives are conditioned by historical and cultural factors, as well as pragmatic factors such as the convenience of use of one formalism over another.

---

are local interventions, which actually have further consequences well beyond the intended domain of impact, and are not predictable because they fall outside the scope of scientific awareness—outside the set of factors that their implicit metaphysics tells them *could* be relevant.

16. See Chirimuuta (2023) for the comparison of ideal patterns to the "real patterns" of Dennett (1991). These are all indeed "quasi real."

Perspectivism is a popular approach within current philosophy of science. It is a broad church bound together by the thought that scientific knowledge is *situated* within a historical era and social context (Massimi 2018c, 164). In other words, scientific knowledge comes from a perspective and is not a view-from-nowhere, in spite of the attraction of the idea that science involves a transcendence of what is anthropocentric in the experience and understanding of nature, as well as the hope that at its best, science obtains a view on reality that could claim universal assent, even among alien beings. We should note how the more modest approach to science is reinforced by acknowledgment of complexity and the need for simplification. Knowledge of the sort apprehended from the God's-eye view would be possible only if the human mind were infinite or if nature were so simple that it could be grasped in its entirety by collectives of finite minds. The first supposition is obviously not tenable, and the second may have seemed plausible on the heels of rapid advance in the theorization of relatively simple physical systems, but its likelihood fades the more that scientists aspire to comprehensive theories of complex living things. The ubiquity of simplification in science is the ever-present mark of its being the activity and product of limited, situated beings. God, we may conjecture, would have no need for abstraction and idealization.

The version of perspectivism advocated by Massimi is far closer to scientific realism than the one I advance.[17] Massimi's perspectival realism accepts the idea that there is a human- and mind-independent world, and truth involves correspondence with it—in Massimi's terms, there are "perspective independent facts" (2018, 170–171, n2). Furthermore, she endorses the scientific realist's semantic tenet that the language of science should be interpreted literally, a view that I will criticize in section 2.2.1 and chapter 4. According to Massimi, it is a mistake to entertain the idea that perspectives in some way shape facts. However, once it is clear that this shaping has to do with the determination of categories and regularities among the many that are suggested by observational data—data that are frustratingly ambiguous and plurivocal, only yielding discernible patterns after pruning and massaging—then it is also apparent that rejection of the realist metaphysics is not some off-the-wall constructivism. All it does is call into question the "ready-made reality" of knowable objects (Chang 2020,

17. See Massimi (2022) for a comprehensive statement of perspectival realism. See Chirimuuta (2020b) for further discussion of the points of disagreement.

22), and propose that the acquisition of knowledge is as much about structuring as it is about the detection of structures.

At the start of his *Philosophy of Symbolic Forms* (vol. 3), Cassirer dwells on how compelling is the feeling that if knowing is mediated by forms that the subject brings with them, knowledge should still aspire somehow to get at how reality is independently of those forms:

> It would seem as though we could apprehend reality only in the particularity of these forms, whence it follows that in these forms reality is cloaked as well as revealed. The same basic functions which give the world of the spirit [Geist] its determinacy, its imprint, its character, appear on the other side to be so many refractions which an intrinsically unitary and unique being undergoes as soon as it is perceived and assimilated by a "subject."
>
> . . .
>
> Again and again, the basic drive of knowledge makes itself felt: the drive to unveil the veiled image of Sais and behold the naked, unadorned truth. The philosopher desires to apprehend the world as an absolute unity; he hopes ultimately to break down all diversity, and particularly the diversity of symbols: to discern the ultimate reality, the reality of "being" itself. (1929/1957, 1)

Still, it is axiomatic for Cassirer that all knowledge is mediated by "symbolic forms," that to philosophy, the "paradise of immediacy is closed" (Cassirer 1929/1957, 40). This rejection of the immediacy aspired to in the correspondence model of truth will be axiomatic for this study of scientific knowledge, too. So how to get rid of the aspiration for scientific knowledge that it should at least aim to correspond with "the reality of 'being' itself," unadulterated? Once again, shifting from a visual to a tactile metaphor is a helpful way to buttress against traditional realist instincts. The metaphors of perspectives and refractive media are, of course, visual. It pays to supplement these by thinking about how this knowledge is touchlike. What comes across is that we can know the empirical world only by virtue of being there, part of it, being able to pick things up and play about with them. Thus, it is not plausible to set up the world as it is independently of the human situation, our actual engagement with it, as the object for empirical knowledge.[18]

18. I am talking here of what is plausible in a philosophical interpretation of scientific knowledge. I grant that it may be expedient for scientists to aim at a very high standard of objectivity in their representations, even if they can never transcend their human situation. In many cases, it may be a good research heuristic to aim at theoretical unification and not settle for a plurality of perspectives, even if full unification is not a realizable goal.

Under haptic realism, scientific theories and models are handlike. Hands are both sensory receptors and motor effectors. And it is not just that the hand is a multitasking device, switching between these two operations. Its sensory and motor roles are intertwined—how the hand senses is molded by what it needs to do, and vice versa. In the same way, the epistemic and the instrumental roles of scientific theories and models mutually condition one another. There is no such thing as pure, disinterested science because there is no way to extricate scientific knowledge from its relation to practical purposes. This reinforces the point that different traditions of scientific research (i.e., perspectives) coalesce around different aims of inquiry, though in some cases, practical problems will foster integration of perspectives (Mitchell 2020).

In its formulation by Ron Giere, perspectivism was intended as a via media between "objectivist" scientific realism, which claims that theories can in principle provide "a complete and literally correct picture of the world itself" (Giere 2006b, 6), and a constructivist antirealism that asserts that "scientific claims about any reality beyond that of ordinary experience are merely social conventions" (Giere 2006a, 26). Likewise, in my view, science can neither deliver an objective, univocal truth nor is it just conventional in its representation of things beyond ordinary observation. However, a via media, it must be said, is a path on the same plane as the two ones that it goes between. To end this part of the chapter, I will now discuss how the proposed view actually stands out in greater relief against scientific realism and its traditional antagonist, empiricist antirealism.

### 2.1.3 Standing Out against Scientific Realism and Empiricism

Most presentations of scientific realism set it out only in relation to the tradition of logical positivism, empiricism, and instrumentalism, made famous by the Vienna Circle, and against which postwar realism reacted. This is unfortunate because restricting ourselves to these two options has constrained the discussion of the questions of what scientific theories and models tell us about the world and about how they should be interpreted. The standard dialectical plane can be envisaged as having one axis, the realist one, with which it is asserted that the representational content and truth aptness of scientific theories outstrip what is given observationally, and another axis, the empiricist one, by which it is asserted that the empirical given—the observed data—suffices to fix the representational content and give the truth conditions for a theory. With this two-dimensional plane in mind, it falls

to perspectivism (insofar as following the middle path) to accept different strands of both the realist and empiricist positions, thus taking up the space in between them—for example, by endorsing the realist proposal that the representational content of scientific theories outstrips the empirical given, while siding with the empiricist in setting the criterion of truth of the theory as no more than empirical adequacy.[19]

However, it needs to be appreciated how much scientific realism and empiricism have in common. Their shared starting point is a normative picture of science being an absorption of natural facts; they differ over whether they restrict the facts to empirical observations (empiricism), or whether they take those facts to be the unobservable states of affairs that are, in the best case, represented by mature theories (scientific realism). It is a conception of the task of knowledge formation in which it is at its best when it is most passive: science succeeds in its epistemic goals when either the empirical given or unobservable reality impresses on the theory so that scientific knowledge can conform to its objects. The Kantian alternative emphasizes, instead, activity and the constructive engagement that brings about knowledge. Science, through interaction with things in nature, works to construct objects and patterns that conform to its demands. This third axis, which asserts that knowledge formation is active, is orthogonal to the realism-empiricism plane.[20] Thus we can appreciate that versions of the Kantian alternative—for example, the haptic realism that I am endorsing—can be as different from both realism and empiricism as those views are from each other. The basic position is that it rejects the passivity that goes with the correspondence ideal of truth: the knower has to be doing something and be engaged with things for anything to be known scientifically.

---

19. This is a hypothetical position, not one I'm attributing to any philosopher in particular. However, a fair amount of the criticism of perspectivism insists precisely on the point that perspectivism either collapses into traditional realism (Chakravartty 2010) or instrumentalism (Morrison 2011). See the discussions by Chirimuuta (2016) and Massimi (2018).

20. It is doubly unfortunate that the textbook restriction to the scientific realism-empiricism dialectic neglects the variety of views that took up the Kantian lead—not only the neo-Kantianism of philosophers like Cassirer, with its emphasis on the spontaneity of the understanding, but also the pragmatist and phenomenologist philosophies of science, which have a strong commitment to construing knowledge formation as active. In pragmatism, the intellectual sphere is shaped by the demands of material praxis, while phenomenology emphasizes the embodied and embedded character of the scientific intellect.

One payoff of this reorientation is that it helps makes sense of the puzzle over how it is that models of physical phenomena can be so successful (i.e., yield satisfying explanations, be predictively powerful, and fruitful in the development of new experiments and models) while being so full of "distortions" such as idealizations and grossly simplifying abstractions. Philosophers like Batterman (2010), Bokulich (2012), and Potochnik (2017) have frequently argued that these distortions are essential to the models' explanatory success, and yet it has been hard to envisage any connection between deliberate distortion and somehow getting a better account of the target when thinking within the traditional constraints. Once knowledge-building activity is acknowledged and emphasized, we can think of models as devices that aim to achieve a certain compromise or balance between a natural system, the scientific collective mind, and some material purposes. Explanatory, predictive, and practical success are a matter of achieving the right kind of fit, not of the attainment of a God's-eye view on the subject. There can be various ways to succeed (a plurality of perspectives), and sometimes the best way to achieve alignment between the target system, human conceptual resources, and material goals is through deliberate distortion.

Some direct criticisms of scientific realism and empiricism will appear later in this book. The empiricism developed by Ernst Mach will be under scrutiny in chapters 8 and 9. His mistake was in thinking that science can be purged of metaphysical commitments, such as the need to refer to an ontology of things beyond the data points. We will see how automated science, which uses the massive data-processing capabilities of machine learning, is an instantiation of Mach's ideal, and as such reveals its limitations. Chapter 8 will also discuss the relationship between science and technology, challenging the common assumption that there is a nonporous boundary between pure and applied science. The primary argument in favor of scientific realism, originally put forward by Hilary Putnam, is that every other account leaves the predictive and technological successes of science an inexplicable miracle (Psillos 1999, chapter 4). But this assumes that technological success can only be a downstream consequence of the acquisition and application of true scientific theories. It in fact begs the question since it presumes a realist picture of science—aiming at correspondence with nature and insulated from practical concerns—that the argument is supposed to establish. According to haptic realism, the epistemic and the instrumental are two faces of the same thing: science. Technological success is attributed to the achievement of an adequate fit between the constraints stemming from the parts

and processes at work in a target system and those due to the limitations of human cognition. Technological success is achieved more often than not by simplification and deliberate distortions, and for this reason should not be taken as dispositive evidence that the theory employed has encapsulated some absolute truth about the workings of nature.

## 2.2 Proceeding with the Account

In this section, I am going through some common points of debate within philosophy of science, which will be important in the rest of this book, outlining the viewpoints that are provided by haptic realism. I will discuss in turn scientific analogy, mechanistic explanation, and scientific understanding and control. A way to sum up, in general terms, the account that unfolds over the course of the book is that science is to be conceived as a project of domestication in which wild things and processes are altered, reconstructed, so that they are knowable and usable for some people's purposes. This supposes an opposition of the human and the natural that may seem dissonant, given that the human brain itself is an object of investigation for neuroscience. The thing to keep in mind is that the distinction is no more than the marking of a simple difference between an object that has been worked over, subjected to modification through experimental procedures and schematized through modeling, and one left alone. Even in noninvasive neuroscientific experiments, such as functional magnetic resonance imaging (fMRI) of the human brain, neural activity is altered due to the person being in the unusual environment of the scanning tunnel; and the data gathered are further subjected to processing so that conclusions may be drawn from them. Thus, the distinction between the "natural" or "wild" and the "artificial" or "worked over" is comparative when employed in this picture of scientific activity as domestication. The brain of a genetically modified lab mouse is "artificial" compared to that of a field mouse, but that same brain is in a "wild" state when the lab mouse is left to its own devices in a cage, and again in an "artificial" state when the mouse is made to perform an experimental task.

### 2.2.1 Analogies So Considered

In chapters 4 and 6, I will be drawing on Marry Hesse's classic work on analogies and models in science. Scientific analogies domesticate what is

incomprehensible in nature by fixating on similarities with what is to the scientist more familiar and better understood (Hesse 1955, 353). Analogical inferences are conclusions drawn about the unfamiliar system on the basis of its similarity with one that is relatively well characterized. A basic, first-pass way of thinking about this pattern of inquiry is to take the similarities to be preexisting facts about the objects of investigation. Comparing two objects, some properties will be shared and others will not be. You might picture a Venn diagram where two classes of properties, each associated with one of the objects, partially overlap. The greater the extent of the overlap, the greater the number of analogical inferences that one may draw, and so the more revealing the comparison will be.

Hesse's early account of analogy did in fact posit an identity of structure between an analogy source, the model system, and the analogy target:

> The most obvious property of a satisfactory model is that it exhibits an analogy with the phenomena to be explained, that is, that there is some identity of structure between the model and the phenomena. Now one may say in a straightforward sense that there is an analogy between two branches of physics if the same mathematical structure appears in the theory of both, for example, the theories of heat and of electrostatics can be formulated in the same equations if one reads "temperature" for "potential," "source of heat" for "positive electric charge," and so on. When there is an analogy of this kind, one theory may be used as a model for the other. (Hesse 1962, 22)

Emphasizing that the similarity between model and target could never be total, Hesse distinguished between the "positive analogy"—the respects in which the two systems are similar or share the same structure—and the "negative analogy," the respects in which they are unlike one another (the disanalogies). She gives an example of analogical models closer to the ones under discussion in this book when she mentions "electronic tortoises," cybernetic robots that were receiving much publicity at the time. In such cases, she writes, "there is an obvious negative analogy in certain biological and chemical respects between the model and the animal, but a positive analogy of unknown amount in some aspects of behaviour" (Hesse 1962, 24).

This account of analogy presupposes a formal realism—indeed, a structural realism in the sense common in philosophy of science these days, whereby physical systems really do have the structural features that are represented in the well-established mathematical theories used to describe them and predict

their behavior.[21] My own account of neuroscientific modeling begins with a rejection of formal realism, so the account of analogy will diverge from this one. I will not be assuming that epistemically appropriate, useful scientific analogies depend upon a preexisting structure in common (a homomorphism) between the model and the target system. As discussed previously, determining structure in the target of investigation is an active, and to some extent constructive, business. The structure that is said to be shared with the analogical model is massaged out of the system—it is an ideal pattern; what is shared between the brain and a computational model is the ideal pattern, an idealized version of neural activity. This means that the predictive and explanatory success of the model does not justify the inference that the brain is literally performing the computations attributed to it by the model.

My position is consistent with some of the themes from Hesse's later work. She concurs with Hans-Georg Gadamer's rejection of formal realism, his "assumption that no ultimate order can be apparent to finite minds," which motivates the position "that there is a fundamental inexactness of all human knowledge," and therefore that the commonly held ideal of scientific language being univocal, literal, and maximally precise, is misguided (Hesse 1995, 363).[22] Instead, Hesse argues that all language, including scientific language, is rooted in the figurative rather than the literal (1994, 453) and can never completely abandon metaphor and analogy. This does not prevent it from conveying knowledge (Hesse 1995, 352) or being the medium for rational argument (Hesse 1994, 453). What it does stand against, however, is the literal interpretation of scientific models and descriptions and the tendency to take scientific accounts of the world as providing a purified account of the plain facts, as both positivists and their realist antagonists would have it.

In this way, Hesse's later work attacked the common idea that there is a sharp separation between literal and figurative language, the former being

21. The work of contemporary structural realists has frequently referred to the early structuralism of Ernst Cassirer. The critical difference is that Cassirer was not a formal realist! While Steven French (2014, 100) thinks that he can pick and choose the bits that he likes from Cassirer's structuralism, a key move he employs, which is to let metaphysics be brought into line by epistemology (59–60), is unprincipled unless some sort of idealism comes into play. Cassirer's own "critical idealism" amounts to his constant rejection of the correspondence theory of truth.

22. An interesting point of connection with section 1.5.3 in chapter 1 is that Gadamer (1975/2004, 435) aligns this position with the ideas of Nicholas of Cusa.

factual and the latter expressive. This picture of scientific language, as imbued with analogy and metaphor, allows a richer understanding of neuroscientists' use of everyday psychological terms, as well as terms borrowed from the technical domain of computer science, and has profound implications for how philosophers should interpret their theories.[23] One recent book addresses the question of whether neuroscientists' ascription of psychological terms like "thinks" or "perceives" to neurons instead of whole animals should be taken literally or treated as a mere metaphor (Figdor 2018). But according to Hesse's considered account, this is a false dichotomy: even at its most literal, scientific language remains metaphorical in some respects.

This argument ultimately leads back to metaphysics, and indeed theology. Leibniz serves as the prime example of how commitment to scientific language being ideally literal and univocal—as best achieved via artificial languages like mathematics—rests on a background worldview in which the universe is inherently intelligible to human reason because created in that way. As Hesse (1962, 159) describes, "Leibniz . . . is committed to the belief that rationality, that is, intelligibility in detail to the human mind, is part of the essence of things, and he therefore requires that descriptions of the ultimate structure of matter should be understood literally and not metaphorically."[24] A worldview like Leibniz's conditions a certain semantics: terms can have maximally precise meanings because the things they refer to are themselves precisely demarcated (this being one way to express formal realism). This then fosters certain views in philosophy of science. If your semantics tells you that the best scientific descriptions can be univocal, literal, and precise, you will interpret the preeminent theories and models in a literal way. The theological and metaphysical foundations are now mostly forgotten, but still a basically Leibnizian approach has dominated twentieth-century philosophy

23. Of the two major cases of analogy featuring in part II, computers and representations, the latter is much more fuzzily defined. There is no shared definition of neural representation among researchers (Vilarroya 2017). This does not by itself undermine the practice.

24. Cf. Gadamer (1975/2004, 416): "The ideal language that Leibniz is pursuing is a 'language' of reason: an 'analysis notionum' which, starting from 'first' concepts, would develop the whole system of true concepts and so be a copy of the universe of beings, just as is the divine reason. In this way, the world—conceived as the calculation of God, who works out the best among all the possibilities of being—would be recalculated by human reason."

of science, from Carnap's fetishization of artificial languages (Friedman 2000, Carus 2007) to the scientific realist's belief that physics is near enough to an account of how things just are. In this widely held view, only literal language meets the ideals of transparency such that only that mode of representation would be said to deliver a proper understanding of nature. But if we live not in Leibniz's world, but one in which intelligibility to human reason is not built into the fabric of things, then we should not expect the basic form of language and description of reality to be literal, nor should we retain nonmetaphorical language, purified of ambiguity, as the ideal case.

### 2.2.2 Mechanisms and Perspectives in Neuroscience

> "Life understands not death, nor death life." So said an ancient philosopher. Yet in his unceasing desire to diminish the boundaries of the incomprehensible, man has always been engaged in attempts to understand death by life and life by death.
>
> —Ernst Mach (1895, 186, translation modified)

The analogy with central status in the history of modern science is that of the world, and its inhabitants, as being machine-like. The clockwork universe gives a generalized picture of cosmology, while the presentation of organic bodies as self-moving machines, *automata*, has had a life of its own in the history of biology.[25] As Falkenburg (2019, 76) observes, mechanical analogies go together with a dissecting methodology, which is one reliant on structural and functional decomposition of the target systems—related in important ways to the reductionist simplifying strategy discussed in chapter 1. A different central analogy, common in premodern natural philosophy, was that of the world being like an organism. The use of the living body as an analogy source to conceptualize nonliving material systems is now obsolete—although the Gaia hypothesis might be classed as an exception here since it compares the whole planet, including both living things and nonliving weather and geological systems, to an integrated organism. According to Hesse, it was the rise in significance of another notion of simplicity in the seventeenth century, one shaped by a revival of Pythagorean and atomistic ideas that led to the fading of the world-organism analogy. On that notion, attribution of properties to inanimate matter was restricted

25. See Riskin (2016) for an extended treatment of automata in the life sciences.

to plain geometrical and mechanical qualities.[26] That said, there is a continuity between the world-organism picture and antimechanistic approaches within modern biology, variously known as *organicist, holist*, or *vitalist*. These approaches tend to stress the insufficiency of the dissecting method whereby organic systems are literally broken into pieces and get their parts examined independently of context within the body and ecosystem.[27]

In the *Critique of Pure Reason* and *Metaphysical Foundations of Natural Science,* nature is constituted in such a way that Newtonian mechanics is the right physics for it—that is, it is a causal nexus of spatiotemporally located objects interacting with one another. As such, the "mechanistic worldview" appears in Kant's philosophy,[28] but it is important to remember that for Kant, this is not an absolute reality. Furthermore, as noted by Cassirer (1950, 211), the preeminent position given to mechanistic explanation in Kant's philosophy is not consistent with developments in twentieth-century physics. Kant's original account has been given a pluralistic treatment by twentieth- and twenty-first-century philosophers of science. Along these lines, it is natural to say that mechanism is one scientific perspective among others, denying it exclusive explanatory rights and declining to ontologize its explanatory posits.

This has been the claim of my previous papers, where mechanistic approaches were treated as one distinct way of modeling neural processes, standing alongside computational and dynamical systems perspectives (Chirimuuta 2014, 2018).[29] It stands in contrast with the monistic ambitions of much of the work on "new mechanism" in philosophy of neuroscience,

26. "Such an understanding of simplicity in terms of the minimum number of mathematical variables consistent with the subject-matter, at once rules out organic analogues which generally contain more variables than the inorganic situation with which they are compared" (Hesse 1962, 99–100).

27. Examples of these antireductionist arguments will appear in the next chapter. See Peterson (2016) on organicism in prewar Britain, and Harrington (1996) on holism in Weimar Germany. Vitalism is controversial. The term is normally used disparagingly, but see Canguilhem (1965/2008a) and Wolfe and Donohue (2023).

28. Note that Kant uses the term "mechanistic" to refer to the Cartesian mechanics that he rejected, whereas I am using it more generally to include Newtonian ("dynamic") mechanics (Warren 2001).

29. In those earlier publications, I give more of a literal interpretation of the computational models than I now think is correct. See also Lee and Dewhurst (2021) on the "mechanistic stance" and Buzzoni (2019) on mechanism as a perspective.

which claims to show that other explanatory frameworks can be subsumed within the mechanistic one, to the extent that they offer genuine explanations (e.g., Kaplan 2011). New mechanists have downplayed the role of the machine analogy in contemporary science (Craver and Tabery 2017), where biological systems are routinely described as mechanisms without comparison to any specific artifacts. In my view, "biological mechanism" is not actually a dead metaphor, one that can be employed neutrally as not implying anything machine-like (cf. Falkenburg 2019, 85). This is because, as I discussed in chapter 1 and elsewhere (Chirimuuta 2020d), the general machine comparison underlying the notion of a biological mechanism imports an ontology that serves a simplifying purpose. In a machine, the construction is modular and the operation of parts (e.g., the causal interactions within a modular subsystem) can be understood independently of detailed knowledge of the workings of other modules.[30] Thus, a machine is, at least in principle, susceptible to a form of reductionist, bottom-up explanation in which the whole system is decomposed into subsystems and each subsystem is characterized in turn.[31]

---

30. One can thus see the connection to the characteristics of mechanisms posited in natural science. This is clear in a discussion by Dieks (2019, 56) with reference to Glennan's (2002) definition of mechanisms:

> "A mechanism for a behavior is a complex system that produces that behavior by the interaction of a number of parts, where the interactions between parts can be characterized by direct, invariant, change-relating generalizations."
>
> This definition accords well with the nature of the mechanisms that passed review in our historical sketch of classical mechanics. In particular, the behavior to be explained is *produced* by the *interactions* between the *parts*; and as Glennan explains, these parts must be *objects* with a high degree of robustness or stability, which are generally spatially localized. The interactions bring about changes in the properties of one part as a consequence of changes in the properties of another (Glennan 2002, 344). . . . . It is important for this new mechanicism, as it was for older forms of mechanicism, that the interaction between any two components should be the same as in the case in which these components are the only systems present: the interactions should not be 'holistic', depending on the behaviour of the complex system that is to be explained. The mechanistic intuition is that the global system should be reducible to its parts.

Against those who endorse a less restrictive characterization of mechanisms in biology, not requiring the system to be decomposable or have localization of function, I would respond, with Silbertstein (2021) and Taylor (2021), that this makes the notion of biological mechanism quite vacuous.

31. Note, however, that the understanding or reverse engineering of a machine of any complexity depends on a top-down approach as well, referring to the function of the entire system (Dennett 1995). Only the causal interactions within the simplest machines can be comprehended without any reference to overall function. Most of the mechanistic explanations discussed in the philosophy of neuroscience

But there is good reason to think that this approach, when imported into biology, underestimates the amount of integration and context-dependency at play in organic systems (Green, Serban, et al. 2018, 1752; Silberstein 2021). As such, the mechanistic description of a living process should be treated as only a first approximation.

Bill Bechtel is one philosopher of biology and neuroscience whose account has been in this vein. He treats mechanisms as models that idealize, and hence produce explanations by way of simplification. The key move is in positing that a biological system is bounded, in terms of the number and spatial and temporal scales of the processes that affect it, such that the researcher considers only a small number of the system's actual interactions. A set of entities and activities in a living system can only be taken to be a mechanism following the imposition of some hypothetical boundaries. Since the model of the bounded system misrepresents the number of interactions that are known to occur, Bechtel classifies it as an idealization: "While not arbitrary, mechanism posits are nonetheless idealizations in that they misrepresent the behavior of the mechanism as due solely to its components and their organization; they neglect the roles interactions with other entities play in determining the mechanism's behavior." (2015, 85)

Bechtel notes that it is not practically possible to extend the posited boundaries too much before the system becomes unworkable as a target of investigation. A mechanistic explanation, in Bechtel's epistemic sense, could not include all the factors relevant to the behavior of the system.

My endorsement of Bechtel's account, consistent with the Kantian position outlined in this chapter, is to be taken as read. Furthermore, my position aligns with interpretations of mechanistic explanation in neuroscience that depend on an epistemic theory of causality, rather than the view that causal relationships simply exist in objects of investigation (Falkenburg 2012; Taylor 2021). On this modernized Kantian account, people make causal claims not because they have latched onto some mind-independent causal connection, but because they have learned some facts about their environment that are

---

are therefore "multilevel," involving characterization of the whole system as well as at various levels of subsystem. For this reason, proponents of mechanistic explanation emphasize the difference between that and reductionist explanation (Craver 2007, 9–16). See Bickle (2008) for a more favorable account of reductionism in neurobiology.

most conveniently exploited for practical purposes if represented as causal dependencies (Williamson 2004, chapter 9).

This picture is in the background of my account in chapter 6 on the complementarity of explanations in neuroscience that posit either intentional or causal relationships between neural activations and the stimuli that are associated with them. Chapter 3 will examine the limitations of the mechanistic simplifying strategy that was employed in the now-defunct reflex theory of the nervous system. Since mechanistic explanation in contemporary neuroscience has already received a disproportionate amount of attention, I will not be focusing on it in part II. The distinct perspective offered by dynamical systems theory will be considered in chapter 7.

### 2.2.3 Explanation, Understanding, and Control

Much previous work in philosophy of neuroscience, including my own, has focused on questions of explanation: *What are the norms for satisfactory explanation? Does all explanation refer to causes? Can abstract and idealized models still explain things?* In this book, I will be changing the target slightly, with more of an emphasis on scientific understanding and on the instrumental goal of neuroscience, which is acquisition of the ability to control the workings of the brain. I take these to be the two core aims of the research, and in chapter 8 I will show how this conception follows from a more general account of the origins of modern science. It is important to appreciate here that understanding and control, the epistemic and instrumental aims, are two faces of the same activity. This follows straightforwardly from haptic realism since it makes clear that the processes of discovery and application mutually condition one another and do not have independent lives of their own. I will now say a little more about each of these terms.

I am following Potochnik (2017) and Elgin (2017) on the identification of understanding, rather than truth, as the primary epistemic aim of science. Potochnik's argument begins with the observation that idealizations are rampant and unchecked within scientific representations. That is, scientists present the world in ways that are known to conflict with their empirical findings, and yet they do not in most cases attempt to deidealize their models to make them more complicated but more consistent with the observed phenomena. To Potochnik, this suggests that truth—in the sense of faithfulness to the observed facts—cannot be the primary epistemic aim of science. Instead, idealized representations are retained because they aid

understanding by virtue of their greater simplicity. This is the indication that the aim of understanding trumps that of truth.

The reader may be wondering what the difference is between the terms "explanation" and "understanding." The preceding paragraph could be written with "explanation" replacing the word "understanding" without much change in meaning. The thing to appreciate is that in the philosophy of science, explanation is often treated as something totally objective, not standing in any relation to human psychology. This was a feature of Hempel's (1965) classic but now unpopular covering law account of explanation, in which a scientific theory was said to explain a phenomenon if its laws could be used to predict the occurrence of the phenomenon. Carl Craver (2014) has defended an ontic sense of explanation in which states of affairs in the world are said to be the explanation of a phenomenon, independently of anybody's reference to them. Even though other philosophers, like Bechtel, restrict explanation to an epistemic sense (see Illari 2013), the prevalence of the Hempelian and ontic notions makes "explanation" an ambiguous term. In contrast, understanding is always taken to bear an ineliminable relationship to a human investigator. When understanding is recognized as the epistemic aim of science, it is clear that this aim does not transcend human standpoints and agendas. Science, in its pursuit of knowledge, is not a collective march toward truth in an absolute sense (truth pertaining to the thing in itself), or even in the empirical sense of consistency with the maximum number of observations. Instead, it is a striving toward a knowledge of things that makes sense to and for the people who are producing and using it, people who exist in groups and societies with collective aims that include the instrumental ones of manipulation and control.

Chapter 8 will explore the point that scientific understanding is directed toward control rather than the more contemplative purpose of understanding nature for its own sake. This again makes sense of the prevalence of "unrealistic" simplifications—ones recognized to be oversimplifications—within the store of neuroscientific models and theories. A toy model that makes clear a manageably small number of dependency relationships will suffice for many practical purposes where the excluded details are assumed to be irrelevant to the task. This emphasis on the interrelationship between the aims of understanding and control will feed into the account of understanding presented in chapter 8. It is based on the *verum factum* of Giambattista Vico, which is the idea that what is comprehensible to human beings is what is made

by them. The basic idea is that neuroscientific understanding is achieved through the building of proxy systems—not only computational models, but also concrete neural systems in which complexity has been reduced—that are simple enough to be understood, and thereby manipulated.

To conclude, the epistemic concerns of neuroscience are not to arrive at some disinterested kind of knowledge of the inner being of its objects. That is, the target is not knowledge for its own sake, of the brain as it is in itself, but of ways that neural processes may be put to the external ends of the investigator. These may well be humanitarian aims, like the cure of psychiatric disease and prevention of neurodegeneration. At the same time, it is obvious that manipulation and control are not unqualified goods. The instrumental achievements of neuroscience are as yet fairly modest, but it is clear that in other domains of inquiry, the power to alter nature has gone beyond what is beneficial for humans and other life forms. The point of my inquiry is to show how the drive toward instrumentality and the appetite for knowledge work together in neuroscience, as well as how they mutually reinforce each other. This must be recognized if there is to be a proper interpretation of what neuroscience is telling us about the brain, and hence the mind. The thing to be remembered is that through the interactions and iterations of experimental research, science constitutes its objects with a view to manipulation. The object will be presented, by science, as a manipulandum. That goes as much for whole humans as for a few of their nerve cells, or an ear of wheat.

# Part II

What follows is a smattering of episodes from within the last 100 years of work in the neurosciences, chosen to illustrate the way that the demands of simplification have shaped both theory and experiment. I draw primarily from motor and visual neuroscience in mammals. Similar stories could be told about many other domains of research. Visual neuroscience has long been the test bed for theories of cortical function, which makes it a rich source of examples for my study.

# 3 The Reflex Theory: Misleading Simplicity in Early Neuroscience

> Let us observe how the mechanical genius of our time has diffused itself into quite other provinces. Not the external and physical alone is now managed by machinery, but the internal and spiritual also.
>
> —Thomas Carlyle (1839, 318)

Most recent analyses of simplifying strategies, such as abstraction and idealization, examine cases of prima facie successful science and offer accounts of how it is that the science works as well as it does, given that its theories and models involve such drastic simplifications. In this chapter, my subject is a now-obsolete branch of neuro- and behavioral science, which is not, in hindsight, considered very successful. This serves to make the point that simplicity is not an unqualified epistemic good—it can sometimes be misleading. It will also illustrate the claim made in chapter 2, that scientific perspectives often take divergent forms because of differences in the goals that drive them. The conclusion of that chapter was that neuroscience as we know it today conceptualizes its objects of investigation in such ways as to facilitate their manipulation. Here, we will focus our study on a mechanistic perspective within historical neuroscience that shows the stark connection between simplification and instrumentality, as well as how these together shape the scientist's vision of what is being discovered.

My topic is *reflexology*, the now-outmoded term for the branch of neurophysiology and psychology centered around the simplifying assumption that *complex* patterns of behavior are concatenations of *simple* reflex responses, exemplified by the sensorimotor reflex arc discovered in the 1830s.[1] In the

1. See Canguilhem (1994) and Clarke and Jacyna (1987) on the early discoveries in reflex physiology.

hybrid discipline once known as "physiological psychology" (Smith 1973), the concept of "reflex action" referred both to a pattern of nerve activation and to an involuntary kind of behavioral response. In his history of the reflex theory, the psychologist Franklin Fearing writes: "For those sciences which are primarily devoted to the study of the integrated responses of living organisms, the concept of reflex action has played, in the 19th and first twenty-five years of the 20th century, a dominating role, comparable, perhaps, to the influence of the Newtonian hypotheses in physics" (1930, 4, cf. Skinner 1931/1961, 319).

Unlike Newtonian mechanics, which is now deemed usable and valuable, even if strictly speaking false, reflex mechanics is no longer employed by behavioral scientists and neuroscientists. It is not featured in textbooks and has not been part of the training in theoretical neuroscience for decades. This retrospective lack of success makes reflexology worth considering as a case of simplification-gone-wrong. We can ask whether the in-aptness of the simplifying assumptions contributed to the failure of this program. In addition, while the reflex theory was the mainstream framework in the period that we are considering, the first three decades of the twentieth century, it did have its critics. It is valuable to consider the skeptical voices—the reasons given for rejecting the simplifications presumed by proponents of reflex theory.

Rather than involving deliberate omission of detail (abstraction), or introduction of falsehood (idealization), within a mathematical model, reflex theory employed the simplifying strategy of decomposition of, alternately, the nervous system or the behavior of an animal, into putative elementary parts—the reflex arc or the reflex response. These reflexes were metaphorical elements or atoms—parts whose properties were posited to stay the same regardless of varying conditions around them or observable changes in the animal's pattern of movements. This simplifying strategy offers a kind of reductive explanation, as argued by Bechtel and Richardson (2010, xxxvii). One of the proponents of the reflex theory, Jacques Loeb, stated the fundamental intuition quite succinctly: "It is better for the progress of science to derive the more complex phenomena from simpler components than to do the contrary" (1912, 58); and "The progress of natural science depends upon the discovery of rationalistic elements or simple natural laws" (1912, 59).

An important question, raised by the work of Loeb and others, is whether the belief in the fundamental simplicity of nature that these statements seem to express has anything more than a pragmatic status. By the end of

this chapter, we will see that the tenets of the reflex theory were justifiable only on instrumentalist grounds. But it turns out that the framework was not as serviceable for manipulation of behavior as had been hoped. Needless to say, the engineering of a scientific framework to facilitate prediction and control does not guarantee that technological targets will be met.

### 3.1 Atoms of the Nervous System and Elements of Behavior

In my exposition of the reflex theory, I will focus initially on two influential scientists, active in the early decades of the twentieth century—Jacques Loeb and Ivan Pavlov.[2] Loeb's research in this area was primarily in the physiology of invertebrates, whereas Pavlov is still renowned for his experiments on learning in dogs. Both envisaged that the concept of the reflex could be the basis of an integrated and complete explanation of the brain, nervous system, and behavior. Another shared characteristic of their outlook was that they sought to model biology and psychology on the physical sciences which, in particular, meant applying the analytical methods of mechanics to the animal. That is, they aimed for a decomposition of the nervous system and movements of animals into the components (i.e., reflexes), whose occurrence could, when taken together, account for the activity of the whole nervous system or whole animal.[3] The science of Loeb and Pavlov was self-consciously, and literally, mechanistic.

The very first words of the introduction to Loeb's *Comparative Physiology of the Brain* are as follows: "The understanding of complicated phenomena depends upon an analysis by which they are resolved into their simple

2. See Pauly (1987) for an account of Loeb's career and methodology and Todes (2014) on Pavlov; also Smith (1992) on Pavlov and his predecessors. For ease of exposition, I have smoothed over much of the variety of opinion within reflexology. To get a better sense of the range of views within this program, see Fearing's (1930) discussion of many now-forgotten figures. Note also that in my account, Pavlov's classification of *unconditioned* and *conditioned reflexes* is treated as equivalent to the distinction between *simple* and *complex reflexes* by Sherrington and others. Both terminologies capture the notion of elementary versus nonfundamental reflexes but this papers over the differences between these sets of concepts and the roles they play in their respective theories.

3. See Falkenburg (2019, 76–79) on these methods of decomposition and explanatory reconstitution (aka "analysis and synthesis") and their relation to mechanistic science.

elementary components. If we ask what the elementary components are in the physiology of the central nervous system, our attention is directed to a class of processes which are called reflexes" (Loeb 1900, 1).

Loeb goes on to give two examples of reflexes—the eyelid closing on the advance of a foreign body and the narrowing of the pupil in response to light. He then defines the reflex by describing what is common to these examples: "In each of these cases, changes in the sensory nerve-endings are produced which bring about a change of condition in the nerves. This change travels to the central nervous system, passes from there to the motor nerves, and terminates in the muscle-fibres, producing there a contraction. This passage from the stimulated part to the central nervous system, and back again to the peripheral muscles, is called a reflex" (Loeb 1900, 1–2).

And he adds that "there has been a growing tendency in physiology to make reflexes the basis of the analysis of the functions of the central nervous system" (Loeb 1900, 2).

In the first lecture of Pavlov's *Conditioned Reflexes*, we find a comparable definition: "An external or internal stimulus falls on some one or other nervous receptor and gives rise to a nervous impulse; this nervous impulse is transmitted along nerve fibres to the central nervous system, and here, on account of existing nervous connections, it gives rise to a fresh impulse which passes along outgoing nerve fibres to the active organ, where it excites a special activity of the cellular structures" (1927/1960, 7).

In addition, Pavlov emphasizes the "necessity" of the connection between stimulus and response, a characteristic that ensures that the reflex is a "genuine scientific conception" (1927/1960, 7). Like Loeb, Pavlov describes reflexes as having an elementary status—he refers to them as "the elemental units in the mechanism of perpetual equilibration"—the mechanism through which an animal adapts its behavior to its surroundings (1927/1960, 8). A distinctive feature of Pavlov's theory is the division between *unconditioned* and *conditioned reflexes*.[4] Ones of the former type are those that exist from the birth of an animal and persist after the removal of the cerebral cortex,[5] and the

4. Todes (2014, 1) points out that this is a mistranslation of Russian terms, better conveyed as "unconditional" and "conditional." However, I keep to the standard translation as the difference in meaning is not pertinent to my discussion.

5. Pavlov describes the unconditioned sort as the "numerous machine-like, inevitable reactions of the organism—reflexes existing from the very birth of the animal, and due therefore to the inherent organization of the nervous system." (1927/1960, 8)

second kind are the central topic of Pavlov's research. Conditioned reflexes are said to be learned through the creation of an association between an arbitrary stimulus (e.g., the ticking of a metronome at a particular frequency) and a stimulus that innately causes a reflexive response—such as food, the stimulus for the unconditioned reflex of salivation. The cerebral cortex is said to be the neural substrate of conditioned reflexes since these responses do not persist after surgical removal of this structure. Pavlov posited no additional, "higher" mental structures that exert control over reflexive behavior—all actions are said to be determined by these two kinds of reflexes.

This radical conclusion did find adherents, such as Hull and Baernstein (1929, 14), who write, "It is believed by increasing numbers of students of human and other mammalian behaviour that the conditioned reflex, with its power of substituting one stimulus for another, is the basic mechanism not only of ordinary habits but of the entire mental life."

The proposal of these authors is that the conditioned reflex should be replicable in human-made device. Their idea was to build a reflex machine and assess its capacities for learning and intelligence as a test of this generalized theory of the mind. Still, the thesis that the reflex is the basis of "mental life" in its entirety was never endorsed by the British physiologist Charles Sherrington, who himself was a major contributor to the field. Sherrington's own way of characterizing the elementary status of the reflex was very influential: "The reflex-arc is the unit mechanism of the nervous system when that system is regarded in its integrative function. The unit reaction in nervous integration is the reflex, because every reflex is an integrative reaction and no nervous action short of a reflex is a complete act of integration." (1906a, 7).[6]

His view was that these fundamental units of activity, the reflexes, were compounded together during the evolution and development of the nervous system to generate the basis for more complex behavior. The motions of running, walking, and leaping are the outcome of reflexes concatenated together.[7]

I have highlighted how proponents of the reflex theory took reflexes to have an elementary status. Relatedly, Kurt Goldstein, a major critic of reflexology (see section 3.2), refers to its methodology as "atomistic." To understand the claim that the reflex approach is atomistic, we must take the

6. See Casper (2014) for a useful overview of Sherrington's career and the notion of "integration."

7. See Graham Brown (1914, 19–20) for discussion.

original meaning of "atom," where "atoms" are the ultimate simples—being indivisible and unchanging, they are the basic constituents of more complex wholes. Just as there can be metaphorical atoms, outside of physics, we have metaphorical elements outside of chemistry by extending the idea that there are fundamental, simple kinds of substances out of which more complex compounds are made. To treat simple reflexes as the atoms or elements of nervous system and behavior is to assert that these processes, or patterns of response, are stable (at least after they have been established, in the case of conditioned reflexes) and that they underlie the apparent complexity of behavior that is manifestly varying. The atomistic methodology is therefore a simplifying strategy—it is an attempt to explain complex appearances in terms of simpler fundamental components.

Fearing writes of reflexologists such as Pavlov that "it is characteristic of this point of view that the 'simple' reflex is described as it appears in the lower animals or in the spinal animal,[8] and there is a tacit assumption that these characters are the same for the more complicated types of nervous action, e.g., those involving the cerebrum" (1930, 296).

The idea is, first, that the scientist can discover these atoms of the nervous system, or elements of behavior, by examining experimental subjects such as invertebrates ("lower animals") or brainless vertebrates, incapable of any complexity of action; it is presumed that those actions reveal the fundamental components of behavior in all their raw simplicity. For instance, it is assumed that invertebrates, lacking a cerebral cortex, will only demonstrate responses attributable to unconditioned reflexes. Likewise, vertebrates like dogs and cats that have been experimentally prepared by decapitation or removal of the cortex will manifest only those simple reflexes. Second, the assumption is that in animals who can and do demonstrate more complex behavior (e.g., dogs and cats with their brains unharmed), the simple reflexes are still there, with the same physiological characteristics as in the "prepared" dog or cat, but somewhat masked by the overlay of complex, conditioned reflexes, as well as top-down inhibition of simple reflexes from the brain (Fearing 1930, 296). While Pavlov emphasizes the *necessity* of stimulus-response cause-and-effect connections, which for him make the reflex a properly scientific concept, he does not propose that these connections will always be *observable*. Even

8. That is, one whose entire brain has been removed (Sherrington 1909).

conditioned reflexes, he reports, are subject to countless influences, "disturbing factors," which interfere with their manifestation (Pavlov 1927/1960, 20). For this reason, he conducted his experiments on the conditioned reflex in dogs in a special isolation chamber in which the animal had no contact with its fellows, or even with the experimenter.[9] Pavlov's reliance on such scenarios is significant. As we now turn to criticisms of the reflex theory, we will see that much of the skepticism has focused on the assumption of the existence—*outside controlled laboratory conditions*—of any of these simple and elementary reflex responses.

## 3.2 The Criticisms of Reflexology

The supremacy of the reflex theory did not go uncontested. In this section, I summarize a number of the objections that were leveled at the ontological and methodological assumptions of reflexology, including the interventions of two philosophers, John Dewey and Maurice Merleau-Ponty. My summary is not exhaustive. For example, I do not include the criticisms offered by the physiologist Thomas Graham Brown, precisely because he does not take issue with the simplifying assumption of decomposition into fundamental units, but instead proposes an alternative kind of element, the "half centre" (e.g., Graham Brown 1914).

Six strands of criticism can be determined. I will discuss each in turn:

1. Empirical findings of the lack of stability of simple reflexes and conditioned reflexes
2. The ad-hocness of the postulates added to the reflex theory to achieve consistency with those empirical findings

9. It is worth reading Pavlov's own justification for the selection of such unnatural conditions for his experiments on conditioning:

> It was evident that the experimental conditions had to be simplified, and that this simplification must consist in eliminating as far as possible any stimuli outside our control which might fall upon the animal, admitting only such stimuli as could be entirely controlled by the experimenter. . . . . The environment of the animal, even when shut up by itself in a room, is perpetually changing. Footfalls of a passer-by, chance conversations in neighbouring rooms, slamming of a door or vibration from a passing van, street-cries, even shadows cast through the windows into the room, any of these casual uncontrolled stimuli falling upon the receptors of the dog set up a disturbance to the cerebral hemispheres and vitiate the experiments. To get over all these disturbing factors a special laboratory was built at the Institute of Experimental Medicine in Petrograd . . . (Pavlov 1927/1960, 20).

3. Questioning of the reductionist methodology, which seeks explanation of behavioral wholes in terms of simple parts
4. Dubiousness of the extrapolation from the neurophysiology of the periphery and spine, assumed by the reflex theory, to anatomically unknown structures within the brain
5. The lack of ecological validity of physiological experiments performed on surgically altered animals, as well as of behavioral experiments performed in highly artificial laboratory conditions
6. Dubiousness of the notion of the simple reflex, even when construed as an abstraction

The purpose of this section is to give a presentation of these criticisms without evaluation or endorsement. Assessment of the cogency of the criticism will be deferred to section 3.3.

One sustained case against the reflex theory was put forward by the German neurologist Kurt Goldstein in his book *The Organism* (Goldstein 1934/1939; see also Goldstein 1940, chapter 5). His first point of attack is that experimental reports of the putative simple reflexes do not show the stability, the constancy of response, that is postulated by the theory (Goldstein 1934/1939, 69ff.). Summarizing the results of various researchers, including some proponents of the reflex theory, it appears that the simple responses, like the patellar reflex, are altered by bodily posture and attentional state. As reported by Sherrington (1906b, 49), the *receptive field* of the scratch reflex in dogs—the area on the skin in which a stimulus can elicit the scratching movement of the leg—varies in threshold sensitivity from day to day. Regarding the conditioned reflex of Pavlov, Merleau-Ponty (1942/1967, 58), following Buytendijk and Plessner (1936), argues that it is too unstable to do the theoretical work required of it. A striking case of instability comes from Pavlov's reports of the behavior of two dogs that had been subjected to repeated conditioning experiments. They appear to fall into a hypnotic stupor and fail to give the expected reactions to either the conditioned or unconditioned stimuli.

This brings us to the next allegation, that the reflex theory is full of ad hoc modifications–the unprincipled use of terms such as "excitation, inhibition and disinhibition"—brought in to mask the disagreement between theory and observation (Merleau-Ponty 1942/1967, 58ff.; 19–20; cf. Buytendijk and Plessner 1936). In particular, when the usual stimulus fails to elicit the

expected reflex, it is posited that a process of inhibition has been activated, preventing the response; but independent evidence for the inhibitory mechanism is not established. Goldstein contends that with the proliferation of hypotheses accounting for the modifications of normal, simple reflexes, the theory loses its justification for drawing a distinction between the normal reflex and variants of it.[10] This lack of justification goes unnoticed because researchers automatically classify the responses produced in certain kinds of artificial experiments as the normal ones (Goldstein 1934/1939, 80–81).

Given this, Goldstein is skeptical of the classification made of the "normal" reflex versus its variants, and of the simple reflex versus the complex patterns of behavior that they are said to comprise. Put together with the abovementioned observation of lack of stability of the supposed elementary responses, Goldstein calls into question the reductionist methodology that attempts to explain a complex behavioral whole in terms of simpler parts. The following passage is worth quoting at length:

> The customary method attempts to reduce variable to constant reactions, seeing, in the latter, the basic ones, and regarding the former as modifications. This tendency is understandable as a very natural desire to deal with constant factors. The supposedly greater simplicity of constant reactions lends itself as a starting point for a theory, in that the variable responses can then be understood as complexes derived from the more simple and constant ones. However, there is no question but that the so-called variable processes are, in reality, no less constant, if one takes into consideration all their causal conditions. Concerning the question of simplicity and complexity, and whether the complex can be deduced from the simple, we shall see, in our later discussion, that the converse view is probably nearer the truth. (Goldstein 1934/1939, 80)

What this suggests is that Loeb's assertion, that it is good scientific method to derive more complex phenomena from simpler ones, has met an obstacle: Goldstein argues that in the science of the nervous system and behavior, no foundation can be found in a substrate of components or processes, both simple and stable.

Along with the attempt to derive the complex from the simple, the reflex theory proposes to infer facts about brain processes from observations of

10. Goldstein lists these hypothetical factors: "inhibition, facilitation, neural switching or shunting of different kinds, influence through peripheral factors, such as the state of tension of the muscles, position, enforcement or diminution through other reflexes, 'central' factors, and amongst these, particularly, psychic factors" (1934/1939, 80).

the more accessible processes in the spine and peripheral nervous system. This extrapolation is challenged by Theodore Hough in an address to the American Association of the Advancement of Science:

> we miss entirely the satisfaction of seeing the cerebral functions clearly pictured in terms of neurone structure. We trace the "way in" and the "way out"; we see that the connection between the afferent and efferent nerve fibers is in the cortex; but what takes place in the cortex? Is it objectively nothing more than our typical reflex raised to the *n*th power of complexity? Perhaps it is; but does any one feel reasonably sure of it? For one, I confess I do not. (1915, 408, quoted in Fearing 1930, 287)

It is significant here that the modeling of cortical neurophysiology, as merely a more complicated kind of reflex, is reported as if it has been taken as a matter of faith. At this time, the neuroanatomy and physiology of the brain were uncharted in comparison with that of the nervous system below the neck. Given their ignorance, reflexologists made the parsimonious assumption that there was nothing radically different going on in the brain. But research later in the twentieth century showed that in this case, parsimony was misleading.

A core challenge to reflexology turns on its deficiency in what we would now call *ecological validity*—the lack of applicability of experimentally generated phenomena to the explanation of the intact nervous system or unconstrained animal behavior.[11] Critics of reflexology go so far as taking the central phenomena of the research program to be experimental artifacts. Having rejected the notion that reflexes are the elementary components of nervous system and behavior, Goldstein concludes that they are no more than an "expression of experimentally produced injury" (1934/1939, 157), especially because of the manifest difference between reflex reactions of the legs and the normal flow of movements of an animal walking over its usual ground (1934/1939, 169–170). The failure to notice the lack of similarity between ordinary movements and reflexive ones stems, Goldstein argues, from the fact that in their research, many physiologists never deal with intact animals (1934/1939, 90).[12]

---

11. Note that a version of ecological validity was a core "methodological postulate" for Goldstein: "no phenomenon should be considered without reference to the organism concerned, and to the situation in which it appears" (1934/1939, 25).

12. Canguilhem (1965/2008b, 113) gives an interesting commentary on Goldstein's view on experimental conditions: "The situation of a living being commanded from the outside by the milieu is what Goldstein considers the archetype of a catastrophic

Although Pavlov performed his conditioning experiments on animals whose nervous systems were unharmed, he still faced the criticism that the artificiality of his experimental setup—such as the confinement and isolation of the dogs—placed limitations on what could be inferred from his results about learning and behavior in general. Goldstein (1934/1939, 175) argues that the precise, repeated paring of unconditioned and conditioned stimuli does not occur in the lives of animals away from human control. Thus, they do not help to explain animals' learning in the wild, but they do shed light on the processes in play during human training of animals (Goldstein 1934/1939, 178). A comparable point about difficulties arising with use of artificial stimuli had been made by Herbert Spencer Jennings, a zoologist and former student of Dewey, in response to Loeb's attempt to make *galvanotropisms* (reflex responses to electric currents) fundamental to the explanation of movement (see Loeb 1900, chapter XI). Not only did the movements elicited in those experiments appear highly unnatural, but the electrical stimulus was one that simply did not occur in the environment of the organism—how could it then form an explanatory basis for the account of ordinary locomotion? (Jennings 1906, chapter XIV; and see Pauly 1987, chapter 6). Dwelling on Pavlov's report of the two dogs for whom repeated conditioning experiments led to their entering a hypnotic stupor, Buytendijk and Plessner (1936) conclude that his research on conditioning can only be informative about the genesis of neurosis!

We now move to the final point, on the lack of utility of the reflex concept, even as an abstraction. Dewey (1896) was one early critic of reflex psychology. He argues that the assumption foundational to the theory, of a clear distinction between stimulus and response, sensory and motor operations, is an artificial, misleading abstraction that masks the concrete fact of the interdependence of sensation and movement. In *The Integrative Action of the Nervous System*, Sherrington does indeed admit that the notion of

---

situation. And that is the situation of the living in a laboratory. The relations between the living and the milieu as they are studied experimentally, objectively, are, among all possible relations, those that make the least sense biologically; they are pathological relations."

Elsewhere in this essay, the changing fortunes of reflexology are discussed (Canguilheim 1965/2008b, 107–111).

the simple reflex is an abstraction, but makes the claim that it is at least a "convenient fiction":

> A simple reflex is probably a purely abstract conception, because all parts of the nervous system are connected together and no part of it is probably ever capable of reaction without affecting and being affected by various other parts, and it is a system certainly never absolutely at rest. But the simple reflex is a convenient, if not a probable, fiction. Reflexes are of various degrees of complexity, and it is helpful in analyzing complex reflexes to separate from them reflex components which we may consider apart and therefore treat as though they were simple reflexes. (Sherrington 1906a, 8; cf. 114)

This is one way to deal with the objection that stable and constant reflexes are never actually observed (point 1). Given their ubiquity in other branches of science, one may rightly ask what is wrong with "abstractions" or "fictions," so long as they are recognized as such.

Goldstein rejects even the "fictional" concept of the simple reflex because (for reasons just discussed) he does not think that it delivers the requisite understanding of the intact organism.[13] Merleau-Ponty delivers an involved response to Sherrington's deployment of the notion of abstraction. He counters Sherrington's claim that the reflex is an abstraction by asserting that it is actually a concrete occurrence, albeit one that is contrived experimentally and lacking more widespread significance: "But neither is the reflex an abstraction, and in this respect Sherrington is mistaken: the reflex exists; it represents a very special case of behavior, observable under certain determined conditions. But it is not the principal object of physiology; *it is not by means of it that the remainder can be understood*" (Merleau-Ponty 1942/1967, 46).

Merleau-Ponty criticizes Sherrington for his deployment of the abstract idea of the reflex to preserve an ontology of animal machines in the face of countervailing evidence (Moinat 2012, 95–97). Yet these appeals to the reflex are not adequate to account for Sherrington's key discoveries of "integration"—the coordination of movement required for adaptive behavior. As Merleau Ponty puts it, "It is paradoxical to conserve the notion of the reflex arc theoretically without being able to apply it anywhere in fact. As in all the particular questions which we have mentioned, in his general conception of nerve functioning Sherrington seeks to save the principles of

13. Elsewhere, I give a more detailed discussion of Goldstein's views on abstraction (Chirimuuta 2020a, 2020d).

classical physiology. His categories are not made for the phenomena which he himself has brought to light" (1942/1967, 33).

And here, we are brought back, in a roundabout way, to the beginning of our list of criticisms: the failure of the reflex theory to properly meet empirical facts.

It is not obvious what actual impact these criticisms had, and this is not a matter that I wish to settle in this chapter. For instance, we should not imagine that Sherrington's defense of abstraction is a response to Dewey's charges. Eventually, the reflex theory was eclipsed when computation came to provide an alternative simplifying framework for neuroscience and cognitive science in the mid-twentieth century. Chapter 4 discusses the rise of computationalism, arguing that its appeal rested not least in the simplifications that it offered to neurophysiologists. In terms of its core tenet—that nervous system and behavior can be explained via decomposition into elementary reflexes—reflexology is a theory without retrospective success,[14] although at the end of this chapter, I will discuss the afterlife of the reflex theory. Now it is necessary to do some evaluation of these criticisms. We will find that some of them miss their mark because of a more fundamental disagreement, between critics and proponents of the reflex theory, concerning the aims of research. This provides a telling illustration of the way that experimental methodologies, simplifications, epistemological standards, and instrumental goals, are woven together to constitute scientific perspectives.

## 3.3 Arbitration

### 3.3.1 Realist and Instrumentalist Stances toward the Reflex Theory

One way to summarize Goldstein's complaint against reflexology is that it is reductionism gone rogue, misled by the attraction of parsimonious explanations. The reflexologist employs a *reductionist methodology*—opting to study parts (simple reflexes) in isolation, with the aim of seeing how their operation together will yield an explanation of the whole nervous system.[15]

14. See for instance, Todes (2014, 300–302) on the failed ambitions of Pavlov's project.

15. This formulation of reductionism is very much in line with the one presented by Bechtel and Richardson (2010): reduction as decomposition of the living system into component mechanisms.

Furthermore, the reflexologist tends to assume a *reductionist ontology*, supposing that reflexes are elementary components, which when aggregated together comprise the whole nervous system; he takes it that the organism is "a bundle of isolable mechanisms which are constant in structure, and which respond, in a constant way, to events in the environment (stimuli)" (Goldstein 1934/1939, 67). That is, the possibility of context dependency for these responses—of parts behaving differently when situated in their wholes—is not considered by the reflexologist. Much of the content of Goldstein's magnum opus, *The Organism*, is a statement of the importance of context dependency in biology.

According to Todes (2014), the picture of the organism as a mere aggregate of physicochemical mechanisms, the reflex being the relevant one for the nervous system, was indeed foundational for Pavlov. He was a reductionist in both the methodological and ontological senses. The failure, described in section 3.1, of detailed experimental work to provide support for the existence of stable, elementary reflexes therefore stands as a challenge to Pavlov's theory. However, not all practitioners of reflexology took up this realist ontological stance. In the US, a distinctly instrumentalist version of reflexology was propounded by behaviorist psychologists. B. F. Skinner is an important case in point. Under the influence of ideas from Mach and Bridgman,[16] he asserted an operationalist philosophy of science in which it was unnecessary and misguided to entertain the question of whether a simple reflex *really* exists, and whether any experiment has been adequate to reveal it:

> Is a reflex a unitary mechanism? Is behavior a sum of such mechanisms? Then, if by reflex we mean a hypothetical entity which exists apart from our observations but which our observations are assumed to approach, the questions are academic and need not detain us; if, on the other hand, we define a reflex as a given observed correlation or as a statistical treatment of observed correlations, the questions are meaningless, for they ignore the process of analysis implied in the definition. A reflex, that is to say, has no scientific meaning apart from its definition in terms of such experimental operations as we have examined, and, so defined, it cannot be the subject of questions of this sort. (Skinner 1931/1961, 341)[17]

16. See Moore (2005) on Skinner's philosophical influences. The connections between Mach, Loeb, and Skinner's teacher, Crozier, are especially interesting.

17. Also, "the notion of a reflex is to be emptied of any connotation of the active 'push' of the stimulus" (Skinner 1938, 21).

Thus, it becomes clear that the first line of criticism, that simple, stable reflexes do not exist and a fortori, the nervous system and behavior are not compounded from them, can only be leveled at the ontologically committed version of the reflex theory, but not the operationalist version put forward by Skinner. This is exactly what Skinner (1940, 463) points out in a review of Goldstein's *The Organism*: "The possibility is not entertained [by Goldstein] that, short of believing in reflexes as 'things,' one may still hold to the predictive value of correlating responses with stimuli and with other variables. The probably meaningless and certainly unimportant question of the *existence* of reflexes bears the brunt of the attack."

With this stance there comes a shift away from the justification of the reflex theory in terms of *simplicity in nature* and toward an emphasis on *simplification*—the production of simple phenomena and cause-and-effect relationships that are not claimed to be part of the fabric of nature (because such claims would veer from positive, factual science into metaphysical speculation) but nonetheless serve some practical purposes. This position is consistent with the ideas of Mach and Bridgman. According to Mach, the "task of science is to provide the fully developed human individual with as perfect a means of orienting himself as possible" (1886/1914, 37; quoted in Smith 1995, 42). That is, the aim of science is not to supply disinterested knowledge of nature, but to furnish the agent with tools for effective practice. As such, Mach's doctrine of science as "economy of thought"—where science is to provide "the concisest and simplest possible knowledge of a given province of natural phenomena" (1883/1919, 6–7)—cannot be taken as the claim that science must reveal order and simplicity in nature, but that simplicity is strived for because of its utility.[18] Similarly, Bridgman (1927, 51–52; cf. 204ff.) remarks that any scientist's conviction in the fundamental simplicity of nature—owing, for example, to a belief of there being only a small number of elements—has no more than a pragmatic status. Bridgman also observes that methodological reductionism has practical appeal because of its ease of application, but then it "will appear to be of disproportionate importance" (1927, 221–222).

To tie together these threads, we see that the charges against reflexology regarding the reliance on methodological reductionism, lack of ecological

18. See Smith (1995, 45), which argues that Skinner follows Mach in this science-as-economy view.

validity, and even the nonexistence of the simple reflex are not devastating against an operationalist and instrumentalist construal of the reflex theory. If simplifications are justified for the part played in a quasi-engineering project, the production of specific responses and behaviors, then these criticisms are not pertinent. The reflexologist-as-technologist no longer sees himself as a disinterested researcher seeking the truths of nature, but more as an investigator whose goal is to take command of natural processes.

Not coincidentally, the behaviorists in the US were enthusiastic about the second of these two job descriptions. At the head of John Watson's behaviorist manifesto, it is declared that psychology's "theoretical goal is the prediction and control of behavior" (1913, 158).[19] As Pauly (1987, 174) reports, Loeb's model of biology as a discipline aiming at control of nature was a direct influence on his student, Watson:

> Watson's central innovation was to place the control of behavior at the foundation of psychology as a science. By arguing that control was knowledge, he broke down the barriers between the aims of pure psychology and those of behavioural technology. In this sense behaviourism was a model Loebian science, organized around the desire "to get life phenomena under our control." In both its positivistic methodology and its radical social claims it was the direct descendant of the ideas developed by Loeb in the early 1890s. For Watson himself, the engineering standpoint represented independence and excitement—from the level of laboratory innovation to that of power for social change. He saw himself in opposition to the received wisdom of his field; like Loeb he would cut through complexity with continuous experimental activity.

Watson's uptake of the idea of the simple reflex as a means to analyze behavior was directly influenced by Loeb (Pauly, 1987, 175). The key idea is that complexity is reduced ("cut through") and simplicity is generated—not discovered—through experimental activity. Another important point is that the "natural" state of things—how organisms are independently of experimental manipulations—is not privileged in the Loebian view.[20] Against this

19. And indeed, Titchener's (1914, 14) response to Watson's attack on the structuralist (introspectionist) psychology was to say that behaviorism is technology, whereas structuralist psychology is an actual science.

20. Pauly (1987, 199) summarizes that "the original organization and normal processes of organisms no longer seemed scientifically privileged; nature was merely one state among an indefinite number of possibilities, and a state that could be scientifically boring."

stance, the criticisms of reflexology that are centered on the artificiality of the experimental preparations do not have force. They merely avert to a bigger dispute between the proponents and critics of reflexology—a disagreement about what biological science fundamentally is and how it should be conducted.

### 3.3.2 Divergent Perspectives

We have just seen that some of the tenets of reflexology that appeared to its critics as simplistic and misguided can be more charitably regarded as simplifications subject to justification, not by their closeness to nature but by their utility within certain practical projects. This is not the attitude taken by all reflexologists toward their posits—Pavlov has already been mentioned as an exception here—but it is attributable to Loeb and the behaviorists influenced by him. However, there are some problems with this analysis, insofar as it depicts the most defensible version of the reflex theory as a pure instrumentalism and operationalism and casts its detractors as scientific realists able to interpret and evaluate theories only in terms of how close they are to the truth about nature.[21]

While it is true that behaviorist psychology developed in a context in which the fruitful application of science was highly prized,[22] the case for their subscribing to total instrumentalism is overstated. One complication is that Skinner himself does at times talk like a realist regarding simplicity ("order") in nature: "I never face a Problem which was more than the eternal problem of finding order. . . . Of course, I was working on a basic Assumption—that there was order in behavior if I could only discover it" (Skinner 1961, 112; quoted in Moore 2005, 100).

One might attribute this inconsistency to the fact that this was written long after Skinner's reflexology research of the 1930s. Still, it fits with the point that Skinner was throughout his career a follower of Francis Bacon's philosophy of science (Smith 1996, 65). The Baconian scientist is not a thoroughgoing antirealist since this vision of science and technology is guided by the maxim that nature, in order to be commanded, must be obeyed. The investigator, therefore, sees himself as striving to reveal the underlying properties and causal structures within natural systems in order that they

21. I have published this analysis elsewhere (Chirimuuta 2021).
22. As discussed by Boring (1950, 551), Mills (1998, 4), and Edwards (2016, 179).

may be bent to human purposes. Although in the first half of the twentieth century, various scientists did on occasions espouse stringent instrumentalism and operationalism, the practice rarely conformed to the purity of the preaching. Doing science involves dealing with and thinking about things, which in turn requires that these objects of investigation be conceptualized in some way beyond the operational definitions and equations linking observations. As Stein (1989) has observed, many scientists seem to alternate between realist and instrumentalist interpretations of their theories, depending on the stage of research. As much as Skinner might have thought the reflex theory free of any ontological commitment, it at least came with a prior conception of what an animal fundamentally is: a machine-like system whose behavior can in principle be determined through precise control of external stimuli.

This is a fundamental difference in perspective between proponents and critics of the reflex theory. For the purposes of drawing the contrast, I will focus here on Goldstein. As mentioned in section 3.2, Goldstein was willing to grant that research on the conditioned reflex afforded insights into human interventions on behavior. On his own conception of biology, however, the goal is to understand the "intrinsic natures" of organisms (1934/1939, 3–9), as they are independently of experimental influence. For example, in his book *Human Nature,* Goldstein draws a distinction between processes of training and drilling and asserts that the establishment of conditioned reflexes, being only a drill, is tangential to the proper task of neurobiology:

> *Training* attempts to achieve . . . [performances] by exercising the natural capacities of the individual organism and by bringing them to the level of greatest efficiency. The performances in question are related to the nature of the organism, and the intended effect is the highest possible adequate relationship between the individual organism and the environment. In *drill* the performance aimed at is not related to the nature of the organism. (1940, 135–136)

Talk of "natures of organisms" sounds quite metaphysical, and indeed this is the first objection raised in Skinner's (1940) critical comments on *The Organism*. Goldstein even uses the terms "potentiality" and "actualization," reminiscent of Aristotle's metaphysics of living beings.[23] But we should not, because

23. For instance, "The organism has definite potentialities, and because it has them it has the need to actualize or realize them. The fulfilment of these needs represents the self-actualization of the organism. Driven by such needs, we experience ourselves

of this, be misled into taking Goldstein's framework as a purely contemplative natural philosophy with no eye to application. Goldstein's practice as a neurologist did include the goal of therapeutic intervention on brain-damaged patients. It was different from the Loebian instrumentalism because Goldstein did not think that therapeutic success could be achieved through piecemeal, reductionist methods as opposed to consideration of the patient's personality, deficits, and capabilities in their entirety. Moreover, Goldstein took recovery to depend on the capacity of the patient to achieve self-actualization, to fulfill his or her nature, with the interventions of the physician being intended to aid this process.[24] As such, health and illness are defined in terms of whether the condition permits or impedes the self-actualization of the patient, not in purely objective, physiological terms (Goldstein 1959).

Instead of imposing the dichotomies of realist and instrumentalist, or pure versus applied science, to the perspectives represented by Goldstein and Skinner, it is better to draw the contrast in terms of their different stances toward agency,[25] both of the scientist and the organism under investigation. In the reflex theory, and especially in Skinner's behaviorist development of it, we are expected to be satisfied by knowing only, from the outside, the relationships between causes and effects, stimuli and responses, that go in and out of the black box that is an animal's nervous system. To the extent that inner tendencies, drives, and predispositions of the animal are to be noted, these are but the material for more exhaustive research into the observable effects of external causes, or the springboard for new forms of control, as in operant conditioning. This picture elides the agency of the creature under investigation, treating the person or animal as passively molded by its environment and subject to the agency of the investigator.

In contrast, for Goldstein, the inner tendencies of the organism, as well as behaviors generated aside from causal interventions of the investigator (all the behaviorist's stimuli and reinforcements), are a central matter for

---

as active personalities and are not passively impelled by drives that are felt to conflict with the personality" (Goldstein 1940, 146).

24. For example, Goldstein writes that the doctor-patient relationship is a situation "in which the one wants to help the other gain a pattern that corresponds, as much as possible, to his nature" (1934/1939, 449). This point is discussed further by Canguilhem (1989/2012, 63).

25. I use the term "agency" in an expansive sense, not restricted to conscious, goal-driven behavior of human beings.

research. Genuine agency is granted to the organism under investigation.[26] Because of the importance he gives to this notion of agency, which we may think of as the spontaneous, self-determined activity of the organism, Goldstein goes so far as to say that the effects of externally imposed causes, such as the stimuli in conditioning experiments, always need to be interpreted in relation to the whole organism and what it is trying to achieve at that time (1940, 133). For example, he writes that genuine habits can be acquired through conditioned reflexes only if they help the child toward the "actualization of its personality" (Goldstein 1940, 158).

We can think of this difference in terms of the degree of abstraction away from the dependency relationships that can be observed empirically. The naive observation of an animal like a rat finds it being both causally affected by its surroundings and also as a cause and initiator of events within those surroundings due to its acting to further its own needs. Behaviorism, to quite an extreme degree, abstracts away from, or black boxes, the causes initially perceived to be initiated within the mind of the animal, and it treats the animal merely as a conduit for effects passively received from the environment, especially those stimuli and reinforcements selected by the experimenter. Even though operant conditioning uses the self-initiated behavior of the animal, the point of the research is to bring the behavior as far as possible under external control.[27] This downplaying of influences stemming from the experimental subject simplifies the picture because there is no need to countenance the possibility of reciprocal causation between the animal and its surroundings, nor the possibility of hidden springs of agency (i.e., mental causation). A linear causal scheme—the stimulus-response pairing—papers over those tricky possibilities. Goldstein's framework, instead, envisages a more complicated picture of circular causality, both within the organism and

26. This contrast between theories that do and do not grant genuine agency to organisms tracks the difference within evolutionary biology discussed by Walsh (2015, chapter 10), between theories that posit organisms' goal-directed behavior as playing a role in the shaping of evolutionary processes, and nonagential theories, such as the Modern Synthesis. The parallel is not coincidental, since Goldstein was a member of the holist and organicist movements in biology, in which organism-level explanations are privileged, whereas the Modern Synthesis neglects organism-level explanations in favor of gene- and population-level ones.

27. Here, it is useful to think of behaviorism as offering externalist explanations in the sense discussed by Godfrey-Smith (1996).

covering organism-environment interactions. This, he believes, is closer to the observed phenomena: "We are dealing with a system in which the single phenomena mutually influence one another through a circular process, which has no beginning and no end. If, starting with the observation of reflexes, we try in unbiased fashion to understand the behavior of an organism, the facts everywhere force such a point of view upon us" (Goldstein 1940, 127).

This picture of the nervous system as comprising these paths of circular causation comes from taking in the complexity more as it stands and not attempting to simplify very much via experiments or modeling. But the price of this acceptance of complication is that the scientist's descriptions will not be as clear as they would be if more simplifications were introduced. Another of Skinner's charges is that Goldstein's theory of the nervous system lacks clarity. Yet the clarity and intelligibility of the reflex theory came from presupposing that holism, the characteristic of widespread mutual influence asserted by Goldstein, is not true.

We can also understand these two perspectives as being shaped by two kinds of real-world ambitions. The reflex theory, especially in its behaviorist incarnation, seeks control of an animal for the investigator's purposes, which are extraneous to the animal itself. A long-term goal of the program was social engineering through control of human behavior by the selection of environmental stimuli and conditioning regimes. The agency relevant to the ends of the program is only that of the investigator, who is external to the target of control. In contrast, the agency of the organism is fundamental to Goldstein's conception of the practical goal of research, which is to enhance self-actualizing processes within the patient in order to bring about recovery. These contrasting approaches to agency account for the striking difference in the importance granted to ecological validity within these two programs. (I include here "ethological validity," the question of whether an experimental paradigm allows display of the typical behaviors of the animal.)

For the reflex approach, lack of ecological validity is a problem only insofar as it might interfere with the goal of achieving external control. If the lack of it is due to simplifications that aid instrumental success, then the deficit is actually a benefit. But for Goldstein, lack of ecological validity is necessarily a problem because it will prevent the researcher being able to observe the behaviors that are the clues to the inherent nature of the organism. This risks both the epistemic aim of the research (learning about those natures) and the

therapeutic aims. The behaviorist emphasis on external control also puts a premium on discovering stable, manipulable, and linear causal relationships at the expense of recognition of the array of less stable, less manipulable, and more interconnected dependencies that are there to be observed under ethologically valid conditions.

### 3.4 The Afterlife of the Reflex Theory

In the previous section, we saw that the force of the attack on the reflex theory, as presented in section 3.2, is somewhat lessened if we acknowledge that the criticisms stemmed from a very different scientific perspective, without commonality in some fundamental assumptions. Scientific traditions come with their own standards of success, and we might take up a charitable interpretative stance by judging reflexology according to the standard held by its own practitioners, that of prediction and control. An immanent critique, so called; yet even by this benchmark, the reflex theory cannot be judged a success. Skinner is quite notorious for his claims made for the potential of behaviorism to bring about a utopian world through social engineering (Smith 1996). But these goals were, alas, not met. Indeed, other than the flourishing of the industries of marketing and advertising (Buckley 1989, chapter 8), reflexology does not have a hallmark success comparable with those of other areas of research associated with Loeb, such as Gregory Goodwin Pincus's invention of the contraceptive pill (Pauly 1987, 194).

An illustrative case in point comes from the work of Keller and Marian Breland, two former assistants on Skinner's so-called pigeon project. As Skinner (1947/1961, 227) wrote on the true agenda of experimental psychology, "the basic engineering problem is to acquire control. . . . . It is not a matter of bringing the world into the laboratory, but of extending the practices of an experimental science to the world at large." The Brelands took up this challenge, setting out to mass-produce novelty displays of conditioned behaviors, in a variety of animal species for commercial purposes. Yet, the theoretical predictions derived from laboratory experiments were overwhelmed by "animal misbehavior"—the failure of animals to learn simple, reinforced actions because of the interruption of instincts (Breland and Breland 1961). In a detailed study of this episode, Ramsden (2021, 89–90) explains that "as the Brelands took operant conditioning beyond the confines of the laboratory,

Skinner's tidy system began to fracture, and the 'nature' of the organism began to override the machine-like predictability of conditioned behavior."

To Skinner's consternation, the Brelands reached the conclusion that ethology—the study of the natural behaviors of animals within their environmental niches—is indeed indispensable in the study of animal psychology. What this attempted application of reflexology shows is that in the end, ecological *invalidity* bit back: when reflexology was extended beyond its niche experimental conditions, the oversimplified poverty of this conceptual framework undermined its ambitions.

From a vantage point built from over 100 years of further research on the central nervous system, it might seem incredible that scientists of stature ever believed that the reflex arc would be the one key to demystify brain and behavior. To our retrospective view, the reflex theory now appears obviously too simple to account for the phenomena that it was supposed to—it looks *simplistic*. We might ask what it was that made the reflex theory so appealing. There is some suggestion that its value was precisely in its being so simple—an attractive oversimplification. Hough, in his critical piece on the theory, writes of how its "diagrammatic clearness" has shaped researchers' "mental approach" to their problems, and how it naturally aligns with textbook expositions that begin with peripheral neuroanatomy and end with the physiology of the brain (1915, 408). Fearing observes, "The reflex arc is easily diagrammed in the textbook," but he also warns that "such a diagram readily forms the basis for a discussion of simple stimulus-response relationships, which is misleading even in connection with the simpler animal responses, and positively inapplicable to the more complex organic responses" (1930, 288). Karl Lashley (a onetime student of the behaviorist psychologist Watson) relates that the passing down of the textbook picture across generations has given it an entrenched, unquestioned, status:

> In the course of time there has been built up a *simple*, traditional, textbook account of the mechanism of reaction, prepared for students' consumption. Repeated copying from one text to another has crystallized it, and early instruction has given us faith in it. The original sources have been almost lost to view and with them the appreciation of the difficulties, the uncertainties, the many unsubstantiated assumptions which underlie every assertion of the classical account. (1931, 16, emphasis original)

A cautionary lesson here is that the scientists' instinctive tendency to head in the direction of simplicity—their *simpletropism*, so to speak—can sometimes send them in the wrong direction.

However, it is not that the reflex theory and its offshoot in behaviorism died out completely. Arguably, there is continuity in an ethos, stretching from reflexology to cybernetics, and from there to cognitive science and computational neuroscience, which also seeks the duplication of cognition in machines, as we saw with the work of Hull and Baernstein (1929) on conditioning in a mechanical device.[28] The key point is that while strict behaviorists, following operationist and instrumentalist principles, eschewed the positing of hidden variables within the brain or mind of the animal, there were always strands of reflexology, such as Pavlov's, that envisaged the strengthening and weakening of reflex connections within the nervous system as mechanisms underlying behavioral conditioning. There was, therefore, always a strand of reflexology amenable to the positing of connections and variables, causally mediating stimulus and response, that are not directly observable in behavior. This is the approach that carried through into later computational theories of the brain, especially the connectionist or neural network models. For example, Donald Hebb's (1949) "neural-assembly" theory of associative learning within the brain is considered an important precursor to connectionism.[29] He drew heavily from the reflex tradition, though clashing with behaviorist orthodoxy (Hebb 1960). Certainly, there is continuity between the methodologies of the reflexologists and experimental neuroscience today, such as the division between stimulus and response criticized by Dewey, and the various learning paradigms discussed by Machamer (2009). Reinforcement learning is all the rage both in AI research and neuroscience today, and its two parents are neoclassical economics and behaviorism (Castelle under review).[30]

Perhaps the most fundamental point is that current neuroscience is continuous with the reflex tradition in its adherence to the picture in which the inner agency of the animal is left on the margins of research. Indeed, the idea that a brain could be conceived as an organ in a self-actualizing system that seeks to realize its own nature now sounds somewhat unscientific,

28. See Carr (2020) for an account of the shift from behaviorism to cybernetics, and ultimately to cognitivism, with more emphasis on discontinuity.

29. Rosenblatt, the inventor of the Perceptron, an early neural network model, writes, "Hebb's philosophy of approach seems close to our own, and his work has been a source of inspiration for much of what has been proposed here" (1958, 407).

30. See also Haugeland (1978, 225), Dennett (1981, chapter 5) on the continuity between behaviorism and computational cognitive science.

and this shows how today's neuroscience is so definitively estranged from the organicist tradition of Goldstein.[31] In his own lifetime, Goldstein's theories were not considered unscientific or overly metaphysical. He had a successful research career, although it was disrupted by exile and emigration to the US. Goldstein's neurological work is now best known to philosophers through his collaboration with Adhémar Gelb, which features prominently in Merleau-Ponty's *Phenomenology of Perception*. Self-actualization did have one widely known descendent in Abraham Maslow's theory of the hierarchy of needs. But a neurobiology that paid equal heed to the agency of the experimenters and the subject of an experiment was not to carry the day. The spirit of the times, we may surmise, was with the externalizing approach of the reflex theory.

More concretely, we should appreciate how later computational theories grew out of the need to look inside the behaviorist's black box, without really abandoning the idea that the behaving animal is more of a passive machine than an active agent. Goldstein (1940, 129–131) discusses Tolman's (1938) experiments on rats' learning paths through mazes, noting his recognition of the need to posit internal factors ("intervening variables") in order to account for the data. Goldstein takes these observations to be vindication of his own rejection of reflexology. But it is telling that this is the same research often identified as the start of the cognitive revolution, the herald of the new era in which behavior would be explained in terms of manipulation of representations (such as spatial maps) by the onboard computers within the rat's brain. The successor theory to reflexology was therefore dominated by the analogy between the brain and a particular kind of machine, the computer. This analogy excludes the very feature of organisms that is central for Goldstein—their self-driven activity—for computers, like all other machines built to date, do not have agency of their own but are projections of the agency of the people who make and use them. They are highly sophisticated but inert arrangements of matter, which passively receive input and deliver output at their users' bidding, a bit like the simple reflex, awaiting a stimulus to trigger its response.

31. But see section 5.2 in chapter 5 on the return of ethological ideas within twenty-first-century neuroscience.

# 4 Your Brain Is Like a Computer

> I believe that the model is a useful and indeed unescapable tool of thought, in that it enables us to think about the unfamiliar in terms of the familiar. There are, however, dangers in its use: it is the function of criticism to disclose these dangers, so that the tool may be used with confidence.
>
> —Percy Bridgman (1927, 53)

The relationship between brain and computer is crucial to the interpretation of theoretical neuroscience, but it has received relatively little attention from philosophers of neuroscience. This chapter argues that much of the popularity of the brain-computer comparison can be explained by its utility as a way of simplifying the brain. I will argue that the relation between brain and computer should be understood as one of analogy, whereby comparisons are drawn between electronic systems—engineered to be somewhat functionally similar to biological ones—and the vastly more complicated organic brain. The implication, to be pursued in chapter 9, is that there are limitations to the scientific understanding of the brain and cognition, including consciousness, which stem from the radical abstraction imposed by the computational framework.

Section 4.1 gives a brief history of the development of the computational theory as it originated within the cybernetics movement, a program of research very much fixated on the building of self-regulating, lifelike machines, and on the perceived equivalences between machines and organisms. Section 4.2 describes how the brain-computer comparison motivates a distinction between the aspects of neuroanatomy and physiology that are *for information processing*, as opposed to *mere structural and metabolic support*. This makes research in neurobiology more efficient by channeling the possibly

endless delineation of biochemical interactions along the paths carved out by hypotheses arrived at by first assuming that the brain is a computer. However, the empirical successes of this research program, made possible because of this gain in efficiency, do not warrant the conclusion that the neural systems themselves compute the functions specified in the models, or that the brain itself is literally a computer. Section 4.3 presents an alternative to the literal interpretation of neurocomputational models, based on accounts of analogical reasoning in science and the formal idealism introduced in chapter 2. Finally, section 4.4 argues that my analogical interpretation fits better with philosophical accounts of modeling practice elsewhere in science, while avoiding metaphysical quandaries about the conditions for computational implementation. Even if the neurocomputational theory remains the only game in town, we should not be tempted to think that the undeniable differences between brains and computers do not make a difference to cognitive capacities and experiences, as they are to be found in living creatures.

## 4.1 From Cybernetics to the Computational Brain

> My aim . . . is simply to copy the living brain.
> —Ross Ashby (1954, 130)

A conclusion of chapter 3 was that there is continuity in a tradition that stems from early-twentieth-century reflex physiology to computational neuroscience as we know it today. A characteristic of this tradition is that it applies methodologies borrowed from the physical sciences to biological objects. Relatedly, it sees no obstacle in principle to the replication of the operations of neural systems in nonliving, artificial devices. However, reflexology, as exemplified in the research of Loeb, Sherrington, and Pavlov, was largely empirical rather than mathematical. It attempted to theorize the nervous system directly, as it were, not via the intermediary of physical models analogous to the nervous system. This section offers a very potted account of the rise of computationalism, the successor to the reflex theory, which came to serve as an overarching framework for building explanations of how the brain and nervous system give rise to cognition and behavior. As we will see, the cybernetics movement of the mid-twentieth century is the connecting link, for it was imprinted with ideas from reflexology while at the same time originating concepts for the neural network modeling that

dominates computational neuroscience today. Given the limited scope of this section, I am ignoring the path that leads to computational neuroscience from mainstream post-Chomskian cognitive science and symbolic AI. Furthermore, I am neglecting to discuss the economic and political conditions that nurtured cybernetics and computationalism in the UK and US.[1] To repeat the warning of chapter 1, my focus on abstraction in neuroscience is itself an abstraction from the myriad background factors that have in combination shaped the discipline as we know it today.

The topic of this chapter is really that of simplification through mathematization, the process that normally goes under the heading of modeling and idealization in philosophy of science. However, the situation is complicated in neuroscience because of the way that researchers have relied on the comparison between brain and computer to scaffold the process of mathematization. But it is not that the computer-analogy is the only path to mathematization. The most successful quantitative model in the history of neuroscience is arguably the Hodgkin-Huxley model, and this achieves a mathematization of the action potential by way of an analogy between the neuron and a simple electrical circuit (Levy 2014). Yet the dominant, quantitative approach in neuroscience does not reside with the characterization of the brain as a conglomeration of such circuits, but rather with the brain as a computer. This was not always the case. The first serious steps toward mathematization in neurophysiology did not involve analogies with artificial devices. I am referring here to the work of Nicolas Rashevsky, a Ukrainian émigré to the US who originally trained as a physicist and founded a laboratory for biophysics at Chicago University before World War II. He motivated the development of highly idealized, biologically implausible models of such targets as neural excitation and the conditioned reflex by invoking the methods that had been for centuries employed successfully in physics to manage the complexity of observed phenomena (Rashevsky 1938). He wrote in an article for the journal *Philosophy of Science*:

> The important thing in the mathematical method is to abstract from a very complex group of phenomena its essential features and thereby to simplify the problem. The more complex features are then taken care of gradually, according to the degree of their importance and complexity, as second, third, and higher approximations.

1. See, for instance, Galison (1994), Edwards (1996), Gerovitch (2002), and Medina (2014) for contextual histories of cybernetics in various world regions.

> True, by abstracting, we lose, so to say, contact with reality; but no harm is done by this as long *as we keep it in mind*. We thus see that the complexity of biological phenomena is rather an argument *for* the use of mathematical methods than *against* it. In the case of a simple phenomenon we may hope to understand it without the use of mathematics, by simple inspection. But in a complex case we are left hopeless without mathematics. (Rashevsky 1934, 178, emphasis in original)

This paper then continues with a long defense of idealization, showing its contribution to the past successes of physics. Despite Rashevsky's spirited support of such methods, he suffered continual criticisms from colleagues who were more knowledgeable in biology that his models were too unrealistic to be of any use, being rather disconnected from empirical observation and lacking in predictive power (Abraham 2004, 336).

Rashevsky's approach was to try to mathematize the nervous system directly. This contrasts with an indirect approach in which one builds a physical device in some respects functionally equivalent to a neural system and uses the mathematical description of the device as the first approximation of the original neural target. It was the indirect approach that eventually took center stage within the cybernetics movement, and this maneuver is nicely described by the primatologist Solly Zuckerman as follows: "It is unlikely that all this knowledge [of the mind-brain relation] is going to be obtained from a direct attack on the living organism. . . . Fortunately, however, recent developments in electronics allow us to represent at least part of the problem by analogy. Machines can be made—and exist—which exhibit some attributes of mental processes" (Zuckerman 1950, 30).

Zuckerman goes on to mention the negative feedback machines of Ross Ashby and Norbert Wiener as models of core functions of the nervous system, and then digital computers as models of memory. This essay appeared in a volume entitled *The Physical Basis of Mind*, following pieces by Charles Sherrington and E. D. Adrian, two eminences of British neurophysiology, both of which strike a rather pessimistic tone about the prospects for materialistic explanations of the mind. It is this that Zuckerman is reacting to: the "direct attack" of the physiologists may have disappointed, but the indirect strategy of the engineers was ready to take the lead.

Ashby and Wiener were, respectively, representatives of the UK and US movements in cybernetics.[2] Wiener was quite explicit about the employ-

2. See Kline (2015), Husbands and Holland (2008), and Pickering (2010) on these traditions.

ment of invented objects (i.e., "material models") as a method of simplification. For example, Rosenblueth and Wiener (1945, 316) write about the centrality of models in science, either formal or concrete, as simpler stand-ins to facilitate the understanding and control of things in the world that are, in general, too complex to be grasped without abstraction. Of concrete, material models, Rosenblueth and Wiener state that they amount to "the representation of a complex system by a system which is assumed simpler and which is also to have some properties similar to those selected for study in the original complex system" (1945, 317). Such models, when judiciously employed, can allow a phenomenon from a familiar field to replace one from an unfamiliar area of research, and often experiments are easier to carry out on the replacement system. Even though Rosenblueth and Wiener's analysis of material models recognizes that they are stand-ins for more complex, unfamiliar systems, the general tenor of Wiener's cybernetics was to elide differences between the operational principles of living organisms and self-regulating artifacts, as evidenced in the subtitle of his best-selling book *Cybernetics: Or the Control and Communication in the Animal and the Machine*.

The same elision is a feature of the writings of Ashby and another well-known practitioner of cybernetics in Britain, Grey Walter. Both were neurologists by training and hobbyist inventors. Ashby's *Design for a Brain*, first published in 1952, begins by asserting that the difference between the brain and machines invented so far is due to the contingent fact of the inventions not employing an operating principle crucial to adaptation in the brain. Thus, an aim of the book is "to show that by use of this principle a machine's behaviour may be made as adaptive as we please, and that the principle may be capable of explaining even the adaptiveness of Man" (Ashby 1954, 1). Incidentally, the book continues with an invocation of Pavlov's distinction between innate and learned reflexes. With Walter, the connection to Pavlov is even stronger. His first research project, when studying for a master's degree at Cambridge University, was conducted under the guidance of a scientist trained by Pavlov. In Walter's book *The Living Brain*, Pavlov's work is discussed with high appreciation, and a meeting with the man in person is mentioned. Walter's most famous invention was the "Tortoise," also known by the pseudo-Linnean name *Machina speculatrix*. These were simple, autonomous robots that gave the impression of purposeful exploratory behavior, by seeking lights and avoiding obstacles. Reflexology was obviously central to the design, as it was with the machine built for learning, the *Machina*

*docilis*.[3] Their movements were determined by a phototropism, and their model nervous system consisted of "two sense reflexes" (Walter 1953, 125ff). Walter presented them as a proof of principle that it is possible to get complex, unpredictable, and lifelike behavior from the right arrangement of mechanical parts. Although, Walter emphasizes, it would not be possible to build a model nervous system by reconstructing each neuron in an artificial medium, strides can be made by building quite simple models that have an overall similarity to some core functions of the brain (Walter 1953, 117–118). Walter did not present his devices as mere simulations or mimics of living creatures, but actually as instantiating the same "principles" that we find in nature, such as the "internal stability" of organisms (1953, 125–130). The interesting physiological quality of such devices, their claim to be "part of a mirror of the brain," is due precisely to the possible demonstration of these principles (131).

We will now turn to the work of McCulloch and Pitts, which is more connected with neuromodeling as we know it today. Single-cell neurophysiology and the engineering of digital computers both grew into maturity in the early 1940s and significantly influenced one another (Arbib 2016). These were all part of the heady cybernetics brew, and its intoxications can be detected in the landmark paper by McCulloch and Pitts (1943), "A Logical Calculus of the Ideas Immanent in Nervous Activity."[4] McCulloch and Pitts were part of Rashevsky's circle at Chicago, and a point of commonality is their deployment of highly idealized models of neurons. McCulloch and Pitts represented the cells as simple devices that sum over inputs and give all or nothing (i.e., digital) outputs, and these model neurons were, as Kay (2001, 598) puts it, "deliberately as impoverished as possible," comparable with frictionless surfaces and point particles. By showing that, under certain assumptions, small assemblies of connected model neurons could be taken to operate as logic gates, McCulloch and Pitts lent support to the claim that the brain *is*, literally, a computer. McCulloch (1965/2016, 169) would later write that "man-made machines

3. *M. docilis* was a version of the Tortoise with the addition of a "conditioned reflex analogue," a machine that, Walter argues, "behaves astonishingly like an animal" (1953, 178–179).

4. See Abraham (2016) and Dupuy (2009) on McCulloch's presence in the history of cybernetics, and Piccinini (2020, chapter 5) for an analysis of the 1943 paper.

are not brains, but brains are a very ill-understood variety of computing machines."[5]

The idealization of neurons as input-output devices is central to the computational theory of the brain as it has developed since then, even though other assumptions, such as digital coding, have been relaxed. The idea of computation via a neural network is foundational to the connectionist stream of AI, which has now evolved into deep learning.[6] To sum up, claims for the physiological relevance of these artificial, material models of the nervous system turn on the assumption that in spite of the manifest differences between the two kinds of things, at a certain level of abstraction, they share some essential commonality, such as the principles invoked by Grey Walter, or the computational structures that would be appealed to today. This is a strong assumption, and it calls for scrutiny; it is also the basis for the literal interpretation of these models. This interpretation of neurocomputational models has certainly been popular since McCulloch gave voice to it. Computer models of neural systems are taken to be more than mere models in the sense of simulations, like weather models, that represent but do not reenact the processes of nature. Instead, both neural circuits and the computational models of them are thought by the majority of neuroscientists to be doing the same thing—processing information (Miłkowski 2018) and representing states of things in the world beyond the brain (Churchland, Koch, and Sejnowski 1994). We will next consider the advantages offered by computational models, while remaining neutral, for the time being, on whether it is correct to interpret them literally.

---

5. John von Neumann is one well-known figure from the cybernetics movement who was concerned about overestimation of the similarity between neural systems and computational models. For instance, he writes:

> What is not demonstrated by the McCulloch and Pitts result is equally important. It does not prove that any circuit you are designing in this manner really occurs in nature. It does not follow that the other functions of the nerve cell which have been dropped from this description are not essential. It does not follow that there is not a considerable problem left just in saying what you think is to be described. (von Neumann and Burks 1966, 46)

6. A connecting figure is Frank Rosenblatt, inventor of the perceptron. He writes explicitly of these three-layered neural network models being extremely simplified in comparison to actual neural networks, but that they should exemplify "some of the fundamental properties of intelligent systems in general" (1958, 387).

## 4.2 Simplification and the Computational Brain

> What is the importance of machines in the philosophy of mind? I think that machines have both a positive and a negative importance. The positive importance of machines was that it was in connection with machines, computing machines in particular, that the notion of functional organization first appeared.
> —Hilary Putnam (1973/1997, 97)

One might ask why computationalism went on to become the dominant theoretical framework for neuroscience.[7] This is a broad question that deserves a complex answer, referring to historical and sociological factors and staying sensitive to differences between subspecialities within the science. However, for the purposes of this chapter, I offer a simple answer, which boils down to one characteristic of computationalism—that it provides neuroscientists with a very useful, perhaps indispensable means to simplify their object of investigation. More specifically, my claims are that (1) computationalism permits a distinction between the functional ("information processing") aspects of neural anatomy and physiology and what is merely background support,[8] thereby justifying the neglect of countless layers of biological complexity; and (2) computational theory, in giving the specification of neural functions, provides an ingredient lacking in purely mechanistic approaches to neurobiology (like the reflex theory), without which it would be far more difficult to separate relevant from irrelevant causal factors, and hence to state when the characterization of a mechanism is sufficiently complete.

---

7. Note that this should not be confused with the issue of whether the dominant mode of explanation in neuroscience is mechanistic or computational. Those on the mechanist side of this debate, such as Kaplan (2011), acknowledge the importance of computationalism in theoretical neuroscience and argue in addition that computational models provide mechanistic explanations. In section 4.2.2,. I briefly argue that the contribution of computational models is distinct from that of mechanistic models. See Chirimuuta (2014, 2018) for the full arguments.

8. This is often referred to as "metabolic support," but this term is employed with a wide meaning, including not just intracellular metabolic processes or activity of glial cells, but also the vascular and immunological systems that are integrated with the brain. Haueis (2018) also discusses the distinction between cognitive and noncognitive functions of the nervous system.

### 4.2.1 Putting Function in the Foreground

As Kurt Goldstein (1934/1939) repeatedly argued, most of the supposed background factors within an organism are highly relevant to the state of the whole creature in ways that experimental biology largely ignores. Yet, even if, like Goldstein, one doubts that there is an absolute distinction between the phenomenon of interest and background factors, it must still be acknowledged that it is appropriate for the biologist to delineate a phenomenon by means of selective attention, as with a visual image affording figure-ground separation. It should not be news to anyone who has observed the practice of science that part of the task (and art) of devising a new experiment or explanation is the drawing of a distinction between the target of investigation and the additional factors that can reasonably be classified as background conditions. For a system of any complexity (which is all the systems studied in biological sciences), the outcome of the endeavor largely turns on the aptness of the distinction.

My contention here is that much of the value that the computational framework provides to neuroscience is in the distinction that it supports between the putative function of a neural system, the information processing that makes up the target of investigation, and the residual features that can be placed in the background as mere support systems. The classic characterization of the neuron as a device that gathers inputs at the dendrites, calculates a function, and delivers an output (i.e., a number of spikes sent down the axon) is the most prevalent way that this distinction has been put to use in neuroscience. While this picture is much broader than the McCulloch and Pitts formalism, they can be credited with disseminating the idea that the single neuron is an input-output processing unit thereby giving neuromodelers an excuse for abstracting away from most of the cell biology underlying the reception and generation of action potentials. This was the opinion stated by the AI luminary Seymour Papert:

> The liberating effect of the mode of thinking characteristic of the McCulloch and Pitts theory can be felt on two levels. . . . On the local level it eliminates all consideration of the detailed biology of the individual cells from the problem of understanding the integrative behaviour of the nervous system. This is done by postulating a hypothetical species of neuron defined entirely by the computation of an output as a logical function of a restricted set of input neurons. (1965/2016, xxxiii)

The utility of this simple picture goes a long way toward explaining the persistence of the "neuron doctrine"—the thesis that neurons are the

functional unit of the nervous system, whose exclusive job it is to receive, process, and send information—in the face of countervailing empirical findings (Bullock et al. 2005).[9]

The strategy, just outlined, for isolating the functional begins with the concrete neural system and abstracts away from it all features classified as nonfunctional background support. Another modus operandi is to start with the specification of a cognitive task (such as detection of edges in a photograph), then to consider what computations would be needed to achieve the task, and next to build an artificial system (i.e., a computational model) that performs it. With the model in place, the final step is to use it as a template or map when looking for activation and connectivity patterns in the brain that are responsible for the performance of this task. This strategy is described by neurophysiologist Jerome Lettvin in response to the criticism that computational models used in neuroscience—such as connectionist networks—lack similarity to actual neural systems:

> Even if ideally one could record from any element or part of an element in situ, it is not in the least obvious how the records could be interpreted.[10] To a greater degree than in any other current science, we must know what to look for in order to recognize it. . . .
>
> This is where a prior art is needed, some understanding of process[11] design. And that is where AI, PDP [parallel distributed processing], and the whole investment in building [neurocomputational models of intelligence] enter in. Critics carp that

9. Cao (2014) recommends going beyond the neuron doctrine to consider synapses and glia also as functional units of the nervous system. This raises the question of the technical feasibility of gathering synapse-resolution data of neural responses, and attempting to model the brain in such a fine-grained way (noting that each cortical neuron receives, on average, tens of thousands of inputs). If the neuron doctrine provides a good enough framework for modeling the brain, which is especially useful for the activation patterns associated with observable behaviors (e.g., perception, learning, and decision making) that involve large populations of neurons, then there is little reason to attempt the impossible and replace neurons with synapses as the postulated fundamental signaling units, even if it is acknowledged that in the brain, much functional activity does occur within synapses. In later chapters, I take up the issue of the importance of these details that are relegated to the background in the classic neurocomputational picture.

10. This, incidentally, is a point made vivid in a paper by Jonas and Kording (2017), "Could a neuroscientist understand a microprocessor?"

11. Lettvin often uses the word "process" in his characterization of the engineering stance in neuroscience. It should not be confused with the notion of 'process models' in psychology or with other kinds of mechanistic models.

> the current golems do not resemble our friends Tom, Dick, or Harry. But the brute point is that a working golem is not only preferable to total ignorance, it also shows how processes can be designed analogous to those we are frustrated in explaining in terms of nervous action. It also suggests what to look for. (Lettvin 1988/2016: xvii–xviii)

If anything, the problem of "knowing what to look for" is more acute now than when Lettvin wrote this. In the last twenty years, the increase in the variety of tools and methods for observing neural activity (from single cells to whole brains) has surprised and delighted many. However, the downside of these advances is that they bring to light kinds of complexity that were not previously apparent, especially at subcellular scales. This is how neuroscientist Yves Frégnac describes the situation: "Each overcoming of technological barriers opens a Pandora's box by revealing hidden variables, mechanisms, and nonlinearities, adding new levels of complexity. By reaching the microscopic-scale resolution, advanced technologies have unveiled a new world of diversity and randomness, which was not apparent in pioneer functional studies using spike rate readout or mesoscopic imaging of reduced sensitivity" (2017, 471).

Frégnac points to the need for a greater understanding of how mesoscopic and macroscopic regularities emerge from the processes observed microscopically. But a wider point is that if artificial systems, sharing none of the microscopic details of the neural ones, can be built to duplicate some specific functions, albeit roughly, then there is an acceptable excuse for keeping shut the Pandora's box of subcellular neurobiology.

### 4.2.2 Mechanism and Function

We will now consider the mutually supportive relationship between computational modeling and mechanistic investigations of neural systems, ones that aim to uncover regular, delimited clusters of causal interactions that explain physiological and cognitive phenomena, such as synaptic plasticity and particular types of memory (Craver 2007; Craver and Darden 2013). (See section 2.2.2 in chapter 2 for more information.) In response to a criticism of the mechanistic account of explanation, which takes issue with the favoring of more detailed descriptions of mechanisms as providing better explanations than less detailed, "sketchy" ones, Craver and Kaplan (2018) emphasize that their account has never favored more detailed explanations per se, but has only suggested that explanations

including more of the *relevant* details may have the edge over less complete ones. Even if we accept the picture of there being ontic mechanistic explanations, which include only relevant causal details, and exist independently of human scientists, there is still a question of how scientists learn to distinguish the relevant from the irrelevant factors to produce adequate epistemic explanations—that is, true enough representations of those neural mechanisms. In any biological system (and the nervous system especially), one finds a densely interconnected causal web with many layers of structural intricacy, as well as patterns of effect across various spatial and temporal scales. Craver and Kaplan appeal to a "mutual manipulability" criterion that is clear and unobjectionable on the face of it.[12] However, if their norms for explanation are to be applied to practice, it becomes hard to see how only the causal factors in a neural system relevant to a particular phenomenon—as opposed to background factors not constitutive of the mechanism itself—could be isolated if only the mechanistic perspective is employed. An individual neuron will have thousands of feasible targets or handles for experimental manipulation—for example, the different kinds of ion channels, which could be blocked on select portions of the membrane; the various receptors that could be agonized or antagonized; and the countless proteins transcribed in the cell that could be targets of genetic manipulation. One needs to multiply this list of causal variables by 10 or 100 if the system comprises a small population of neurons. One faces a combinatorial explosion of experiments that would be needed to determine the independent causal relevance of each of these factors in a putative mechanism. But, of course, neuroscientists do not plan sequences of experiments according to brute force search. When designing an experiment with the aim of determining which of the many causal variables present in a system are crucial to its behavior (given a certain explanatory question), how does a neuroscientist know which ones to select from an inexhaustible list? We should think of hypotheses regarding the information-processing functions of neuronal structures as heuristics that drastically reduce this search space.

12. Craver and Kaplan write, "A factor is constitutively relevant when (ideal) interventions on putative component parts can be used to change the explanandum phenomenon as a whole and, conversely, interventions on the explanandum phenomenon as a whole can produce changes in the component parts" (2018, 20).

For example, at a fairly high level of abstraction, only net excitation minus inhibition is the causal factor relevant to determining whether a neuron's firing rate will increase or decrease. This abstraction disregards the kinds of neurotransmitters found at the synapse, receptor types, and location of synapses.[13] And, of course, this is the kind of abstraction fostered by the neuron doctrine and fundamental to the McCulloch and Pitts vision of the brain as a computer in which the logic gates are built from neurons.[14] In essence, without any prior assumption in place about what the neuron's function is, and about what aspects of physiology and anatomy are relevant to it, the search for relevant causal factors would have to proceed by brute force or be guided by mere prejudice. What this indicates is that the functional, informational processing perspective on neural systems is an indispensable complement to the mechanistic approach in neurobiology.[15]

13. Craver and Kaplan (2018, 19, n16) appeal to the purely causal notion of "screening off" to address the question of why complete explanations do not have to go down the full reductionist route, referring to the most fundamental particle such as quarks. The idea is that "low-level differences" will be ignored if they "make no relevant difference once the higher-level behaviour is fixed." It is important to appreciate that for the kind of neuronal details discussed here, screening off should not be expected to occur—that is, these excluded details *do* causally affect neuronal behavior in ways that are not fully summarized by the higher-level variables of net excitation and inhibition. This implies that a search for "relevant details" that proceeded only by the method of searching for higher-level causal variables to replace lower-level ones would not result in the abstractions found most useful in computational neuroscience.

14. There is latitude in the abstracting assumptions. I have described a case where total inhibition is subtracted from total excitation, whereas McCulloch and Pitts (1943, 118) posit that inhibitory input at any one synapse will cancel out the effects of excitation.

15. Although he does not focus on computational explanations, this is actually the conclusion reached by Craver (2013, 155), who writes (emphasis in original):

> Identification of functions is a crucial step in the discovery of mechanisms. We no longer speak of mechanisms simpliciter, but rather as mechanisms *for* some behavior. Mechanistic descriptions thus come loaded with teleological content concerning the role, goal, purpose, or preferred behavior of the mechanism. This teleological loading cannot be reduced to features of the causal structure of the word, but it is ineliminable from our physiological, and particularly neural, sciences, precisely because their central goal is to make the busy and buzzing confusion of complex systems intelligible and, in some cases, usable.

His perspective-dependent account of functional attributions is different from that of other mechanists like Piccinini (2020), who are straightforward realists about "teleo-functions."

Another way to make this point is just to say that the boundaries around neural mechanisms are not simply there in the brain, discoverable through a small enough number of causal experiments. There are many justifiable ways for the neuroscientist to carve up the subsystems of the brain into mechanisms and separate them from background conditions. The computational perspective is one approach that has suggested to scientists a particularly fruitful set of delineations.

The difference between the physicist's and the engineer's outlook is a useful analog to the difference between mechanistic and computational approaches in neuroscience (Fairhall 2014). When one considers the structures of the brain as a physical system, it is a nexus of causal interactions in which considerations of function are alien; in contrast, the notions of design and function are inherent to the engineering approach, from which it is natural to regard the brain as a target of reverse engineering (Sterling and Laughlin 2015). With the computational approach, one begins with the consideration of what the neural system is *for*, and the question of how that function is achieved is addressed after this is settled. When dealing with complex, biological systems, any attempt to employ only the function-less physics stance would quickly lose its way among tangled causal details.[16] It is the task of theory in science to provide a synoptic view of the subject matter, and thereby suggest pathways for future experimentation.

In current neuroscience, the computational theory is best developed. I do not claim that this is the only possible theory of the nervous system, and I certainly am not claiming that computational theory should float free from experimentally derived facts about neural mechanisms. Ideally, these two perspectives are mutually constraining and complementary.

### 4.3 Two Interpretations of the Brain-Computer Relationship

The building of machines in order to elucidate processes underlying vital functions, including cognition, is a strategy that goes back at least to the

16. This point is made by the neurologist Francis Walshe (1961, 131). See also Knuuttila and Loettgers (2014, 79) on the contrast between physics- and engineering-based approaches within synthetic biology research. We might also be reminded of the "design stance" (Dennett 1987), though this comes with strong assumptions of adaptationism and optimality.

automaton-makers of the eighteenth century.[17] But an open question here is whether, to understand the efficacy of this pattern of investigation, one must resort to a literal interpretation of the artificial models as duplicating and thereby bringing to light the very same process or function occurring in the living system, or if one can still make sense of the research strategy by taking the machine-organism relationship as one of analogy. That is, saying that the organism is *like* the machine in some way, but making salient the numerous differences (disanalogies) that limit the appropriateness of the machine-organism comparison to the narrow domain of the phenomena explicitly modeled.[18]

I do think that the literal interpretation is the majority view among neuroscientists—given a wide enough definition of computers as systems that encode input, manipulate those representations, and transform them into output according to some specific algorithm (Marcus 2015, 209). Complaints from neuroscientists that the brain is *not* a computer usually just make the point that the brain is not a digital, serial machine, while still asserting that the brain is some other kind of computer. Many assert that any disanalogies between information processing as it occurs in electronic and neural tissue do not present an obstacle to computational replications of brain processes, and these will eventually provide explanations of cognitive capacities, bringing about the reproduction of those capacities in machines.[19]

---

17. As Canguilhem (1963, 510) describes, "Texts, taken from Quesnay, Vaucanson and Le Cat, do not indeed leave any doubt that their common plan was to use the resources of automatism as a dodge, or as a trick with theoretical intent, in order to elucidate the mechanism of physiological functions by the reduction of the unknown to the known, and by complete reproduction of analogous effects in an experimentally intelligible manner."

18. A potential misinterpretation of section 4.2 may push one toward the literal interpretation. If one thinks that the brain—like a digital computer designed to be indifferent to variation in magnetic grains in a hard drive, for instance—is a device that ignores its own complexity, so to speak, then an abstract computational description of the system can be equally, literally true of the brain as of the machine. However, the point of section 4.2 is to explain how and why neuroscientists use computational models to ignore the complexity of the brain, leaving it a live possibility that those details *do* matter to cognition in animals (see section 4.4 and chapters 9 and 10).

19. This strong view is best exemplified in the work of researchers at the interface between neuroscience and the deep learning style of AI, such as Hassabis et al. (2017) and Yamins and DiCarlo (2016). It subscribes to the *computational theory of mind* much discussed in the philosophy of mind, psychology, and cognitive science. The

Drawing on the account presented in chapter 2, the literal interpretation of neurocomputational models should be taken as an instance of *formal realism* (see section 2.1). It supposes that the structure represented in the computational model obtains in the neural system independently of the scientist's experimental and theoretical work. Moreover, this computational structure is taken to be essential to the brain having the cognitive capacities that it does. By itself, formal realism does not entail the multiple realization of a form in different material substrates. But given that the forms in question here are computations, multiple realizability does follow because of the fact that the same computation (e.g., multiplication of 653 x 10) in principle can be performed by a variety of material realizers, including an artificial computer (mechanical or electronic) and biological tissue. This picture of abstract mathematical forms being realized in an array of material substrates—breathing intelligence into them, one might say—has had long appeal in the history of computation.[20]

While the formal realist takes for granted the brute existence of mathematical forms, which are realized equivalently in brains or computers, the formal idealist takes the mathematical forms represented in computational models of the brain not to be straightforward discoveries regarding mathematical structure or information processing in the brain, but rather constructs developed through an arduous process of experimentation, model building, and analogical reasoning. The proposal is that the mathematical structures that make the brain intelligible to the scientist, as an organ whose function is to process information, are to some extent imposed onto the neural system by the scientist and should not be taken as straightforward discoveries of mathematical forms inherent in the system. Since, by hypothesis, neurocomputational models are not discoveries of the inherent computational capacities of the brain but are as abstract and idealized as any other models in science, an analogical interpretation of these models is more appropriate than a literal one.

My analogical interpretation is an alternative to the literal interpretations of neural-computational models.[21] To preempt the worry that there

implications of my interpretation of neurocomputational models for the computational theory of mind will be discussed in chapters 9 and 10.

20. See for instance, Morar (2015, 126) in relation to Leibniz.

21. Another alternative to literalism is the deflationary approach of Cao (2019). Her resistance to literal interpretations of neurocomputational models turns not on

is no substantial difference between the literal and analogical interpretations, I specify at the outset that I am not defining analogies as isomorphisms or homomorphisms that obtain between the brain and its model,[22] since with that definition, the analogical relationship would amount to the instantiation of the same structure (i.e., a computational structure) in the neural system and the model. It would follow, on the assumption of a *mapping* or *structural* account of computational implementation (see Sprevak 2018), that there would be no daylight between the literal and analogical interpretations of neurocomputational models. This is because the literal interpretation is the claim that the neural system and its model compute approximately the same function. From my conception, to say that a model should be interpreted analogically is to say that the target is *like* the model in some way that may turn out to depend on the interests of the scientists and the techniques that they employ. The crucial point is that the structure in the brain found to be relevantly similar to the computational model is not assumed to be an inherent, human-independent fact about the brain. Rather, it is an *ideal pattern*, a regularity in the neural data whose features are determined not only by the brain under investigation, but also by scientists doing the investigating (see sections 2.1.1 and 2.2.1 in chapter 2).[23]

In the classic account, Hesse (1966) charts the structure of analogical reasoning in science using diagrams that compare two systems (the analog source and target) along vertical and horizontal axes. For example, the analogical inference that Mars, because of its similarities with the Earth, *may* support life is depicted in figure 1.

Figure 4.2 presents an example drawn from the work of David Marr and Shimon Ullman, whose framework for computational modeling in

abstraction, but on the failure of these models to support a robust notion of representational content. The issue of neurorepresentations will be taken up in chapter 6.

22. See Knuuttila and Loettgers (2014, 87) on why analogical reasoning in science goes beyond the isolation of structures that map from model to target. The analogical interpretation should also not be confused with analog computation or the analog-model account of the brain (Shagrir 2010).

23. This proposal bears interesting similarities with Sprevak's account of computational implementation, which argues that all descriptions of material systems as computing systems are idealized representations of the concrete physical processes (unpublished manuscript).

| Earth (Source) | | Mars (Target) |
|---|---|---|
| | **Known similarities** | |
| Orbits the Sun<br>Has a moon<br>Revolves on axis<br>Subject to gravity | | Orbits the Sun<br>Has moons<br>Revolves on axis<br>Subject to gravity |
| | **Inferred similarity** | |
| Supports life | ==> | *May* support life |

**Figure 4.1**
A schematic for analogical reasoning, after Bartha (2016).

neuroscience has been highly influential.[24] Because of the "behavioral" similarity observed across the systems (the ability to detect edges), and the similarities in patterns of activation in response to edges, the analogical inference can be made that the neuronal activity in the cat's early visual system—the response pattern of retinal ganglion cells (RGCs) and neurons in the lateral geniculate nucleus (LGN)—is like the computation of a Laplacian of Gaussian function.[25]

In addition to the observation of similar overall behavior, the dissimilarities in the material substrates of the systems may be noted and the *abstractive inference* made that these dissimilarities are *not* relevant to the scientist's investigation of the capacity for edge detection. Importantly, the point is not that the differences in implementation are irrelevant tout court, but that they can reasonably be ignored for this kind of investigation of this particular capacity. The possibility of such abstractions is a precondition for Marr's (1982, 25) distinction between the levels of computational theory and algorithm and that of implementation. This kind of abstractive inference fits with my account of how computational models aid neuroscientists in the simplification of the brain, since the abstractions of computational

24. See Marr and Ullman (1981); Marr (1982, 54–65).
25. See Egan (2017) and Shagrir (2010) for discussions of this example that instead endorse the literal interpretation. Likewise, Marr (1982, 64) makes the stronger (but hedged) claim that these neurons *are* computing the function: "it is not too unreasonable to propose that the $\nabla^2 G$ function is what is carried by the X cells of the retina and lateral geniculate body, positive values being carried by the on-center X cells, and negative values by the off-center X cells." This amounts to a formal realism. I do not suppose my weaker interpretation to be the one put forward by Marr himself.

| **Computer (Source)** *Laplacian of Gaussian Model* | | **Brain (Target)** *LGN or RGC neurons in cat* |
|---|---|---|
| | **Observed similarities** | |
| Detects edges in a photo. Characteristic peaks of model output for onset and offset of edges. | | Responds to moving edges. Average increases in neural activity for onset and offset of edges. |
| | **Observed dissimilarities** | |
| Peaks for onset and offset are *symmetrical.* Implemented in digital computer. | | Peaks for onset and offset are *asymmetrical.* [Ignored] Is an electrically excitable cell. |
| | **Inferred similarity** | |
| Model computes Laplacian of Gaussian function. | ==> | RGC and LGN neurons can be modeled as computing Laplacian of Gaussian function. |
| | **Abstractive inference** | |
| ==> | *Differences in implementation are not relevant to the particular capacity investigated here.* | |

**Figure 4.2**
Abstractive pattern of analogical reasoning. The text in gray indicates features that are observable but ignored for the purposes of modeling and explanation.

neuroscience are licensed by this sort of analogy. But by putting my account of abstraction and simplification in the context of a nonliteral, analogical approach to interpretation of neurocomputational models, there is no commitment to any "computational essentialism" about the brain, or to the idea that the biological functions of the brain are computational functions and therefore *must* be multiply realizable.

According to formal idealism, the relevant similarities between the model and target are not simply there, waiting to be discovered by the scientist; rather in some respect they are constructed or massaged from equivocal data. Some details from our example will reinforce my proposal. Figure 4.3 illustrates the correspondence between the Laplacian of Gaussian model and the neural data (Marr and Ullman 1981, 165; Marr 1982, 65). If one examines the average neural traces depicted here, and in addition the data presented in the original neurophysiology papers from which these examples were taken (Rodieck and Stone 1965, figures 1 and 2; Dreher and Sanderson 1973), it is

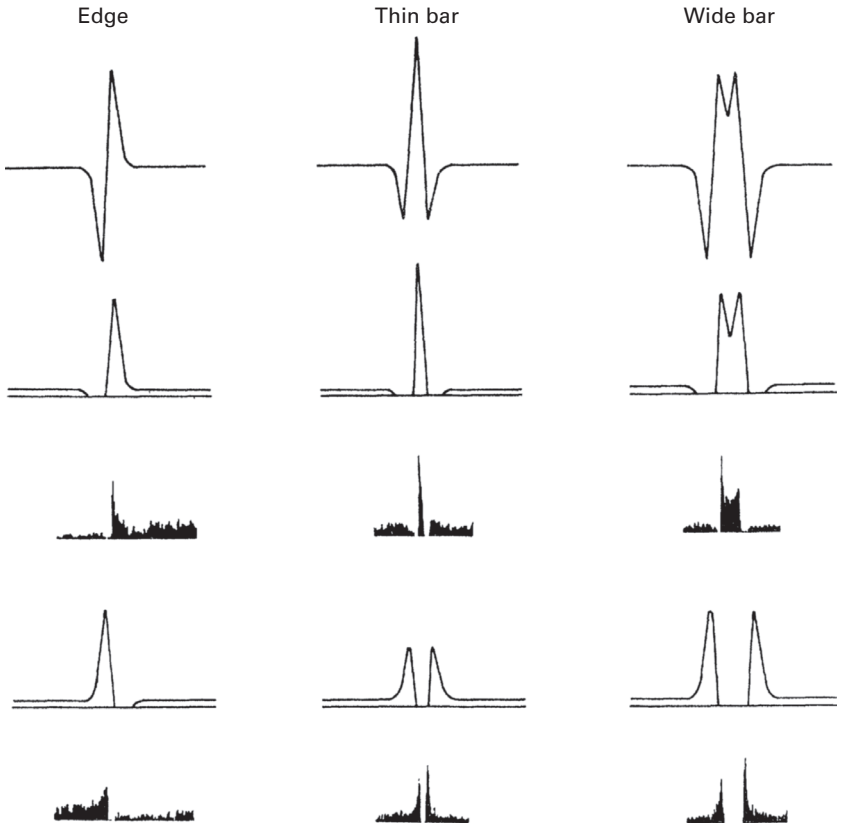


**Figure 4.3**
Comparison between a Laplacian of Gaussian model and neural data. The neural data indicate an unequal treatment of light versus dark edges and bars that is not captured by the model. From Marr and Ullman (1981, 165); Marr (1982, 65).

striking that there is a pattern of the neural response that goes unnoted by Marr and is not captured by the model—the asymmetry of peak response, depending on the polarity of the visual stimulus, and whether the bar stimulus is being swept onto the neuron's receptive field or is leaving the field. For example, the first column of figure 4.3 shows that a light edge on a gray background generates more neuronal response than a dark edge, whereas the model response is exactly equal. The general point is that the positing of an analogy—here, that a common pattern of activation occurs in the model and in the neurons—requires selective attention to certain similarities and the ignoring of dissimilarities. This is a matter of judgment by the scientist, and

the data do not usually, by themselves, force one choice over all others; Marr *could* have taken the asymmetry to be a relevant part of the neuronal behavior and come up with a mathematical model that captured this. One should not think of the structure described in any particular model as simply duplicating a structure that is preexisting in nature, as a formal realist would assert. Instead, what is strictly common between the neural system and model is an *ideal pattern* that does not exist independently of the scientists' activities, data processing, and theoretical choices. My point is, basically, that as a consequence of the complexity of the neural events, and therefore of the data sets gleaned from them, the determination of signal versus noise is not unambiguous, and for that reason, the data sets afford numerous plausible mathematical descriptions. Marr treated the asymmetry in the responses as noise and left it out of his model; another scientist would have been equally justified in treating it as a signal, a feature to be included in the model.

Formal idealism does not suppose that the finding of structure in a target of investigation is purely made up and then projected onto the data, but it does take it to be the result of the researcher's interaction with the target such that the human-dependent element of the structure can never be fully removed. This is reminiscent of the way that the visual system finds shapes in what might appear to be very disordered stimuli, as demonstrated with certain images in Gestalt psychology. While visual Gestalts are in most cases formed involuntarily, I emphasize that the scientist has a certain amount of latitude and choice in the determination of the patterns that are the target of modeling because these depend on methods of data collection, data processing (at minimum, averaging), and style of representation. The role of experimentation in this interaction, toward the shaping of ideal patterns, will be explored in chapter 5.

Another way of describing the difference between formal realism and idealism is that in the first case, the abstractions of computational neuroscience are presented as if the work of the researchers has been to pare away all the extraneous neurobiological details to find the essence (form) of the brain qua information processor. This is something like picking all the leaves off a tree and asserting that the bare trunk and branches are all that is essential to the tree. In contrast, the formal idealist does not assert that the computation described in the model conveys the essential features of the neural-cognitive system. The abstractions introduced by the model are taken to be there for the convenience of the scientist (i.e., to provide an economical

representation that does not overload the scientist with a million details), rather than a means by which the most essential structures of the brain are revealed. A botanist would not insist that the leafless representation captures all that is indispensable to explaining the capacities of a tree; nonetheless, a pared-down representation would be useful, and good enough, for many purposes.

## 4.4 Assessment

In the previous section, I presented the two ways of interpreting neurocomputational models without dwelling on the reasons why my analogical interpretation should be preferred to the more common literal one. This is the task left for the final section of this chapter, where I argue that, first, the literal interpretation leaves theoretical neuroscience hostage to there being an acceptable answer to the philosophical question of computational implementation; and second, the literal interpretation prematurely forecloses consideration of neurobiological details, not shared with computers, that may well be prerequisite for cognition as it occurs in animals.

### 4.4.1 Why Not Interpret Literally?

The analogical interpretation is a doctrine of restraint: it declines to infer from the success of the computational approach in neuroscience that the brain really is a computer—an organic device performing calculations to which the neurocomputational models provide a closer or wider approximation. For one thing, it is wise to remember that a model is only ever a model—that a model can never deliver the full, complete truth of its representational target. We should not be seduced into thinking that neurocomputational models are an exception here. My proposal brings the discussion of neurocomputational models in line with accounts of abstraction and idealization elsewhere in the philosophy of science, where it is recognized that models are always distortions, and as such cannot be read at face value.

Moreover, the literal interpretation requires taking on some difficult metaphysical commitments and philosophical challenges. First, formal realism here presupposes a realism about mathematical structures normally associated with Platonism—the existence of mathematical abstracta outside of space and time. At the same time, mathematical operations are taken to be realized in the material brain, which is located in time and space. The standard answer

to the question of how the abstract and concrete can be related in this way is to point to the concept of implementation, but this raises its own difficulties (Putnam 1988; Searle 1992; Godfrey-Smith 2009). The formal realist claims that a brain area implements some computations specified by neuroscientists. The *triviality objection* to the computational theory of mind asks what warrants the claim that the brain implements *those* ones, but not any of the countless other computations that also map onto a physical system like the brain (Sprevak 2018). The formal realist must appeal to a theory of implementation that would allow her to rule out the trivial computations but retain the claim that the brain does implement certain computations. The pressing challenge is to give an account of the implementation of computations in concrete systems that does not imply pancomputationalism (as conceded by Chalmers 2012), while showing how the computational level of explanation is somewhat autonomous from the implementational one (Ritchie and Piccinini 2018). The selling point of my alternative interpretation is that it does not come with the burden of needing to solve such problems.

The formal idealist is not faced with this challenge because she is not asserting that the brain implements any computations, but rather that it is useful to model the brain as if it were computing. Compare our case with the interpretation of the liquid-drop model of the atomic nucleus (Morrison 2011). A literal interpretation would say that the nucleus simply *is* a liquid drop. A proponent of this interpretation is then committed to explicating what it is that makes liquids different from solids and what the liquidity of the nucleus fundamentally is. Moreover, there is the difficulty of accounting for the predictive success of alternative models of the nucleus that do not make this assumption. In contrast, someone following my proposal can merely say that the nucleus is *like* a liquid drop in some way, that making this comparison is useful to nuclear physics, and then put questions about the metaphysics of liquidity to one side. All that needs to be assumed is that some things are uncontroversially and pretheoretically liquid drops, or computers,[26] while theoretical inquiries about the nature of liquidity and computation are tangential.

It is to be noted, of course, that some current theories of implementation have been tailored to address the question of how the brain can be said to

26. Factory-made computing machines are computers, uncontroversially and pretheoretically.

compute biologically relevant functions, and of course the formal realist may refer to them (see, e.g., Piccinini 2020). I will point out here that no theory of implementation is uncontroversial, and appealing to such a theory cannot by itself make the case for formal realism over formal idealism. One positive argument for formal realism would be to say that if the computational description is a useful simplification—and a good analogy—it must be that it does a good job of capturing the structure of the target system. That, then, is reason to think that the system is literally computational. Conversely, if the target system is not literally computational, then the computational approach must provide a poor simplification and a misleading analogy. But this argument simply assumes that models work—provide useful simplifications—to the extent that they faithfully represent structures that exist in the target system, an assumption at odds with so much work in the philosophy of science on modeling, abstraction, and idealization.

Most of the models that scientists employ, such as the liquid-drop model of the nucleus, represent their target in ways known to be false in some respects. This does not detract from their utility as means for prediction or simplification of the subject matter, but it does mean that we should be wary about making metaphysical claims about the nature of the target on the basis of them. There is no reason to think that models in neuroscience work any differently.

Another issue, noted already, is that the literal interpretation implies the multiple realizability of computations underlying intelligence, and therefore it comes with the expectation that there should be multiple realization as an empirical fact. Polger and Shapiro (2016) present a thorough case that the evidence for multiple realization is lacking, contrary to the expectations of many philosophers of mind. Of course others have a different opinion, and it is not obvious that the challenges are insurmountable (Aizawa 2018). I would not claim that formal realism is untenable just because of the empirical case that has been made against multiple realization. However, the fact that this challenge exists does provide motivation for the development of an alternative that does not need to meet this demand.

### 4.4.2 Beyond the Analogy

> The negative importance of machines, however, is that they tempt us to oversimplification. The notion of functional organization became clear to us through systems with a very restricted, very specific functional organization. So the temptation

> is present to assume that we must have that restricted and specific kind of functional organization.
>
> —Hilary Putnam (1973/1997, 97)

According to the literal interpretation, the features of the brain that are essential to cognition are such that they can be captured in a computational model and realized in a nonliving machine. This viewpoint does recognize that there are countless differences between neural and artificial substrates (the "hardware" of these systems, so to speak), but it denies that these are truly relevant to the explanation of cognition. Yet, given the number of central, unresolved questions in neuroscience concerning the relationship between brain, mind, and behavior, it is much too soon to exclude so many factors from consideration. It is a selling point of the analogical interpretation that while it sees the value of the simplifications afforded by the computational theory, it does not rule out a priori that researchers may need to look beyond computational analogies to answer these central questions. This strikes a balance between accepting the pragmatic necessity of abstraction and retaining awareness that every simplifying framework comes with inherent limitations. Next, I will mention some of the most salient differences between brains and computers and why they should not be neglected in neuroscientific and philosophical investigations of cognition.

The neurophysiologist E. A. Adrian (1954) once quipped that "what we can learn from the machines is how our brains must differ from them."[27] The most glaring difference is that the brain is an organ in a living body, made of metabolizing cells, whereas the computer is an inorganic machine. Various important disanalogies stem from this point. Sprevak (2021 draft, 2) points out that compared with brains, computers such as electronic PC's have a quite simple internal physical structure. The relevant notions of simplicity here concern the homogeneity of parts and their arrangement, and the fact that in the computer, it is straightforward to identify the processes implementing a given computation and to set them apart from background physical activity. The brain is the extreme opposite, in terms of heterogeneity of parts and organization, even within one individual, as well as the difficulty of identifying the cognitive processes against background biological processes without the imposition of a simplifying schema. Another point

27. This was quoted approvingly by Canguilhem (1963, 516).

is that the brain, unlike the computer, lacks an unequivocal hierarchical structure. As neuroanatomists have noted, there are multiple ways to chart the hierarchy of feedforward and feedback connections within the brain, given that the "wiring" is so complicated (Hilgetag and Goulas 2020). A simplification is made available by treating the brain as if it were a parallel computer, with determinate, hierarchical processing streams, even though it departs from this picture in many respects.

Digital computers are physical systems whose operation depends on there being an invariance in their functioning, across a range of different physical states. For instance, fluctuations in voltage levels get ignored in digital signaling and are classified as either 0 or 1. More dramatically, two machines can be extremely different physically, but identical with regard to what they compute. This means that the workings of computers afford a separation into distinct levels of analysis. Most obviously, they can be described at the level or hardware or software. Neuroscientists often suppose the brain is like that (e.g., Marr 1982, Carandini 2012), but this is more of a simplifying scheme than a discovery of neurophysiology. There is an argument originating with Herbert Simon (1962), that for complex biological systems to be robust and evolvable, they must be roughly modular and separable into levels, just as artificial devices are. Similarly, Ballard (2015) argues that for brains to function as control systems for the body, they must have these computer-like characteristics, where description of their operations separates cleanly into different levels of abstraction. However, these arguments do no more than show that there must be at least some tolerance for differences in the fine-grained states of neural components, and some modularity (i.e., anatomical separation of function), but not that the brain must have these architectural and functional characteristics of computers.[28] Crucially, these considerations do not get around the basic fact pointed out by Sprevak—namely, that brains, as physical systems, are a whole lot more complicated than computers.

One significant point of difference is that the hardware of electronic computers is engineered *not* to undergo material changes with use, whereas there is an inherent tendency for biological cells, whose material constitution is changing as they metabolize, to undergo use-based plasticity (Chirimuuta 2017a; Godfrey-Smith 2016a). Thus, it should not surprise us that

28. This issue is discussed in more detail in Chirimuuta (2022a).

the plasticity shown by the brain, with ordinary development and deliberate learning, is very much unlike what is seen in computational machines, even in artificial neural networks designed to simulate synaptic plasticity. Another point is that functional characteristics are far more apparent and clear-cut in artificial systems than in living systems because they have been designed by humans with specific functions in mind. What this means is that the usefulness of engineering analogs for understanding the "principles of neural design" (Sterling and Laughlin 2015) is tempered by the way that they impose an engineer's template, in which structure-function relationships are fixed and transparent, and this may mask important considerations of variability and multiplicity of function.

One consequence of the literal interpretation of the computational framework in neuroscience is that the explanatory focus on commonalities between computers and brains closes down the possibility of understanding how the specifically biological properties of brains are fundamental to cognition. As Godfrey-Smith puts it, philosophers of mind seek to understand how mind arises from matter, and one important option to keep open is that life is the bridge between matter and mind—that all living systems have protocognitive properties—characteristics whose explanations could be steps toward understanding cognition in animals with nervous systems. But this is not a viable option if it is assumed at the outset that in principle, a nonliving computer could have all the cognitive capacities of an animal. Godfrey-Smith therefore invites us to attend to the differences:

> Part of the message . . . is the enormous functional difference between a living system and this AI system, despite any coarse-grained cognitive similarity. This difference can be hard to keep in focus because the AI system, imagined or real, has been designed as a non-living analogue of a living system. It's only a partial analogue, though; it has a combination of no metabolism but a lot of information-processing. In the living system, the information-processing side of its activity is integrated with the metabolic side, so the two can only share coarse-grained functional properties. (Godfrey-Smith 2016a, 502)

Therefore, he argues that computational models are far more limited in their potential to explain cognition in living systems than is normally assumed.

One further observation is that prevalence of the literal interpretation may well be due to motivations that are extrinsic to consideration of the models and target systems themselves. Canguilhem (1963, 514–515) observes that

in biology, but not in physics, researchers tend to overinterpret analogical models. His point is that in physics, the analogical use of quantitative models does not invite researchers to project the ontology of the analog source onto the analog target, a caution that is often lacking when such models are used in biology. The difference is that the use of an inorganic system as the analog source for an organic target carries with it a promise of a reduction of the organic to the inorganic—that is, making sense of the organic in purely physical terms—which is why the literal interpretations are so alluring. Canguilhem goes on to say that cybernetic models are a good example of this tendency, especially when the models' actions (e.g., in a robot) tend to simulate or mimic natural behavior.

To conclude, formal realism offers the promise that it is possible to devise quantitative, formal, and perspicacious models that faithfully mirror the processes in the nervous system that underlie cognition. This invites people to interpret neurocomputational models literally, and when this interpretation holds sway, there is a tendency to downplay the disanalogies between brains and silicon computers (even if the official doctrine is that the brain is *not* like your PC), and moreover to keep the processes relegated to mere metabolic support on the sidelines of theoretical neuroscience. It remains to be seen whether the mysteries of biological cognition will open up to an approach that treats organic intelligence as sui generis, not sharing essential commonalities with the functions of computing machines. But the replacement of formal realism with an approach that pays attention to the various modes of analogy and disanalogy between brains and computers will at least help us avoid any false directions indicated by overreaching, literal interpretations.

# 5 Ideal Patterns and "Simple" Cells

> The scientist, lacking the power to reign over a part of the world as master and possessor, constructs an image of it that is simpler but at the same time as faithful to it as possible, acquiring in this way a mastery over the image . . . if not over the world.
>
> —Jean-Pierre Dupuy (2009, 138)

The quotation marks in this chapter title are because the cells in question are simple only under a certain manner of investigation. Chapter 3 was about simplification introduced through experimental design. The simple reflex was a creature of laboratory induced dissociations between parts of the nervous system; it was not an original, latent element as the reflex theorists had sometimes supposed. Chapter 4 was about the simplifications afforded by modeling the brain as a kind of computer. We will now examine how these two strategies come together. In a study of early research on primary visual cortex, we will see how the simplifications of experimentation and data analysis lead to the production of the ideal patterns, the phenomena that are the target of computer modeling. The argument of section 5.1 will be that the simple cell—one of the major classes of neurons introduced sixty years ago in the standard account of primary visual cortex—was in some sense a creation, something artificial, but that this did not undermine its theoretical and technological usefulness. However, there were legitimate concerns raised by neuroscientists because of its nonnatural status.[1]

1. Of course, the cell is not "nonnatural" in the same sense that a human-made object, like a computer, is. In chapter 4, the important distinction was between an organ of a living body (the brain) and a purpose-built tool (the computer). Here, we

Section 5.2 will show how the turn away from the methods of classic visual physiology were in part prompted by these worries. Within the last decade, there has been a shift toward big data, ethological methods, but an irony of the story, as I will relate, is that artifice has not gone away but only changed its role.

## 5.1 The Creation of the Simple Cell

> All good experiments are good abstractions.
> —Arturo Rosenblueth and Norbert Wiener (1945, 316)

An essential, though easily discounted feature of most laboratories is that there are walls and a roof, built not only to keep the researchers dry and warm, but to create a controlled environment for the experimental occurrences themselves, shielding them from the instability of the weather and, where required, from other influences such as electromagnetic radiation, which could introduce unknown complications to the processes under investigation. As Cartwright (1983, 1999) has argued, the successful demonstration and explanation of physical phenomena in the laboratory are no warrant to think that those same phenomena can occur elsewhere.[2] The phenomenon of the simple cell is in its own way dependent on shielding, although not against rain and radiation, but instead from interfering factors generated elsewhere in the brain due to the inherent tendency of the cortex to shape itself to the complexities of its environment.

Before setting out this line of argument, some preliminary information about visual physiology needs to be provided. The *receptive field* (RF) of a visually responsive neuron represents how the cell responds to patterns of light present in the visual field. It describes the kind of stimulus that activates a neuron—a moving bar, a red patch, or whatever, at a particular location in space. The term was introduced by Sherrington in the context of the reflex theory, where the receptive field of the scratch reflex was the area

attend instead to the difference between a brain area in its "natural" state outside the laboratory, as opposed to its condition when affected by experimental manipulations.
2. See also Sullivan's (2009) foundational work on experimental neuroscience, which highlights the problem of translating experimental constructs from one laboratory to another.

of skin on the dog's back, within which a stimulus could trigger the reflexive movement of the hind leg, at least in the "spinal dog" (Sherrington 1906b).[3] It was later used to characterize the responses of optic nerve fibers (Hartline 1938) and retinal ganglion cells of the frog (Barlow 1953).

Primary visual cortex (V1)[4] is the region that receives input from visually responsive cells in the lateral geniculate nucleus (LGN) of the thalamus, which in turn have received input from retinal ganglion cells (RGCs). The iconic studies of Hubel and Wiesel (1962, 1968) mapped the receptive fields of neurons in this region of the cat and monkey brain, respectively. Whereas RGC's and LGN cells had been found to have circular receptive fields, in which light falling in a small central "ON" region excites the neuron, and light in a concentric surrounding "OFF" field inhibits the neuron,[5] the receptive fields of V1 cells were elongated, consisting of parallel bar-shaped areas of ON and OFF regions (see figure 5.1a).[6] This suggested to Hubel and Wiesel that certain VI neurons acquired their response characteristics by receiving input from a series of LGN neurons whose circular receptive fields were spatially aligned (see figure 5.1b).

Such neurons were termed *simple cells*, and their receptive fields were noted to have spatially segregated and spatially antagonistic ON and OFF subregions. They also showed summation within each region, which is to say that there was a cumulative effect of increasing the number of spots of light falling within the subregion, with more spots in the ON region leading to a larger response. It was emphasized that the simple cell's response to any stimulus could be predicted from its arrangement of ON and OFF

3. Note that this is the RF of a *reflex arc*, not an individual neuron. Sherrington's introduction of this term leads Yuste (2015, 487) to state (erroneously) that Sherrington (1906) treated the single neuron as the basic functional unit (namely, the "integrative unit") of the nervous system.

4. This is also known as *striate cortex* or *Brodmann's area 17*. Strictly speaking, the term "V1" should be reserved for primates, but I will use it more widely than that.

5. And the opposite is the case for OFF-center cells. This juxtaposition of excitatory and inhibitory subregions in the RF is known as *spatial antagonism*.

6. See Hubel and Wiesel (1998) for an overview of their work, including the "accidental" discovery of elongated RFs; also, their (1977) Ferrier Lecture includes a summary of findings relating to functional architecture not discussed in this chapter, such as ocular dominance columns. The concepts of ocular dominance and orientation columns certainly relate to Hubel and Wiesel's sense of "order" (i.e., simplicity) in the visual cortex (1977, 55).

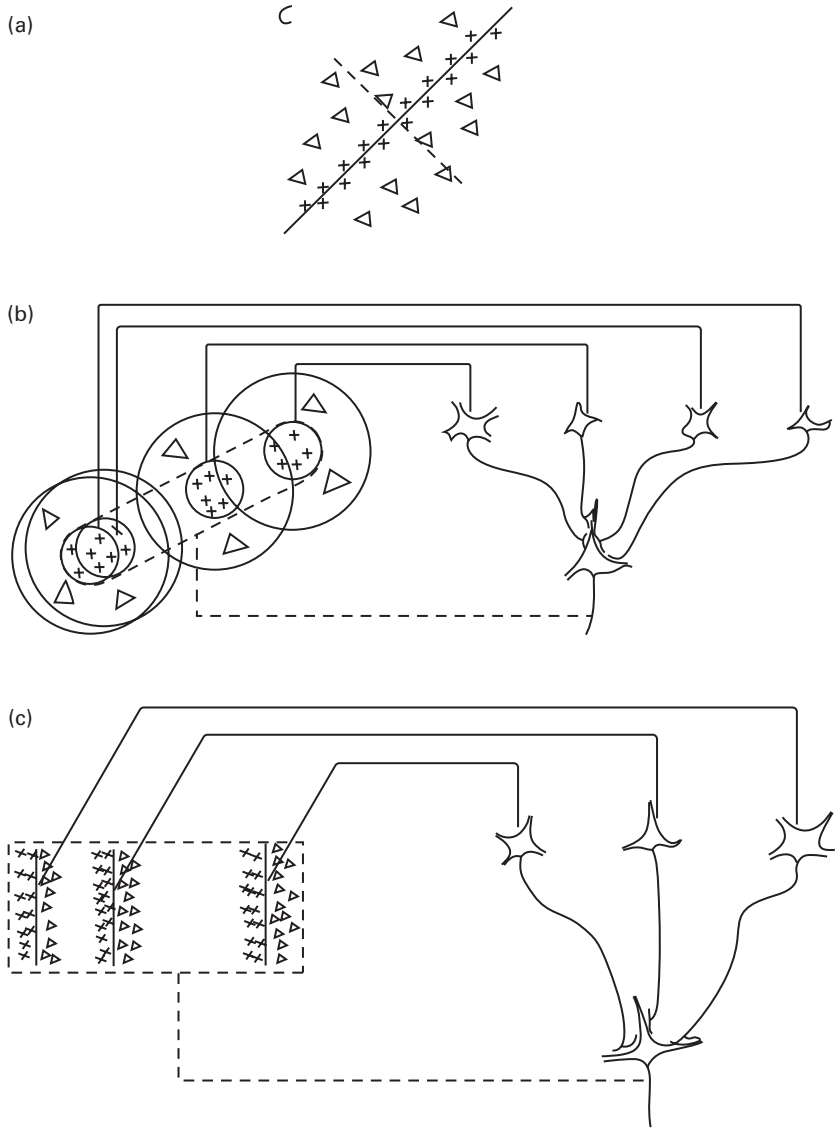


**Figure 5.1**
(a) An "ON"-center receptive field of a simple cell. Here, the *x*'s represent the area in the visual field that gives excitatory responses when light is flashed there, and the triangles represent areas that produce inhibition. From Hubel and Wiesel (1962, figure 2). (b) Schema for explaining simple cell RFs. Four LGN neurons with circular-surround RFs (top right) give input to one simple cell (bottom right), giving an elongated RF (left). From Hubel and Wiesel (1962, figure 19). (c) Schema for explaining complex cell RFs. Three simple cells with phase-sensitive, elongated RFs (top right) give input to one complex cell (bottom right), giving a phase-insensitive elongated RF (left). Hubel and Wiesel (1962, fig 20).

subregions.[7] The cells termed *complex* failed to show at least one of these characteristics. Complex cells were not sensitive to the exact placement of a bar within the receptive field, and this suggested to Hubel and Wiesel that their response profiles were due to their receiving inputs from a series of simple cells whose elongated receptive fields formed a small array (see figure 5.1c).[8] Thus we have a feedforward, hierarchical picture of the visual system in which the response profile of cells at each stage is assumed to be determined by the input received from neurons lower in the hierarchy.[9]

Given that the characterizations of Hubel and Wiesel were exclusively qualitative, other research groups developed quantitative approaches to receptive field mapping and to the simple/complex distinction. Later work showed that the various characteristics definitive of simple cells were all manifestations of an essential *linearity* of spatial summation (Movshon, Thompson, and Tolhurst 1978): the simple cell just adds up the total of light falling in its ON region and subtracts from it the total of light falling in its OFF region, such that its firing rate gives a running reflection of this calculation.[10] Linear systems have appealingly simple characteristics, in particular that they are amenable to reductive analysis. Knowledge of the behavior of a linear system can be achieved through piecemeal examination of its parts, or as we should say here, its partial responses to pared down stimuli such as spots of lights, small white bars, sinusoidal gratings, or Gabor patches (see

7. See Mechler and Ringach (2002, 1017) on these characteristics.

8. Hubel and Wiesel (1968) also used the classification of *hypercomplex* to refer to cells showing "end stopping"—inhibition when a stimulus is placed at the ends of the receptive fields. For ease of exposition, I will restrict my discussion to simple and complex cells.

9. It is worth mentioning Hubel and Wiesel's anatomical evidence for the hierarchy—the finding that simple cells were more likely to be found in layer 4 of V1, the layer that receives subcortical input. Also, Hubel and Wiesel present their scheme with some diffidence: "proposals such as those of Text-figs. 19 and 20 are obviously tentative and should not be interpreted literally" (1962, 144); the hierarchy is "possibly over-simplified" (1968, 217).

10. Movshon et al. (1978), and subsequent researchers also posit that simple cells have an output nonlinearity, *rectification*, whereby negative values of summation due to inhibition are converted to a response of zero spikes. This is necessary because V1 cells have a low baseline firing rate. Thus the standard model of the simple cell is known as the *linear-nonlinear* (LN) model. See Carandini et al. (2005, figure 1) and Butts (2019, figure 3) on the LN model.

figure 1.1 in chapter 1 for examples).[11] Hence the pressure to examine the system as a whole, or in more complicated circumstances, is alleviated—or so the story would go.

### 5.1.1 The Experiment-to-Model Pipeline

I will now explain how it is that the simple cell—a V1 neuron whose activity reflects an essentially linear sum of light falling within its receptive field—is a creation of the laboratory. Masland and Martin (2007, R581) write that "experimental physiologists know all too well that sensory systems are only linear when the experimenter forces them to be so." This might be common knowledge to experimentalists, but it demands some explication. The notion of "forcing" linearity amounts to the selection of experimental conditions that induce more simple behavior in sensory systems than would occur otherwise. The two important simplifying strategies here are *restraining behavior* and *constraining stimulation*.

The experiments of Hubel and Wiesel (1962, 1968) and Movshon et al. (1978) were performed on anesthetized animals. Behavior, one factor that modulates responses in visual cortex, was eliminated. There would not be trial-to-trial variability due to behavioral and attentional shifts, as there would be in an awake, behaving animal. The result was a more regular data set.[12] Hubel and Wiesel (1962, 123) justify this experimental choice by referring to an earlier study by Hubel (1959) on awake, loosely harnessed cats, which reported no qualitative difference in neural activity between those and anesthetized animals. Still, general anesthesia does cause a general suppression of spontaneous activity in the cortex, which would be a matter of concern for future physiologists, as it was for Kuffler (1953). Regarding stimulus constraints, the fact that use of simple artificial stimuli, like spots and bars of light, leads to simpler (i.e., more linear) neuronal responses is due to the inherent plasticity of the cortex. Neuronal responses tend to adapt themselves to the statistics of the prevailing stimulus regime and will show a more complex response profile when the stimuli themselves are more

11. See De Valois and De Valois (1988) on the application of linear systems theory (Fourier analysis) to vision science.

12. This is not to imply that trial-to-trial variability is not still very large, even in anesthetized animals. See Arieli et al. (1996) for the case that ongoing network dynamics account for this variability.

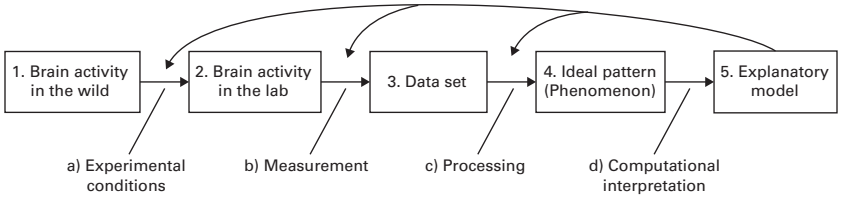


**Figure 5.2**
The experiment-to-model pipeline. This is the schematic path from an experiment measuring neural activity during a specific task or stimulus regime to a model that interprets the neural activity as carrying out certain computations, offering an explanation of the brain's involvement in cognitive performance. The computational models are themselves abstract and idealized representations of neural activity, but they also depend on simplifications introduced early in the pipeline. As Butts (2019, 469) writes, "The use of simple stimuli leads to simple models of neural computation."

complicated, as with images encountered in the environment outside the lab (David, Vinje, and Gallant 2004; Butts 2019, 454). Even more important, the use of stimuli, like oriented bars, that generate a peak response in only one population of neurons selective for that particular width and orientation means that the modulatory effect of neurons selective for different patterns (which would be activated by a complex image, containing a range of features), is suppressed.

I am not claiming that the constraints on behavior and stimulation were consciously employed by Hubel, Wiesel, and other physiologists of this era in order to make neural activity more simple. It happens that just by following the ideal of a well-controlled experiment, the investigator will be led to these choices. The aim is to eliminate factors that are both unobserved and uncontrolled. The use of anesthesia and artificial stimuli are two ways to make a better-controlled neurophysiological experiment. But because of the adaptiveness of the brain, the tendency for its activity patterns to shift with different behavioral and environmental contexts, these practices will have the effect of inducing a more regular and more linear response profile than would occur without them.

Figure 5.2 depicts what I call the *experiment-to-model pipeline*.[13] By "brain activity in the wild" (1), I just mean the neural activity associated with

13. My discussion of it is limited to the example of classic V1 physiology, although this template should be applicable more widely. Indeed, the postclassical research

cognitive performances, such as vision, that an animal undertakes spontaneously or in response to nonexperimental situations. This is contrasted with "brain activity in the lab" (2), which is the neural activity associated with cognitive performances occurring under experimentally controlled conditions. Simplifications again occur at the stage of measurement (figure 5.2b). The extracellular detection of action potentials of single neurons was the recording technology available at the time of Hubel and Wiesel's studies.[14] With single unit, as opposed to population measurement, the physiologist is only examining one tiny component of the visual system at a time. While the practice of single-unit measurement was imposed by the technology available in the mid-twentieth century, it does in any case have a simplifying effect. This mode of recording does not allow observation of the interaction between the neuron and any other parts of the brain, so it forces the physiologist to treat all recorded activity either as stimulus-driven or noise. In addition, single-unit physiology involves cell selection. An article from this era would report at most the result of recording from hundreds of neurons out of the population of many thousands, if not millions.[15] The experimenter would shift the position of the electrode, hunting for the cells that make large, visually driven responses. Thus, the sampling from the population was not at all random. Masland and Martin (2007) argue that physiologists have a tendency to record from neurons that can be classified as a certain homogeneous type of standard cells, like simple and complex ones. The outcome of measurement is a data set that abstracts away from the heterogeneity of response profiles and receptive field types in the original cortical population.

Various techniques fall under the stage of "processing" (figure 5.2c). Responses of any one neuron, to a given stimulus, would be averaged across trials; data would be analyzed with the aim of making a principled classification of cell type. Olshausen and Field (2006, 190) highlight the significance of this stage, writing that "the way in which response properties are characterized can have a profound effect on the resulting theoretical framework that

discussed in section 5.2 would be another good example of the pipeline. Motor cortex presents other examples (see chapter 7).

14. Note here the invention of the tungsten microelectrode by Hubel (1957), discussed by Bickle (2022).

15. There are estimated to be about 140 million neurons in V1 of an adult human (Leuba and Kraftsik 1994).

is adopted to explain the results." I will mention one important classification tool: the $F_1/F_0$ ratio.[16] When V1 neurons are stimulated with a sinusoidal grating that drifts across their receptive fields, some of them will give a response that is strongly modulated by the alternation of light and dark bars, whereas others show an activation that is steadier during the stimulation window (i.e., phase invariant). The $F_1/F_0$ ratio measures whether the cell is modulated or more phase invariant, and it was widely used to classify cells as linear or nonlinear, and hence simple or complex (Mechler and Ringach 2002, 1017). Use of this ratio effectively sharpens the category boundary between simple and complex cells, creating a distribution of two discrete populations out of an original population in which response properties form a blurred continuum.

We now reach stage (4), in which we have a phenomenon, an *ideal pattern*. The philosophical dimensions of this notion will be provided in the next section. Here, I will just point out that it is at this stage that the simple cell makes its appearance. Certainly, there were cells in primary visual cortex already before the experiment, but the particular profile definitive of the simple cell did not predate the operations of fixing experimental conditions, taking measurements, and processing the data. The ideal pattern is what is represented in the receptive field map of the simple cell (see figure 5.1a), which is itself a summary representation of neuronal responses to visual stimuli elicited in the course of an experiment. The ideal pattern or phenomenon can be the target of an explanatory model. When the phenomenon is given a computational interpretation—when it is hypothesized to perform a certain information-processing task in the economy of the brain—it is described mathematically as taking inputs from elsewhere in the brain and computing some output, a signal to be read out in a downstream area. As argued in chapter 4, modeling introduces another mode of simplification—abstraction away from the organic underpinnings of the brain's operations. For visual neurons, the explanatory model takes the form of an *encoding model*, a function over visual stimuli that makes a prediction of how the neuron will respond to any arbitrary configuration of light falling within its receptive field. The theoretical content of visual neuroscience is centered around such models (see section 5.1.3). I have included an arrow leading back from the model to stages

16. This is calculated from the Fourier components of the cell's average response to a drifting grating. It is the ratio of the first harmonic, $F_1$, over the mean spike rate, $F_0$ (Movshon et al. 1978, 59).

(a)–(c). This is because explanatory models and theoretical ideas shape the processes through which the ideal pattern is formed.[17] Over the years, this sequence is iterated and refined many times.

### 5.1.2 Scientific Phenomena as Ideal Patterns

The initial outcome of the experiment is a data set generated through the measurement of neural activity. Since Bogen and Woodward (1988), it has been commonplace to say that phenomena, not raw data, are the target of explanatory models.[18] Phenomena are the simplified patterns created by processes such as averaging and curve fitting that can potentially be explained by a theory, unlike the messy patterns of raw data. In their intuitive example of the melting point of lead, Bogen and Woodward describe how a series of thermometer measurements would show some variance, noise due to uncontrolled factors in the laboratory. As every high school experimentalist has been taught, the thing to do is take the mean of the raw data points. That value, Bogen and Woodward argue, is *the* melting point of lead—the phenomenon to be explained. By their account, such phenomena are regularities that exist in nature independently of experimental controls and statistical processing, though they are somewhat masked by the uncontrolled occurrences that inevitably occur and lead to noise in the data.

---

17. To give an example, we see in this passage from Hubel and Wiesel (1962, 145) that the theoretical proposal of a feedforward hierarchy settles concerns about the trial-and-error method that they used to search for effective stimuli for complex cells:

> The method of stimulating the retina with small circular spots of light and recording from single visual cells has been a useful one in studies of the cat's visual system. In the pathway from retina to cortex the excitatory and inhibitory areas mapped out by this means have been sufficient to account for responses to both stationary and moving patterns. Only when one reaches cortical cells with complex fields does the method fail. For these fields cannot generally be separated into excitatory and inhibitory regions. Instead of the direct small-spot method, one must resort to a trial-and-error system, and attempt to describe each cell in terms of the stimuli that most effectively influence firing. Here there is a risk of over- or under-estimating the complexity of the most effective stimuli, with corresponding lack of precision in the functional description of the cell. For this reason it is encouraging to find that the properties of complex fields can be interpreted by the simple supposition that they receive projections from simple-field cells, a supposition made more likely by the anatomical findings of Part III.

18. Phenomena are often equivalent to data models (McAllister 2007), but I do not restrict my use of the term to data models. See also Feest (2011) and Colaço (2020) on the characterization of phenomena in psychology and neuroscience.

My proposal is that neurophenomena must not be thought of as "real patterns," regularities that exist as much in the wild as in the laboratory. Instead, neurophenomena are *ideal patterns*, the regularized products of the series of simplifying procedures of the sort outlined in this discussion (and see section 2.1.1 in chapter 2). The intended contrast is with Potochnik (2017, chapter 2), who proposes that "real causal patterns" are the representational target of scientific models. With Potochnik's use of the notion—which is a departure from Dennett's—these patterns have a reality that does *not* depend on there being an agent capable of discerning them (Potochnik 2017, 28).[19] According to Potochnik, observed phenomena comprise real causal patterns mixed with "noise" (which may be other causal patterns, not currently of interest to the researcher). The job of idealized models is to separate the wheat from the chaff, making the real pattern of interest more salient, comprehensible and useful. The notion of an *ideal pattern* is actually closer to Dennett's "semirealist" intent than Potochnik's "real causal pattern," but I name it "ideal" to mark its distinctness from the commonly used sense of "real" as "just out there in nature" and "independent of the scientist." Ideal patterns do not arise from the superposition of a perfect regularity with noise or with interfering regularities; instead, they are the product of scientists working on a material system to regularize its behavior and then enhancing this regularity through data processing.[20] This approach fits with the *haptic realism* introduced in section 2.1 of chapter 2, whereby scientific knowledge is conceived as the product of the interaction between scientists and their target of investigation. Here, we see how this works in

19. The term "real pattern," of course, is famous from Dennett (1991), who used it in a less robustly real sense than Potochnik. I argue elsewhere that the notion of "ideal patterns" is required to account for the nonfactive understanding provided by simplified models such as the LN model (Chirimuuta 2023a). That paper contains further discussion of ideal patterns and how they relate to scientific understanding. See McAllister (1997) and Massimi (2011) for additional arguments that phenomena be treated as dependent on scientific activity, to some extent.

20. The second stage is crucial. If these patterns were only causally dependent on the scientists' activity, they would be as concretely real as any material artifact. But in addition, they come about through the scientists' processing and interpreting the data derived from the controlled material system in particular ways that are underdetermined by the system itself. This justifies the label "ideal pattern"—these patterns are not only causally, but also constitutively dependent on the scientists.

practice through the production of knowable objects: the simple cell as an ideal pattern.[21]

The motivation for this shift is clear when we consider how neurophenomena are different from the straightforward example of the melting point of lead. In the observation of complex, biological systems, there is no clear-cut difference between signal and noise. There are innumerable processes in play, some of more interest to the scientist than others. The partitioning of signal and noise is therefore driven by theoretical expectations and by pragmatics. In electrophysiology, some trial-to-trial variability is certainly due to experimentally introduced noise (i.e., instrument noise); but much is due to the endogenous variability of the system, such as behavioral and attentional context (Stringer et al. 2019, Musall, Kaufman, et al. 2019), and such modulations occur even in anesthetized animals (Arieli et al. 1996). The important point is that this variability is *not* all "noise" from the point of view of the brain. Unlike in the case of melting lead, which can intuitively be depicted as a melting phenomenon separated in nature from background occurrences,[22] the specification in neurophysiology of what is the phenomenon—the process of significance standing out against a background of irrelevances—is always a matter of the researcher's discernment. Out of countless, mostly uncharted occurrences, a neurophenomenon is honed and sculpted through the tools of experimental practice and data analysis. It should not be assumed to exist independently of those methods.

---

21. I reiterate that haptic realism is not supposed to be an intermediate position between standard scientific realism and antirealism (aka "empiricism" or "instrumentalism") (section 2.1.3 in chapter 2). Instead, it stands out against both kinds of views by emphasizing the interactive processes that bring about scientific knowledge, explanation, and understanding. Against standard realism, this means that the imprint of the human scientist in shaping knowledge can never be discounted, as is required to bolster claims that the best science is an unfiltered representation of human-independent nature. Against standard antirealism, science can still be taken to offer representations of portions of nature (that go beyond data summaries), and which are constrained by the behavior of the target system. We must also keep in mind the message of chapter 1—namely, that each portion of nature is vastly more complicated than will be possible to depict in any scientific representation, hence the partiality and limitations of these representations.

22. I do not mean to affirm that some version of these issues—to do with the problem of isolating a "clean" phenomenon against background factors—do not arise in the investigation of physical phase transitions (see Chang 2004), but just that it is relatively easy to *think* that this is not the case, as per Bogen and Woodward (1988).

To bring focus to the sense in which the simple cell is an ideal pattern, we can consider a question posed by Olshausen and Field (2006, 190), who ask, "Are these categories [simple and complex] real or a result of the way neurons were stimulated and the data analyzed?" As I see it, there is a false dichotomy in their holding that *either* the categories are real (preexisting in the wild) *or* not real and merely the result of experimental practice, in essence an artifact. My account enables us to see that indeed, the simple cell is a result of certain procedures, but it is not merely an artifact. The ideal pattern is not fictional. The simple cell is a creation, but not ex nihilo. The categories of simple and complex cells are suggested by things that neurons actually do (in the lab), with certain characteristics of their response profiles highlighted at the processing stage. More generally, we should think of ideal patterns as taking shape through a process of interaction between scientist, lab equipment, target system, data, and modeling. It must be appreciated that the simple cell was not the result of the scientists projecting their categories onto the inert substance of the brain, but rather the outcome of a certain agency on the part of the visual cortex, the neurons' tendency to adapt themselves to their context.[23] Without this adaptiveness, the simple—or rather, *simplified*—cell would not come about, for experimental constraints would not so readily lead to a reduction in complexity of neuronal activity.[24]

The regularization that comes with the generation of ideal patterns can be seen as foundational to the scientific study of objects that do not present themselves as already neatly packaged and classified. Lorraine Daston (2016) writes about the cloud atlases of nineteenth-century meteorology, in which the shape-shifting structures of clouds, whose forms blur continuously from one kind to another, were codified into the paradigmatic types of *cumulus*,

23. This point is comparable with some ideas from actor network theory. There is a focus on experimental interaction between scientist and target system, with agency acknowledged on both sides (Pickering 1995). Like Latour (1992), I seek to avoid both a pure constructivism (whereby the simple cell would be a fiction generated by the scientist) and a robust realism, whereby the simple cell would be an object that exists independently of scientific practice, with the aim of science being to generate accurate representations of it. See Nordmann (2006, 17–18) on the thread between Kantian epistemology and Latour's account of experimentation.

24. The idea that simple experimental conditions lead to simplified patterns of neural activity is a point supported by the recent theory of *neural task complexity* (Gao and Ganguli 2015), to be discussed at the end of this chapter.

*nimbus*, and so on, through "description by omission." The nimbus cloud is an ideal pattern, with the additional characteristic of the simple cell being that it is not only observed under a certain manner of abstraction, but it is experimentally shaped *and* observed. In an early report on retinal physiology, Kuffler (1953, 61) had emphasized the "flexibility and fluidity of the discharge patterns," and how the heterogeneity of responses was such that the cells defied imposition of a classification scheme.[25] Hubel and Wiesel did not have such compunctions. By 1977, they had arrived at a neatly parameterized picture of V1 neurons, where in an analogy with manufactured items—ready-to-wear suits that are identical except for differences along prespecified lines—cells are said to be fully describable in terms of their position along five axes of variation:

> We may think of a cell in area 17 of the monkey as responding optimally when a number of stimulus variables are correctly specified. The cell may be described, then, by its degree of complexity,[26] the x-y coordinates of the position of the receptive field, the receptive field orientation, the ocular dominance, and the degree to which there is directional preference to movement. Such a list of specifications is analogous to the tag showing the price, sleeve length, percentage of wool, and so on, attached to a suit in a department store. (Hubel and Wiesel 1977, 11–12)

By Kuffler's account, which did not abstract away from the heterogeneity of the population, each single neuron would be a bespoke outfit, unique and sui generis, though sharing some common features of general organization, such as center-surround antagonism.

### 5.1.3 Theoretical and Technological Sequalae

We will now examine the theoretical significance of the simple cell, considered as an ideal pattern. From the 1970s onward, a standard model of primary visual cortex evolved, and the conception of the simple cell as an essentially linear filter in a feedforward processing hierarchy was central to it (Carandini et al. 2005). The LN encoding model of simple cells (see footnote 10), in conjunction with the energy model of complex cells (Adelson and Bergen 1985) dominated the field. So much has been written about primary visual cortex

25. Kuffler (1953, 62) writes of the cat retinal ganglion cell that "there seems to exist a very great variability between individual receptive fields and therefore a detailed classification cannot be made at present."

26. Namely, whether it is simple, complex, or hypercomplex.

that I could not hope to cover a range of theoretical perspectives in any detail. Instead, I have selected one article by Horace Barlow as representative of the way that experimental and modeling work on the visual cortex fed into a theory of the neural basis of visual perception.[27] With the simplicity of this neurophenomenon in place, a clear, intelligible account of the operation of the visual cortex was made possible. We will see that many reasonable criticisms could be leveled against the picture of primary visual cortex suggested by research on simple and complex cells, but regardless of these flaws, it led to a significant technological innovation.

Barlow's (1972) article "Single Units and Sensation: A Neuron Doctrine for Perceptual Psychology?" is a manifesto for methodological reductionism, insisting on the adequacy of knowledge of parts in isolation for the understanding the whole. Its central proposal is that the single neuron is the elementary operational unit of the visual system, not any smaller-scale structures or larger populations of neurons.[28] The elementary function of the visual neuron is, Barlow, contends, to detect a specific feature, and as Butts (2019, 465–496) notes, this notion of neurons as feature detectors is grounded in simplified linear models. Each level of the feedforward hierarchy contains neurons responsive to certain kinds of trigger features, with complexity of features increasing as we move up the levels—spots for precortical cells, spatially located bars for simple cells, and involved objects like hands for neurons in the temporal cortex areas near the top of the hierarchy. It is in the inferotemporal cortex (IT) that the infamous "grandmother cell" might appear. The operation of the whole visual system is decipherable from knowledge of what individual neurons respond to. Since Barlow takes it that cells are silent for all but their preferred stimulus, not showing signs of multifunctionality, he is not bothered by antireductionist worries about the insufficiency of the single-neuron perspective (1972, 382).[29] Thus, the methodology

27. See Movshon (2021) for a wider overview of the significance of Barlow's theoretical work, including the notion of *sparseness* also employed in the 1972 article.

28. There is, of course, a debt to the neuron doctrine of Ramon y Cajal (Shepherd 1991).

29. We should note here the connection between the supposition of unifunctionality and the archetypal notion of a mechanism, in which each component has one clearly specified role in the system. Fusi et al. (2016, 66) state this nicely: "The traditional view of brain function is that individual neurons and even whole brain areas are akin to gears in a clock. Each is thought to be highly specialized for specific functions."

of single-unit electrophysiology—the detailed study of one neuron at a time, without regard to the influence of cortical population activity—is vindicated.

This theory was compelling to many, but as Masland and Martin (2007: R578) put it, "an aggressive slice of Occam's razor was required to make sense of the properties of cortical neurons." They are referring in particular to the way that the theory had to posit homogenous groups of standard precortical RGC and then LGN cells feeding into the simple cells, whereas nonstandard cells are the majority in many animals, including the cat. They also point out that the simplifying assumptions of linearity, and distinct cell classes in the cortex, are ill matched to the findings of more detailed anatomical and physiological examination. A yet-harsher attack on the theory came from Olshausen and Field (2006), who make the case that the so-called standard model of V1, though claiming comprehensiveness, accounts for only 15 percent of the behavior of neurons in this area.[30]

Olshausen and Field (2006, 182) conclude that "for all practical purposes, we stand today at the edge of the same dark abyss as did Hubel and Wiesel forty years ago." In saying this, they ignore some technological (i.e., practical) advances that were inspired by this research. In chapter 4, I described the research strategy of constructing artificial devices with similar functionality

30. See also Olshausen and Field (2005), which shares much of the same content as the 2006 publication, and is more widely cited. Olshauen and Field (2006) present five lines of criticism: (1) biased sample, (2) biased stimuli, (3) biased theories, (4) interdependence and contextual effects, and (5) ecological deviance. Point (1) is the bias toward recording cells that are strongly visually responsive and easy to categorize. In point (2), they argue that use of artificial stimuli (see figure 1.1 of chapter 1 for examples) is justified only if V1 *is* basically a linear system because stimulation with a small set of such stimuli would then allow a complete characterization of the system's response profile. But see Rust and Movshon (2005). Point (3) draws on the Mechler and Ringach (2002) result that the $F_1/F_0$ ratio creates a bimodal distribution from V1 data. In point (4), regarding interdependence and contextual effects, Olshausen and Field note that feedforward input from the LGN is responsible for only 35 percent of the behavior of neurons in the "input" layer 4 of V1. The other 65 percent, they argue, must be due to intracortical connections, including input from neurons most responsive to other sensory modalities. In point (5), Olshausen and Field argue that ability to predict responses to natural stimuli is the key test for encoding models. As we will see in section 5.2, this view was shared widely, and its consequences for the field were significant. Lehky and Sejnowski (1988) is a much earlier criticism of the "neuron doctrine" assumption that the function of visual neurons can be determined by examination of their RF maps.

to a neural system, using the artificial, more scrutable system to shed light on the neural one. This was, as it happened, the method that led to the first deep artificial neural network (ANN) models of vision. Here is Fukushima (1980, 193), inventor of the "neocognitron": "If we could make a neural network model which has the same capability for pattern recognition as a human being, it would give us a powerful clue to the understanding of the neural mechanism in the brain."

His feedforward, hierarchical model took inspiration from Hubel and Wiesel's work, as well as Barlow's theories. It is the ancestor of the deep convolutional neural networks (DCNNs) that rose to prominence in the last few decades.[31] The outstanding successes of DCNNs have been in machine vision tasks such as text, object, and face recognition. They contain artificial nodes performing the computations that the old theories attributed to neurons in V1 (Butts 2019, 462). Against attacks on the "standard model" of V1 from Olshausen and Field, others point to the achievements of this technology as vindication of the (strictly speaking) false assumption that the visual system is a feedforward processing hierarchy, lacking top-down (recurrent) input (Vintch, Movshon, and Simoncelli 2015, 14839).[32] This brings to the surface important questions about the value of manifestly false theories.[33]

In particular, we should consider whether the classical theory—which is deemed false because it is severely oversimplified with respect to the anatomical connections and physiological processes thought now to occur in the visual cortex—still provides any understanding of this brain area, especially as it operates in the wild. I have argued elsewhere that ideal patterns

31. See for instance, LeCun et al. (1989) and Krizhevsky, Sutskever, and Hinton (2012); for review, Buckner (2019) and Lindsay (2020).

32. In fact, it is a common view that as yet, the DCNN is the best model of the primate ventral stream, even though it contains these false assumptions (Lindsay 2020). From anatomical and physiological observation, it is well known that feedback connections from the rest of the brain to the visual cortex are significant.

33. I do not concur with the account of Wimsatt (2007, chapter 6), that false theories are a guide to truer ones, since this pulls us back toward a traditional realism that sets up complete, representational accuracy as the ultimate goal. Fruitfulness for empirical research is another recognized benefit of so-called false theories. Of the 1972 "Neuron Doctrine" article, Movshon (2021, 188) writes that "it is quintessential Horace, because even when ultimately proved wrong, his ideas provoked important experimental work that would not otherwise have been done. . . . The fact that Horace was not always correct is what made him a good—no, a great—theorist."

have a role to play in the generation of *nonfactive* understanding of complicated systems.[34] This is a kind of scientific understanding that does not rest on the acquisition of truths about the target system, in all its complexity, but instead on the presentation of a theory or model that is more of a caricature, exaggerating some features to the point of misrepresentation and completely ignoring others. Such theories and models provide understanding because they strike a compromise between the overwhelming complexity of their target and the cognitive capacities of the investigator, limited in their ability to make sense of too many intricacies. Nonfactivists tend to emphasize the pragmatic dimension of this notion. A simplified model will not tell the whole story about V1, but it could be "true enough" (Elgin 2017) for certain tasks. Further, de Regt (2017) insists that scientific understanding enables people to *do* certain things. The caricatured depiction of V1 as containing linear simple cells feeding forward into translation-invariant complex cells enabled people to build artificial visual systems. I will argue in chapter 9 that there is a danger in making too much of this technological feat—the differences between artificial and organic visual systems are stark. Still, this achievement was possible because of the nonfactive understanding afforded by the simplified picture of V1.

At the end of chapter 3, I argued that the reflex theory was ultimately unsuccessful, in that it failed to achieve its stated goal of behavioral engineering. We saw that its lack of technological success was due to its ecological invalidity. The record of the classical theory of V1 is more mixed. Like the reflex theory, it is highly simplified, perhaps simplistic, but it does have a significant spin-off to its credit, the DCNN. In common with the reflex theory, it could not meet the challenge of ecological validity, and its failure to predict responses in naturalistic viewing conditions was a major reason why neuroscientists began to explore methods beyond the classical approach.

Before moving to the next era of research, I would like to point out that my interpretation of the classic era stands against a common view of simplifying methods in experimentation and modeling, one that asserts that they facilitate production of effects outside the lab because they provide knowledge of stable "capacities" operating both in and out of controlled conditions (Cartwright 2009, discussed in Chakravartty 2017, 117). The key issue here is that

34. See Chirimuuta (2020c, 2023). Nonfactivist theories of scientific understanding in recent years have been defended by Elgin (2017), de Regt (2017), and Potochnik (2017).

there are no grounds for asserting that the classical simple cell experiments revealed much about invariant capacities or causal relationships, given that the response profile changes so much under different circumstances (see section 5.2). At most, one could simply posit that the cell has an invariant disposition, which is partially manifest in the classic experiments. But that would be an empty proposition, given that the partial manifestation does not reveal enough to be a basis for robust generalization to new conditions. Northcott (2022) makes a helpful distinction between "master-model" and "contextual" strategies, where the former but not the latter strategy presupposes the stability of the relationships being modeled. Master-model strategies rely on the assumption that nature consists of stable relationships, perhaps remaining hidden behind noise, whereas the contextual approach countenances the fragility and instability of the relations targeted in the model. My position is that a certain instability needs to be accepted regarding neuronal behavior—hidden invariance is an unwarranted posit.

## 5.2 Beyond the Classical Approach

Although many were unmoved by Olshausen and Field's all-out attack on the classical approach, the sentiment that the standard V1 model needed to be tested against natural stimuli seems to have been widespread already around the turn of the century (e.g., Carandini et al. 2005, 10577). A number of experiments recorded from V1 of primates and other mammals, sometimes awake, and mapped neuronal responses to natural images, comparing them with those obtained under artificial stimulus conditions (e.g., David, Vinje, and Gallant 2004, Smyth et al. 2003). This opened a Pandora's box of nonlinearities—contextual effects not seen previously.[35] Still, one can ask why prediction of activity in the wild was so widely accepted as the gold standard for the testing of encoding models. To compare, the inability of classical mechanics to predict the falling trajectory of a paper note dropped

35. This happens because a natural image abounds with features different from the preferred stimulus of any one neuron. The activations of the neighboring cortical neurons that are responsive to those other features will modulate the activity of the recorded neuron, hence there are contextual effects that can be observed just with slightly more complicated artificial stimuli, such as two sinusoidal gratings superposed at different angles (Bonds 1989).

from a high building into a city square is not taken to be disconfirmatory of those laws, but only to indicate a limitation in their domain of application (Cartwright 1999, 27). Perhaps it is because, unlike those physical laws, the encoding model does produce a prediction, but one found to be inaccurate; or it may be because an ultimate goal of neuroscientific research is translational medicine. Theories must therefore be applicable in the wild if they are to form the basis of therapeutic interventions in uncontrolled conditions.

There are some curious similarities between the criticisms leveled against the reflex theory and the reasons given for the turn against the classical approach to research on V1 and other brain areas. The themes that recur are the rejection of the localizationist assumption that brain areas are strictly specialized (e.g., for sensory or motor functions), worries about the artificiality of laboratory-induced simplifications, and assertions of the importance of understanding animal behavior as it occurs in the wild—a revalidation of the science of ethology.[36] These three notes can be seen in the following passage from Sonkusare, Breakspear, and Guo (2019, 699), reviewing functional magnetic resonance imaging (fMRI) research on human perception and cognition:

> Cognitive neuroscience has traditionally relied upon relatively simple parametric tasks using abstract stimuli, delivered with strictly controlled and sparse temporal order. Such designs tightly control the variables involved and isolate targeted behavioural or cognitive constructs as much as possible, classically driven by the "localisationist" objective of assigning specific cognitive processes to discrete brain regions. A suite of carefully designed parametric tasks has been the mainstay of cognitive neuroscience research and enabled fundamental insights into our understanding of brain–behaviour relationships. However, the ecological validity of these abstract, laboratory-style experiments is debatable, as in many ways they do not resemble the complexity and dynamics of stimuli and behaviours in real-life.

In the remainder of this section, we will examine the direction that has been followed on the basis of such concerns. The interesting thing is that the new trajectory of research does not follow a straightforward pattern whereby a rough, oversimplified model (the first approximation) is replaced by ones

36. Olshausen and Field (2006, 206) say: "Reductionism does have its place, but it needs to be motivated by functionally and ecologically relevant questions, similar to the European tradition in ethology."

See also Musall, Urai, et al. (2019), Parker et al. (2020, 581), and Nastase, Goldstein, and Hasson (2020).

that improve accuracy by adding correcting factors to the original (the second and third approximations). Although this approach was tried in various ways (e.g., Heeger 1992), what we see now is a more wholesale rejection of the classical approach, with a complete overhaul of experimental and modeling techniques. And yet, as the story will go, even this has not led to an account of V1 that is free of the intermediary of artifice. Scientists are finding new ways to strip down the complexity of the cortex through the study of artificial visual systems in silico.

### 5.2.1 The Big Data, Ethological Turn

I have noted three important ways that simplification was brought about in classic V1 experiments: behavioral restraint, stimulus constraint, and use of single-unit recording. The practice I call the "big data, ethological turn" departs from each of these methodologies to a fairly extreme degree. From the turn of the century until now, research has been conducted that takes new directions with respect to scale of recording, behavior and stimulus conditions; but that is not to say that classic paradigms have ceased to be used altogether. We will now review these trends.

**Scale of Recording**[37] The fifty years between 1960 and 2010 saw exponential growth in the number of neurons that could be simultaneously recorded (Stevenson and Kording 2011). The first studies of Hubel and Wiesel employed a single electrode, while multielectrode recordings starting in the 1970s increased capacity to tens of neurons at a time. By 2010, use of the 100-electrode Utah array, invented in the early 1990s (Jones, Campbell, and Normann 1992), meant that it was commonplace for publications to report the activity of over 100 neurons at a time. In the last decade, this trend toward larger-scale recordings has only accelerated. As reported by Steinmetz et al. (2018), with a useful review of prior innovations, the Neuropixels recording probe allows for about 1,000 recording sites per device. Sahasrabuddhe et al. (2021) have recently reported the development of the Argo microelectrode array, which offers simultaneous recording from over 65,000 channels. Two-photon calcium imaging is a fundamentally different strategy for the measurement of neural activity, which came into use

37. This section is by no means an exhaustive review of new recording methods. The articles cited here contain much information about additional techniques that I do not mention.

at the turn of the century (Smetters, Majewska, and Yuste 1999; Mao et al. 2001). Instead of directly recording the voltage changes that occur as neurons fire action potentials, it detects the influx of calcium ions occurring when neurons are activated and signaling. Stosiek et al. (2003) demonstrated that calcium imaging could be used *in vivo* to monitor the activity of a network of 100s of neurons in the mouse brain. By the publication of Stringer et al.'s (2019) study of mouse visual cortex, calcium imaging had scaled up to encompass the observation of around 10,000 neurons. To state the obvious, this is an immense increase in the amount of data that can be generated in any one physiological experiment.

The Allen Institute for Brain Science, a research institute launched by a donation from Microsoft cofounder Paul G. Allen, is one of the institutions in which large-scale recording of the visual cortex is now being conducted. For example, de Vries et al. (2020) published a data set resulting from a Taylorist restructuring of the neurophysiological experiment in which 59,610 neurons from six visual areas in 243 mice were observed using two-photon calcium imaging. The article lists seventy-two authors. Their experimental "pipeline" was designed to ensure that data would be standardized across experiments, and it is said to be motivated by the arguments of Olshausen and Field (2006), whom they cite, regarding bias in the old approach. The virtue of their nonselective survey of tens of thousands of neurons, they indicate, is that it reduces these sources of bias.[38]

The development of these methods is perhaps a case of innovation occurring by its own momentum, as well as through targeted allocation of resources form funding agencies.[39] Still, the theoretical motivations and ramifications are highly relevant to our discussion. We find protagonists in the invention of

38. Further, de Vries et al. (2020, 149) write that earlier research "may have failed to capture the variability of responses, the breadth of features that will elicit a neural response, and the breadth of features that do not elicit a response. This results in systematic bias in the measurement of neurons and a confirmation bias regarding model assumptions."

39. Innovation in this area has been viewed by funding agencies as the path for progress in neuroscience toward remedies for prevalent brain disorders. Since 2014, the federally supported Brain Research through Advancing Innovative Neurotechnologies (BRAIN) Initiative in the US has made available millions of dollars for the development of new neural recording techniques, among other tools for experimental neuroscience. See https://braininitiative.nih.gov.

large-scale recording methods rejecting Barlow's (1972) theory of the single-neuron code, and proposing instead that it is populations of neurons that encode information about the world. In a paper that helped prompt the BRAIN Initiative by the National Institutes of Health (see footnote 39), the authors criticize the old methods of recording one or a handful of neurons at a time, on the grounds that the neuron doctrine is likely to be false: "It is probable that neuronal ensembles operate at a multineuronal level of organization, one that will be invisible from single neuron recordings, just as it would be pointless to view an HDTV program by looking just at one or a few pixels on a screen" (Alivisatos et al. 2012, 970).[40]

Fusi, Miller, and Rigotti (2016) explain that in a population code using nonlinear "mixed-selectivity" neurons, the response of any one neuron will be context dependent, and hence uninterpretable without knowledge of the activity of its peers. This, they say, is consistent with "a recent update to the neuron doctrine notion, that ensembles, not individual neurons, are the functional unit of the nervous system" (Fusi et al. 2016, 37). Saxena and Cunningham (2019, 109) similarly draw the connection between the availability of large-scale recording and the "rapidly growing trend in the field towards the *neural population doctrine*."

We have seen that Barlow's neuron doctrine came with reductionist assumptions about the adequacy of knowledge of the parts of the system (single neurons) as building blocks for the understanding the whole. It should not surprise us, therefore, that discussion of population-based alternatives to the neuron doctrine often invokes emergence, the old foe of reduction. Saxena and Cunningham (2019, 105) assert that "decoding accuracy"—the ability to read off the "meaning" of a neural code—"is more than the sum of its parts." The article by Alivisatos et al. (2012) begins with a quotation from physicist P. W. Anderson's (1972) emergentist manifesto "More Is Different." Helpfully, they state what they mean by the claim that neural circuit function is emergent—that it "could arise from complex interactions

40. Cf. Yuste (2015, 488–489) and Yuste and Church (2014). If Barlow's (1972) neuron doctrine were true, the activation of any one neuron by itself would tell a story, so to speak, being selective for a particular kind of feature in the world. Fusi et al. (2016) note that the classical picture of cells being highly selective to one kind of stimulus is more tenable in visual and inferotemporal cortex (ventral visual stream) than in other parts of the brain, such as prefrontal and parietal cortex, where mixed selectivity (i.e., multitasking) appears to be the norm.

among constituents" (970). To elaborate on this, they employ the language of dynamical systems theory (DST), saying that "dynamical attractors" are examples of emergent functional states.[41] Interestingly, these neuroscientists' acceptance of emergence does not lead to critical reflection on the practice of observing neurons to understand the mind, but it is used to justify the project of recording from more and more neurons. In contrast, a more thoroughgoing emergentism would entertain the thought that the whole nervous system, not the circumscribed population, is the true functional unit, perhaps removing the motivation for any monitoring of individual neuronal activity; indeed, it is hard for a consistent emergentism to stop at the brain and avoid the conclusion that the whole organism, and its behavior in context, is the level at which mental life must be studied—an end point in which psychology would reassert itself as the preeminent science of cognition.[42]

**Behavior** In the classical experiments, the modulation of neural responses by the animal's performance of different behaviors—such as orienting or reaching toward a sensory stimulus—was eliminated through general anesthesia. As mentioned previously, Hubel (1959) reported no qualitative differences between the visual cortical activity of an anesthetized and an awake cat, just a generalized dampening of activity. This reassuring finding did not suppress all worries that the response profiles of these neurons could not be comprehensively charted unless the animal was awake. In the early 2000s, there was a trend toward V1 electrophysiology to be performed on "awake behaving" primates—that is, ones with head restraint and behavioral incentives to fixate their eyes on targets imposed by the experimenter. This was not without worry that the experimenter would lose knowledge of where the animal was looking, and hence what the visual stimulus actually was on any given trial. Simultaneous eye tracking was therefore the norm in these experiments.

41. Concepts of nonlinear DST have been employed in numerous recent studies (see chapter 7).

42. This is indeed a point made by a different group of neuroscientists: "The phenomenon at issue here, when making a case for recording from populations of neurons or characterizing whole networks, is *emergence*—neurons in their aggregate organization cause effects that are not apparent in any single neuron. Following this logic, however, leads to the conclusion that behavior itself is emergent from aggregated neural circuits and therefore should also be studied in its own right" (Krakauer et al. 2017, 484; emphasis in original).

It is now commonplace for experiments on sensory physiology in rodents to employ "ethological paradigms" in which the animal is free to move, and so perform the behaviors that would naturally be elicited by the stimuli encountered. One review of research in this vein begins with a critical assessment of the localizationist and reductionist framework assumed in earlier research. Parker et al. (2020, 581) describe how these assumptions supported the feedforward, hierarchical model of sensory processing which is now deemed "an over-simplification, since it overlooks important components such as top-down feedback and various forms of contextual modulation." Instead, experiments in which animals are behaviorally less constrained reveals multitasking in V1 neurons that was not previously observed. For example, movement-related modulation of activity in mouse visual cortex has been reported by various authors (Niell and Stryker 2010; Musall, Kaufman, et al. 2019; Stringer et al. 2019). Fiser et al. (2016) and Saleem et al. (2018) report V1 activity, in a subset of neurons, reminiscent of place cells in the hippocampus, where the level of response to a given stimulus depends on its location on a track. A note of caution is that all these experiments were performed on mice. It could be that there is greater interconnectivity between regions of the cortex in rodents than in primates, which would mean that these phenomena would not be so readily observed in primate visual cortex.

The trend toward richer behaviors interacts importantly with that of larger-scale measurement. With the acquisition of population data sets, the *dimensionality* of neuronal responses has been used as a metric for complexity of brain activity. For example, a raw data set recorded from N=240 neurons will have 240 dimensions. However, the responses of neurons are often correlated with one another, which makes possible the use of dimensionality-reduction techniques such as principal component analysis (PCA) (Cunningham and Yu 2014). A consistent finding of the analysis of data sets of populations of about 100 neurons has been that the dimensionality reduces to around 10, one order of magnitude lower, indicating that the recorded brain activity is far less complex than might be expected. This observation is the starting point for the theory of *neuronal task complexity* (NTC). Gao et al. (2017) propose that the consistent findings of low dimensionality are not due to the inherent simplicity of these brain areas, but rather to the simplicity of the tasks being performed when the data were

collected.[43] Simple tasks have a small number of behavioral parameters, and this puts a bound on the dimensionality of the neural network dynamics, so Gao and Ganguli (2015, 149) argue. Although the NTC metric is not applicable to single neuron data sets, the general point is consistent with the lessons of section 5.1—namely, that the design of classic V1 experiments made the cortical responses simpler than they otherwise would have been. Conversely, NTC is expected to increase for data sets that use richer behavioral tasks (Whiteway and Butts 2019, 91). While this means that the more complex data sets will be harder to analyze and interpret, it is a mountain that will have to be climbed if researchers' understanding of the brain is to be translatable to applications in the wild.[44]

**Stimuli** It was mentioned at the start of section 5.2 that failure to predict responses to natural images was probably the main strike against the classic encoding models of V1 neurons. I will now discuss some of the findings that came with the trend of using richer stimuli—both natural images and artificial stimuli.[45] The classical RF was defined only by the area of the visual field within which stimulation could achieve an excitatory response, consistent with experiments in which simple stimuli (e.g., a grating at one orientation) were presented, and these were small enough not to expand beyond the confines of the area to which the cell was responsive. However, it had been observed since the publications of Blakemore and Tobin (1972) and Maffei and Fiorentini (1976) that the presence of a contrasting stimulus beyond the classical RF would modulate a cell's responses, sometimes increasing but mostly decreasing the cell's responsivity to the stimulus placed within its classical RF. Albright and Stoner (2002) review a range of such contextual effects, thought to be driven by intracortical input, and discuss their role in the perception of forms and contours.

---

43. More specifically, the NTC theory states that "the dimensionality of neural populations dynamics has an upper bound defined by the number of task parameters and the smoothness of neural trajectories across those parameters" (Musall, Urai, et al. 2019, 230), where the number of task parameters indicates the complexity of the task.

44. For example, on the development of brain-computer interface (BCI) technologies to aid patients with spinal cord injuries, Laiwalla and Nurmikko (2019, 234) remark that NTC is expected to increase "as we move from the realm of highly controlled, experimental BCIs to more naturalistic, deployable systems."

45. These are discussed in more detail in Chirimuuta and Gold (2009).

A feature of the classic theory of visual cortex was that neurons had fixed response properties, except for plasticity in early life during "critical periods" of the development of the visual systems (see Hubel and Wiesel 1977, 46–50). However, it turned out that receptive fields were somewhat mutable, with the extent of summation fields dependent on stimulus contrast (Kapadia, Westheimer, and Gilbert 1999), among other findings (e.g., Cavanaugh, Bair, and Movshon 2002). Recent history of stimulation with natural images, as opposed to artificial ones, was found to cause subtle changes in the response profiles of neurons, requiring different models to fit their responses (David, Vinje, and Gallant 2004).

Given the failings of the mathematically simple LN models, much work has gone into the development of alternatives. More complex mathematical tools are needed to model more complex neural responses, and ANN models are now the state of the art for predicting visual neurons' responses to natural stimuli (e.g., Cadena et al. 2019; Yamins and DiCarlo 2016; McIntosh et al. 2016). Use of these machine learning techniques has the drawback that the resulting models are less intelligible than the earlier ones—they do not offer a transparent theory of the computations performed by these neurons, meaning that the selectivity of the artificial neurons is "inscrutable" Butts (2019, 463).[46]

A further move toward naturalistic stimulus regimes comes with the practice not of presenting an image to an animal on a screen, but in allowing an animal to freely explore its surroundings. Of course, this immediately brings up the problem of the experimenter not knowing what the stimulus actually *is* (Parker et al. 2020, 590). This difficulty, however, can be overcome through the use of sensors such as head-mounted cameras that record all that the animal sees. Again, we should note that the break with the artifice of classic sensory physiology has been made possible only through the invention of highly sophisticated technologies (see Parker et al. 2020, figure 3A).

### 5.2.2 Artifice, Old and New

We have seen that there has been a trend away from what was artificial in classic V1 physiology, a shift directly related to concerns about the localizationist and reductionist assumptions of that account and the biases

46. This point is argued at length in Chirimuuta (2020c) and will also be the topic of section 8.3 in chapter 8.

introduced experimentally. Still, a question looms over the new approach—namely, whether it can yield intelligible data sets and models. Frégnac (2017, 471) worries that the industrialization of neuroscientific methods creates a culture in which scientists are pressured "to use mouse-specific state-of-the-art techniques, irrespective of their adequacy," and that "wishful thinking has replaced the conceptual drive behind experiments, as if using the fanciest tools and exploiting the power of numbers could bring about some epiphany."[47]

To draw this long chapter to a close, I will point out that the newest developments have displaced rather than resolved the problems of artifice and brain complexity. First, the neuroethological paradigms, as much as they might evoke wildlife in untrammeled nature, depend on technologies so sophisticated that they could be science fiction. We saw that cameras and other sensors are required to keep track of the stimulus during free movement, and large-scale population recording is needed to sample broadly and gather the activity of cells with a range of response profiles, not just the visually dependent ones. When sensory responses are analyzed in tandem with movements, the experimenter also needs tools to keep track of the animal's behavior. This is done at scale by taking video recordings of the experiment and then using machine learning to classify the movements into behavioral motifs (Juavinett, Erlich, and Churchland 2018, 47). Such experiments generate more complex neural data sets, and these demand the full statistical firepower of modern machine learning for modeling and analysis. For experiments investigating sensorimotor activity and decision making in freely behaving animals, researchers admit that the data are too hard to theorize without the intermediary of an ANN as an "artificial model organism" (Musall, Urai, et al. 2019, 234).

The irony here is that the big data, neuroethological paradigms were a turn away from the artifice inherent in the classic experiments, and yet artifice has returned in a different guise. To get the correct picture of sensory cortex, it was deemed necessary to study the animal performing naturalistic behaviors—perceiving its environment, locating ecologically relevant objects, and making decisions about them. However, the brain in the wild is too complicated to be directly understood. When an ANN is used as an

47. See Churchland and Sejnowski (2016) for a more positive take on these developments.

artificial model organism, this introduces a new way for the neuroscientist to avoid studying brain activity in its full complexity, by targeting a simplified system in its place, for the purposes of theory and explanation.

Admittedly, reverse-engineering ANNs to understand how they perform simulated tasks is hard,[48] but the ANN is still less complex than an actual brain. As Musall, Urai, et al. (2019: figure 2) depict, an ANN is a gray box in comparison with the black box of the brain. Even if it is somewhat inscrutable, at least the full wiring diagram and connectivity matrix of the ANN are known. These words from Haesemeyer, Schier, and Engert (2019, 1130) might remind us of the quotation from Jean-Pierre Dupuy that opens this chapter: "The principles underlying the operation of ANNs on the other hand are likely easier to dissect because they are made by man and because activity states in such networks can be readily queried." Still, Musall, Urai, et al. (2019, 235) admit that there are "profound conceptual differences" between the ANN and the brain—some of which will be examined in part III of this book. For this reason, it is helpful to follow Hardalupas (2021, chapter 4) in treating ANNs serving the role of surrogate brains as *artificial Krogh organisms*. A Krogh organism is useful because of peculiar features that make it uniquely accessible for certain investigations, not because it is ideally similar to some other target of research (Green, Dietrich, et al. 2018).

## 5.3 Conclusion: What I Have Not Made, I Do Not Understand

The departing words of Richard Feynman—"What I cannot create, I do not understand"—are often used as a rallying cry for workers at the intersection of neuroscience and AI.[49] Yet the lesson from this chapter is that this slogan misrepresents what occurs in computational neuroscience. It is better to say, "What I have not made, I do not understand." As far as we can tell from the practices of neuroscience, past and present, scientists' understanding of the brain relies on there being an artificial stand-in—a simplified cell or an artificial model organism—between themselves and their original object of

48. For instance, see Sussillo and Barak (2013) and section 8.3 in chapter 8.

49. For instance, see Arkhipov et al. (2018, 1), Hasson, Nastase, and Goldstein (2020, 423), and Einevoll et al. (2019, 739). They take it in the sense of re-creation of cognition being the test of the scientist's understanding of it. See section 8.2, n4, on the various meanings of Feynman's saying.

investigation. Moreover, technological success has no inherent connection with understanding the brain in its native complexity. The model of visual cortex that led to the creation of DCNNs was a highly abstract and idealized version of the perceptual process. Despite containing false assumptions, it made possible today's machine vision technologies. And all of this is not to say that artificial intermediaries like ANN models are perfectly understandable. They clearly are not, and the significance of that fact will be discussed in chapter 8. Before then, in the next chapter, we will examine one of the conceptual practices that the standard model of V1 helped foster: the treatment of sensory neurons as encoding features of the environment and producing representations of the item in the world that best stimulates them.

# 6 Why "Neural Representations"?

Theories in cognitive neuroscience often posit representations in the brain, and while the appeal to intentional notions is widespread, their status is controversial. This chapter offers a new interpretation of the theory and the practice. I argue that intentional posits are brought in because they allow a workable model or framework for describing relationships between neural activity and extracranial occurrences in which distal relationships, but not proximal causal interactions, have explanatory relevance. This stands in contrast to mechanistic frameworks, which do not license the black-boxing of proximal causes. The positing of neural representations is a powerful way to abstract away from the details of very complex processes that link patterns of neural activity to their distal triggers or effects in the world beyond the brain. I situate my proposal with respect to current realist and antirealist accounts of neural representations and argue for the advantages of my metaphysically neutral proposal over these alternatives.

## 6.1 The Quandary of Neural Representations

Talk of neural representations is in the common jargon of cognitive neuroscience. Over the last few decades of research, cells in various areas of the primate visual system have been said to represent edges (Marr 1982), textures (Freeman et al. 2013), and faces (Tong et al. 2000); population activity in primary motor cortex has been claimed to provide a representation of intended movements (Georgopoulos, Schwartz, and Kettner 1986); certain hippocampal cells have been characterized as representing places in an animal's surroundings (O'Keefe and Conway 1978). Philosophical discussion of this practice has concerned itself with the matter of whether these so-called

neural representations meet the conditions considered necessary for something being a representation, such as semantic determinacy and normativity.[1] In other words, debate has settled around the question of whether neural representations, so-called, really are genuine representations.

The explanatory and predictive success of cognitive neuroscience has pushed some toward the positive answer (e.g., Shea 2018; Colombo 2014). The explanatory value of the term "neural representation" has been taken as evidence that there *are* neural representations. Still, there are two important worries about the positive answer. One may query whether the talk is justified since so-called neural representations arguably only meet the criteria for less demanding causal but nonintentional[2] notions. Some instances of the relationship between an external factor and neuronal response may even fall below the benchmarks for causality, perhaps no more than covariation. Furthermore, ontological commitment to neural representations sets up the need for a naturalistic theory of intentional content, and there is no generally accepted account (e.g., Sprevak 2013; Egan 2020).

Likewise, problems arise if one settles for a negative answer. The move is revisionary, finding fault with a common scientific form of expression. Moreover, those who deny the existence of neural representations need to provide an alternative account of the persistence of this scientific practice. I will say more about the current philosophical accounts in section 6.3. For now, we should note that the controversy over representation-talk also afflicts the neuroscientific community, with publications appearing both in defense and rebuke of intentional descriptions of neural processes (Brette 2019; Kriegeskorte and Diedrichsen 2019).

My account of representation-talk picks up the thread from chapter 4, where it was argued that the reference to neural systems as undergoing computations should be taken as invoking a rough analogy rather than being literally descriptive. The positing of neural representations is of course related to the computational theory of cognition, where it is asserted that animals' mental capacities are the outcome of a series of computations over representations, implemented in neural tissue (Sprevak and Colombo 2019). Here, I

1. See Egan (2019) on these and other conditions.
2. "Intentional" is the philosopher's term for "representational," and it is sometimes used synonymously with "semantic." An "intentional system" is one that employs representations—some physical tokens that have meaningful (i.e., semantic) content.

contend that the models positing neural representations, as well as related intentional notions such as neural codes, invoke an analogy with artifacts including maps, scripts, pictures, and coding systems that are made by people to represent various things. The talk of neural representations also invokes an analogy between neural activity and the intelligent behavior of whole animals—its robust sensitivity to distal goals over proximal disturbances. What both of these analogy sources have in common—the intentional relationship between a sign and its object, the intending behavior of a creature—is their indifference to intermediary factors: the statement that some item is a sign for something can be upheld regardless of the causal relationship between these two; the same is true for the account of a behavior as being directed at a goal. My argument will be that this bracketing, or black-boxing, of intermediary causal factors is what is most useful to the cognitive neuroscientist whose task is to explain the relationship between neural activity and distal objects, states and events in the extracranial world.[3] In my exposition, I will focus on the analogy with representational artifacts or "public representations." The interpretation of representational notions in cognitive science as grounded in an analogy with these has precedent in the accounts of Godfrey-Smith (2004) and Coelho Mollo (2021), although my account is more indebted to the ideas of Mary Hesse.

My account has three basic aims. The first is to give a charitable interpretation of representation-talk in neuroscience. I will be nonrevisionary, accepting the positing of neural representations and leaving representation-talk "as it is." This aim will cover even the hard cases to defend against elimination, showing why representation-talk can be endorsed even when neural activity is involved in cognitive performances not deemed "representation-hungry."[4] Thus, I will focus my attention on the "receptor notion" from sensory neuroscience, whereby representations are posited just when "some sort of internal state reliably responds to, is caused by, or in some way nomically depends

3. To be concise, I mostly just talk of distal objects in what follows; this should be read as including states of affairs and events as well.

4. These are cases where either the task "involves reasoning about absent, non-existent, or counterfactual states of affairs" or "requires the agent to be selectively sensitive to parameters whose ambient physical manifestations are complex and unruly" (Clark and Toribio 1994, 419). Perception and motor control are not generally considered representation-hungry because they involve responses to or activity directed toward objects that are present in the surroundings.

upon some external condition" (Ramsey 2007, 123)[5]—the notion that Ramsey most strongly criticizes as untenable.[6] My interpretative task may be contrasted with the more constructive task that I am not undertaking—that of creating a new definition of "neural representation" to be offered for theoretical employment.

In order for charity not to be quietism for sake of it, I need to satisfy a second aim, which is to account for the epistemic benefit of positing neural representations. Much of the time spent in section 6.2 will be on this point. My third aim is to remain neutral on the metaphysics of representations, intentionality, content, and causation. The third aim of metaphysical neutrality is actually driven by the first aim of charity. I want to avoid making the epistemological justification of this scientific practice contingent on the buy-in to a particular metaphysics of representation, such as a naturalized theory of content. A charitable interpretation is one that does not make the validity of the scientific practice hostage to neuroscientists and philosophers having a correct metaphysics of intentionality or causality, since even just having a consensus account is a remote goal, given the level of contestation over this matter. I will now set out my preferred way to interpret neural representation-talk, and then, in section 6.3, it will be compared with rival accounts. We will see that the biggest point of difference comes from my insistence on metaphysical neutrality. Finally, section 6.4 considers the upshot of these comparisons, especially in relation to the mainstream naturalism assumed by rival accounts.

## 6.2 Neural Representations and Simplification

In keeping with the theme of this book, my interpretation of representation-talk in neuroscience is centered on the claim that what we have here is

5. Cf. Rule, O'Leary, and Harvey (2019, 141): "We take 'representations' to mean neural activity that is correlated with task-related stimuli, actions, and cognitive variables."
6. See also Krakauer (2021) and Barack and Krakauer (2021) for reiteration of these complaints against the receptor notion. In what follows, I argue that Ramsey and Krakauer overlook the utility of the receptor notion and exaggerate the suitability of ordinary causal explanation in sensory neuroscience. I do not disagree with Krakauer that more stringent notions of representation are called for in other branches of cognitive neuroscience, where planning and reasoning are under investigation. See Behrens et al. (2018) on such cases, in relation to the idea of the "cognitive map."

a simplifying strategy.[7] In case this claim seems puzzling, I will go over some preliminaries, making some very general observations about scientific methodology.

### 6.2.1 Proximality as a Research Strategy

A tried, tested, and intuitive methodological strategy in science is to seek local causal explanations. It is a policy associated with the mechanistic philosophies of Descartes and Leibniz, which held that nature was intelligible to the extent that observable physical phenomena could be put down to the pushes and pulls of tiny parts, in contact with one another when they interact–such pushes and pulls being the epitome of efficient causes. By the same token, it is a view that rejects the possibility of action at a distance and for that reason, followers of Descartes, as well as Leibniz himself, would not accept Newton's theory of gravity.[8] I will call this outlook a *proximality*-seeking one. When an apparently remote cause *A* is known to affect *B*, the strategy is to discover all the links in the chain of efficient causes between *A* and *B* so the effect of *A* on *B* can be understood as an unbroken series of actions by contact.

It is a further commitment of this approach that more spatiotemporally remote causal factors are "screened off" by the more proximal ones: if *A* acts on *B* via an intermediary *a*, closer spatiotemporally to *B* than *A* is, then *A*'s effect on *B* can in principle be fully potentiated or overridden by alterations made to *a*. Indeed, the intuition at play in this picture of causal-mechanical systems is that influence scales with proximity: the more remote a cause is from its effect, the more ways there are for the causal chain to be broken, and the less likely it is to bring about its effect, whereas the most proximate cause (if determinate) is guaranteed to be influential.

In *Forces and Fields*, Hesse writes that both the doctrine of action at a distance and its antithesis, the continuous action view, have scientific regulative principles associated with them. The regulative principle associated with the continuous action view is what I call the *proximality principle*. Hesse (1962, 291) articulates this principle as the maxim to "always look for continuously

7. Perhaps this oversimplifies the conceptual territory of representation terms in neuroscience. As I warned in chapter 1, this study of abstraction in neuroscience is itself an abstraction.

8. See Hesse (1955, 1962), Dear (2006, chapter 1), and de Regt (2017, chapter 5) on changing attitudes toward the intelligibility of action at a distance in the history of physics.

acting causes." Hesse goes on to assess the heuristic value of the principle, as due to its recommendation to the scientist to seek out explanations of ever more finely defined phenomena:

> It looks as though the second directive [to seek continuously acting causes] will encourage the construction of more fruitful models, because models conforming to it will have to contain descriptions of "interphenomena" as well as "phenomena," and in general . . . , it is eventually possible to devise further experiments to detect the interphenomena predicted. Continuous action therefore appears to be more powerful as a predictive model, and to make more claims upon the facts. (1962, 291)

But she concludes that this principle does not have overarching superiority over the one allowing action to take place at a distance (1962, 292).

Mechanistic modeling in biology today conducts itself under this guiding picture of nature as an unbroken chain of local causal interactions or a nested set of densely interacting mechanisms. The proximality principle shows itself in microbiological research that determines the causal processes occurring in living systems to ever-greater standards of precision and detail. Craver and Darden (2013) spell out the operating principles of this kind of mechanistic research and note that the prolonged presence of black boxes in a model is a bad thing—a "vice of boxology." We should think of these black boxes as structures in which a nonproximal causal relationship between A and B is noted, but with no detail of how this relationship is mediated. At best, black boxes in a model are pointers for future research to fill in the gaps in information. At worst, they are signs of superficiality and incompleteness, which are two of the three "classes of failure" for models or mechanism schemata. Craver and Darden describe the "vice of boxology" by saying that it is

> the vice of operating with incomplete schemas for which one cannot pass the "And how does that work?" test. But a goal of science is to push beyond the levels of understanding of everyday life to reveal the internal mechanisms by which things work. (2013, 90–91)

The observation that Newtonian gravity should not be understood as *causal* action at a distance (Norton 2007, 16) is another reason to surmise that the proximality principle is tied to causal reasoning about natural systems, and that departures from the proximality principle involve shifts away from application of the familiar notions of efficient causation that constitute a "folk science" according to Norton.[9]

9. But see section 6.2.4 ,where I discuss how the interventionist account of causation departs from the proximality principle.

However, we should note that various branches of modern science do not employ the proximality principle. I have already mentioned Newton's theory of gravity. Current theories of gravity should also be listed here. Quantum mechanics is notorious for its "spooky" nonlocal effects.[10] Closer to neuroscience, we should consider the distinction between proximate and ultimate causal explanation in biology. Ernst Mayr (1961) argued that there are two contrasting and complementary explanations of the warbler's migrating south. The ultimate causal explanation refers to the evolutionary reasons for this occurrence. It begins way back in the mists of time, before the birth of the bird in question, and black boxes the details of genetics and physiology that would explain precisely how the adaptive behavior occurs. The proximate causal explanation begins its story within the lifetime of the organism and reveals at least some of the inner workings of the black box, describing the environmental triggers for the hormonal mechanisms that bring about migratory flight.

The context for Mayr's publication was a concern over the rising status of molecular biology, a discipline with the reductionistic aim of applying only the concepts and techniques of the physical-chemical sciences to the examination of organisms (Beatty 1994). Seeking a role for a more autonomous and less reductive kind of biology, Mayr argued that ultimate, evolutionary explanations could not be tackled with the same molecular approach that had successfully taken on other kinds of problems. We should note here that Mayr's proximate explanations adhere to the proximality principle, whereas ultimate explanations are dissociated from it, for it is impossible to complete the push-and-pull details of a causal explanation that begins millions of years ago, where the initial causes are in deep evolutionary history. We can see why the proximate explanation, as well as various methodologies adhering to the proximality principle, tend to be reductionistic. This norm of explanation is dissatisfied when there are unfilled gaps in the account of causal interactions, so the rule is to search in the micro-details of the system for the processes that will remove the lacuna. Consider the ingestion of a tablet like Xanax and its effect on mood. The details of the causal interaction would be filled in by referring to processes happening at a lower scale than the observed cause and effect—that is, at the level of the microconstituents of the pill and the synapses in the person's brain. A common metaphor for reductive science is

10. The important point here is just that contemporary physics does not rule out action at a distance (Norton 2007, 18).

that of "digging down" to a lower stratum of micromechanisms. Similarly, in causal explanation adhering to the proximality principle, we can think of the scientist as tunneling underneath the surface level and digging through the mediating microcauses to find out how the effect comes about. This reductive tunneling strategy is, we can observe, feasible only if the cause is quite spatially and temporally close to the effect. With cause and effect separated by eons (as in evolutionary biology), it is impossible to tunnel down and reconstruct a detailed causal history—the deeds of all the warbler's ancestors from prehistory to the present day.

Although Mayr's ultimate explanations are still causal ones, we must appreciate that in departing from the proximality principle, they depart from the norm whereby an ideal explanation of a phenomenon would show how it is the result of a series of efficient causal processes. Indeed, they are something of a throwback to Aristotle's explanations in terms of final causes (distinct from efficient causes), except that the ultimate cause is without doubt in the past, preceding the effect. This is consistent with Mayr's program elsewhere, that of modernizing teleological notions by rebranding them as "teleonomic" ones (Mayr 1988). The common feature to be noted in the explanations in physics and biology that depart from the proximality principle is that they decline to represent those phenomena as being governed by a series of efficient causes at some fundamental level. This point will become significant in section 6.4.

### 6.2.2 The Need for Distality in the Research Strategy of Cognitive Neuroscience

When research is directed toward the question of how brain activity underlies intelligent, adaptive behavior—spelling out how the operation of perceptual, motor, and other systems leads to appropriate actions in a complex environment—investigation must focus on the relationship between neural responses and a distal perceptual stimulus or motor action. The details of how that relationship is mediated (e.g., the transmission of light or sound waves from object to eye or ear, the physiology of efferent nerves from spinal cord to arm, the neuromuscular junction, and so forth) are subsidiary. Those details do not have explanatory relevance to the cognitive neuroscientist, in comparison to the nonproximal relationship between brain activity and extracranial state. An explanatory framework bound to the proximality principle is not appropriate since it gives no grounds for

privileging the connection between brain and distal object over the many intermediary structures and occurrences closer to the brain. Moreover, if a neuroscientist attempted to dig the explanatory tunnel from the distal stimulus to the neural response, including an account of all the microinteractions along the causal pathway, she would soon find herself caved in under a mass of mechanistic details, given the intricacy of the processes that occur. Another problem with the proximality principle is that brains differ widely in their details from one individual to another, and even within one individual across a lifetime, such that increases in the detail given in a causal explanation come with losses in its ability to generalize.

A framework is needed for conceptualizing the relationship between brain activity and extracranial state that forsakes proximality and embraces distality in its research strategy. The advantage is that such a strategy can justify the black-boxing of the details of the mediating causal processes, treating the simplification that it affords as virtue and not a vice. It is here that an analogy with the relationship between sign and signified, vehicle and content, in artifacts made to serve as representations, shows its use to the neuroscientist. With these artifacts, the intentional relationship between a concrete sign and the object that it signifies can reach indefinitely far in space and time: pieces of writing, pictures, and numbers can all refer to events at the beginning of the universe and light years away. Indeed, the intentional relationship should not even be treated in the same way as a causal one, one that connects spatially and temporally located items. The relationship between such inscriptions as "the unicorn of the Bermuda Triangle" and its intentional target, which is not an object located in space and time, cannot be a causal one. Since the intentional relationship is not, on the face of it, bound to causation, consideration of the mediating link between neural response and distal item can be bypassed in explanations that cite intentional content.[11]

It follows that if one treats the relationship between neural activity and extracranial state (a stimulus or bodily action) according to the analogy with representational artifacts, as an intentional relationship, the door is open to treating it in a way that departs from the proximality principle. This may not

11. Of course, naturalistic explanations of original intentionality of mental states (as opposed to the derived intentionality of artifacts) do attempt to show that the intentional relationship reduces to certain kinds of causal ones. These attempts will be under discussion in section 6.4.

be the only way to go beyond proximality, but it is the route traditionally taken. The neuroscientist buys into a framework in which it follows naturally that she can ignore the question of how representation and object are mediated, allowing her to focus on the fact of their connection and its functional significance. Moreover, this captures the property that neurophysiologists sometimes report of sensory neurons—namely, that their responses seem to "reach out" to features of the world around them, regardless of the details of the intervening medium. Here is Horace Barlow (1972, 373): "The properties of the retina are such that a ganglion cell can, figuratively speaking, reach out and determine that something specific is happening in front of the eye. Light is the agent by which it does this, but it is the detailed pattern of the light that carries the information, and the overall level of illumination prevailing at the time is almost totally disregarded."

In chapter 5, we saw that Barlow's theory of single-neuron coding has been roundly criticized. But the point about the importance of distal relationships bears just as much on the alternative theories of population coding, only that it is groups of neurons, rather than individual cells, whose activity is said to represent some extracranial object.[12]

I have proposed that representation artifacts are the source for an analogy through which the relationships between distal, environmental events and neural activity studied in cognitive neuroscience can usefully be modeled. There is an attractive synergy with the account offered in chapter 4 of the other dominant technological analogy in neuroscience—the brain as computer. An alternative analogy source would be person-level mental representations, the putative bearers of original rather than derived intentionality. A reason to discount this option comes if we consider that the fruitfulness of scientific analogies depends on there being a *relatively* well understood analogy source. Although mental representations might strike some as being more intimately known than anything else, their operation is opaque, and they may be no more than a hypothetical posit of psychology and cognitive science. Representation artifacts are more perspicacious than person-level mental representations, for we can say what the vehicles, as opposed to their contents, are, distinguish the sender and consumer, and specify the contents

12. For ease of exposition, I will mostly give single-neuron examples in what follows.

(usually) to a high degree of determinacy, whereas all this becomes contentious when talking of mental representations.[13] As Hesse (1955, 353) maintains, analogies serve "to enable the new and unfamiliar to be thought about and described in terms of the familiar." Granted that the point of analogies in science is to put the unknown into view by appeal to similarity with something already well characterized, comparisons between neural and mental representations would not be helpful because both are quite murky.[14] We will see in section 6.2.3 that an analogy with person-level, goal directed behavior does have a crucial role to play in theorizing neural representations. I note here that publicly observable behaviors are more familiar and less hypothetical than the idea of mental representations.

Consistent with my proposal, we see that the use of terminology borrowed from artificial systems of representation is all over cognitive and theoretical neuroscience. The notions of codes, decoding, reading and writing, maps, and even pictures are abundant in mainstream research.[15] Although metaphors drawn from person-level cognitive performances are also present—such as the retinal neuron that "perceives"[16]—such analogies have not received the theo-

13. Much work in philosophy of mind shows this. See, for instance, Neander (2017, chapter 7) on the problem of indeterminacy.

14. A concern to be mentioned here is that the intentionality of public representations is only derived from the original intentionality of mental representations. As such, it can be objected that public representations cannot serve as the familiar analogy source for neural representations—only mental representations could, if anything. My response is that we must appreciate that the talk of familiarity here is a reference to our everyday use of representational artifacts, which can happen quite effortlessly following inculturation into the system of representations and does not, in the everyday practice, raise philosophical perplexities about the source of these objects' intentionality, their grounding in the mental. My point is that in these practices, the intentionality is just taken for granted, and it is obvious to people schooled in a particular system of representation, what, for instance, the content of those items is. It is this everyday familiarity that the analogy of neural representations helps itself to.

15. See deCharms and Zador (2000) and the references therein. See Brette (2019) for dissent from the mainstream positing of neural codes and representations, and section 7.2.2 in chapter 7 mentions some neuroscientists who reject the positing of representation in motor cortex.

16. See Figdor (2018) on attributions like this. Her literalism about them is a different matter from the view targeted in this chapter—which is the literal interpretation of attribution of intentional content to neural states. However, an important

retical elaboration and regimentation granted to the technological analogies, where coding and decoding are given formal treatments.

Critics and doubters of neural representations often argue that less demanding notions of correlation or triggering would be just as adequate to describe sensory neurons' reliable responses to specific distal stimuli, with less of the contestable baggage that comes with the positing of neural activity as representing something. However, we can now see that just invoking the less demanding notion of correlation between neural activity and extradermal state does not facilitate the exclusive focus on the distal object needed by the cognitive neuroscientist. This is simply because there are too many processes, which can equally be said to correlate with the brain response, along the mediating chain from the distal object. Similarly, the notion of triggering does not isolate the distal object. Triggering is a causal concept which means that it opens the door to holding that any item or event along the mediating chain is a trigger of the neural response that it causes.

In short, the proposal here is that appeal to neural representation allows an important simplification in cognitive neuroscience because it facilitates and justifies the black-boxing of intermediaries between brain responses and the distal objects that are relevant to the explanation of behavior. Thus, if the task is to explain how neural responses in sensory cortex enable a mouse to locate food, and how activity in motor cortex enables a mouse to move its body in the direction of that food, the key component of the explanation is the relationship between distal items—the smell of something edible, the layout of the territory, locomotion in the direction of food—and the cortex. The detailed causal mechanisms of olfactory reception, afferent and efferent signaling to and from the cortex, and neuromuscular activation can be relegated to the background to allow focus on the now more significant connections between brain, body, and environment.

But that is not to say that accounts referring to more detailed causal processes have no explanatory relevance and must always be relegated to the background of neuroscientific investigation. An example will show how the account of representation-talk as a simplifying strategy fits with a picture of explanatory pluralism. Take the explanandum "Why does this fusiform face

---

difference between Figdor and myself is that she assimilates analogical interpretations to literalism, whereas I treat them as substantially different from both literal and metaphorical interpretations of scientific theories and models.

area (FFA) cell fire more action potentials when a face stimulus is flashed up?" It affords two different answers: proximate and distal. The proximate explanation traces the immediate causes of the spiking—the release of excitatory neurotransmitters by neurons whose axons terminate at the FFA cell's dendrites, which leads to depolarization of the FFA cell above the threshold of action potential initiation. A proximate explanation might reach to a few steps further back in this causal chain, but the processes are too complex for it to be manageable to proceed very far without detriment to the methodological principle of delivering a gapless account of causal interactions. It is important to appreciate that detailed explanations of extended causal process quickly become unwieldy not because I am attributing to the proximate explanation the requirement that it should bottom out at some lowest level of explanation (e.g., quantum mechanics); for even when we go no further than the standard "bottom-out" level for neurobiology—the level that describes macromolecules and their configurations, and ion fluxes across the cell membrane (Machamer, Darden, and Craver 2000, 14)—extended causal tracing is unmanageable.

The distal explanation refers to the object in the environment that elicits the response, black-boxing the mediating processes. The FFA cell responds to this stimulus since it belongs to a cortical area whose specialized function is to represent faces to enable face recognition. The cell's responses are finely tuned to such stimuli, just as the function of a portrait is to represent a face and thus is sensitive to its features.[17] A key point is that the standard way of modeling and characterizing sensory neurons, in terms of their receptive field, is a nonproximal approach. A visual neuron's receptive field is conceptualized as located outside the animal, in the visual field, and a somatosensory neuron's receptive field as somewhere in the body of the animal. Yet elevated firing rate in response to stimulus is explained by the match between that stimulus and receptive field.

The plurality of explanatory forms is reflected in disciplinary plurality. Cellular neurobiology is the usual source for proximate causal explanations, whereas cognitive neuroscience is most ready to offer explanations of neural activity in terms of its relationship to distal objects and events. The reason for the dominance of the distal kind of explanation in sensory and motor

17. This is the standard account of FFA given by such papers as Kanwisher and Yovel (2006), though it is not universally accepted.

cognitive neuroscience is not just that the explanandum (appropriate behavior in response to environmental stimuli) is defined distally, but also that distal objects provide a highly effective set of control handles for sensory and motor cortex neurons. For many neural systems, the precision and effectiveness of control are not proportional to its proximity to the target. Precise control of states in a person's brain can readily be obtained by modulating quite distant variables (like the contents of their Twitter feed), and this is easier to achieve than through more immediate neural interventions. It is a trivial point, but an important one, that I can modulate activity of contrast sensitive cells in your visual cortex by flashing up this . . . **on the page** . . . from my location in space and time so remote from where you are now. This is better control than I could achieve even with state-of-the-art neurotechnologies such as optogenetics or Designer Receptors Exclusively Activated by Designer Drugs (DREADDs). Indeed, it is currently impossible to cause complex images or beliefs to arise in someone's brain through direct neural stimulation, and yet together all the typed words in this book, if my task is successful, will induce you to have a particular set of beliefs. This fact about the brain—that influence does not diminish with distance—defies the intuition that we have about causal-mechanical systems, that causal relevance or impact scales up with proximity. In situations where this intuition holds, application of the proximality principle is all that is needed. The fact that neural systems do not act in the way that the proximality principle would lead us to anticipate shows us why an alternative frame, not subject to this principle, is required.

To reiterate, this is not to say that neuroscientists do not investigate proximal causes. Scientific disciplines grow around techniques of experimental intervention—sets of control handles that require their own kind of laboratory provisioning.[18] Thus, the cellular neurobiologist relies on high-tech, invasive methods to intervene on neural activity. What is significant about the control handles used to modulate neuronal activity in cognitive neuroscience is that they are distal to the brain, quite low-tech, but extremely effective. As neurophysiologists from the time of Hubel and Wiesel to now have learned,

18. This relates to perspectival pluralism, the idea that different scientific perspectives grow around particular choices in how to simplify subject matter and that a plurality of approaches is needed in neuroscience, given the complexity of the brain. See a related argument for pluralism in Longino (2006, 2013).

the most effective way to modulate the activity of a visual cortical neuron is to find its most preferred stimulus.[19] The putative representational targets of cortical neurons are at the same time the most obvious loci of control.

To conclude this section, I would like to emphasize that my class of explanations in which representation-talk is well motivated is not restricted to cases where a behavior seems to depend on "off-line" reasoning about objects not currently present in the environment—the so-called representation-hungry tasks. A common argument against eliminativism about representations is that the need to posit representations shows itself when a system is engaged in off-line reasoning, and so needs a representation to serve as a stand-in for an absent object (Clark and Toribio 1994; Colombo 2014, 225). Yet much of the representation-talk in cognitive neuroscience concerns neural responses to available objects, and the "stand-in" justification provides no answer to the eliminativist in such cases. By recognizing that the positing of representations offers a convenient simplification, we see that it adds epistemic value, even for "on-line" cognitive activities, like perception and motor control.

### 6.2.3 Tidal Representations?

A concern may have already surfaced in the reader's mind. Earlier in this discussion, it was noted that departures from the proximality principle occur elsewhere in science (namely, in physics and evolutionary biology). My justification for the cognitive neuroscientist positing representations seemed only to rest on the need to depart from the proximality principle. Do I mean to suggest that a physicist would be equally justified in positing that tidal activity represents the position of the moon, since her treatment of the relationship between moon and tides departs from the proximality principle? There is obviously an incompleteness in the account presented so far. I have asserted that the analogy with public representations is apt for neural activity, but I have not said why it is any more apt in neuroscience than in gravitational physics. I need now to mention an additional feature of the relationship between neural activity and distal states that makes the analogy

19. See Hubel and Wiesel (1998) for recollections of the discovery of the barlike stimuli that best activate neurons in primary visual cortex; and also see Bashivan, Kar, and DiCarlo (2019) on the use of machine learning to discover the optimal stimuli for neurons in another visual area, V4.

with the intentional relationship appropriate in a way that it is not for the tidal case. This is due to the robustness of the distal relationship between neural activity and object, over variations in the more proximal connections.

Consider some experiments on movement control using brain computer interfaces, which introduce perturbations to the relationship between activity in motor cortex and the angle of movement of a computer cursor so that the subject can no longer accurately direct the movement toward targets presented (Jarosiewicz et al. 2008).[20] In response to such perturbations, activity in motor cortex plastically modifies itself, restoring accuracy of the movements. It is natural to explain these results by appealing to intentional notions—adjustment of cortical representations to restore a mapping between neuronal activity and movement. Indeed, it is fitting to treat this plasticity and robustness of mapping as characteristic of the system being intentional in the etymological sense—as the motor cortex *aiming* at its intentional object, and hence reorienting itself toward it even after a perturbation. In contrast, the nonproximal relationships posited in physical systems do not evoke any sense of purposiveness or goal directedness.

Returning to the worry over the overgeneralization of my account to "tidal representations," the new thing to consider is that in addition to being distal, the relationship between neural activity and extracranial object exhibits robustness. A moment's consideration will make apparent that the distal relationship between moon and tide does not share this feature. If it did exhibit robustness—with a specific height of the tide robustly relating to a specific position of the moon—what we would find is that when a local perturbation is made to the beach, such as with a wall being built to stop the tide reaching so far, the seawater would somehow adapt its activity to overcome this obstacle and reach the same tidal height as before. That is, someone would have more luck adjusting the course of the seawater by intervening on the position of the moon than by building a wall along the beach. If the relationship between tide and moon did exhibit this robustness, it might make sense to employ the language of representation. But that this scenario sounds so weird in the tidal case is an indication that the intentional notion is not appropriate here.[21]

---

20. See Gilbert, Sigman, and Crist (2001, 684–685) for examples of plasticity in response to lesions in sensory cortex as well.

21. Adherents of dynamical systems theory (DST) who propose to eliminate representation-talk, such as Hutto and Myin (2014), might object to my account that

This weirdness demands further examination. For ordinary physical phenomena, we have the strong intuition that the impact of a remote gravitational influence can be undermined by the effects of closer causal factors (like the building of the wall). And if we consider physical systems understood purely as causal systems, it is impossible for a remote cause to bypass the interference of disturbances in the causal chain closer to the effect and make its presence felt regardless. In contrast, the overriding of proximal factors by distal ones fits an intentional frame naturally enough. In addition, this intentional treatment supports a range of counterfactuals about what *would* happen if the relationship between neural response and distal object were disturbed. Yet here it is apparent that by "intentional treatment," we are now invoking goal-directed behavior, actions, of entire, intelligent creatures showing this characteristic of being distally driven and robust: the homing pigeon that single-mindedly directs itself to the coop in the face of perturbations to the flight path, or the squirrel whose foraging activity is driven by long-term need and not sent off course by immediate stimuli or appetites.

While the notion of robustness might be applied to the intentional relationship that an artifact like a map has with its target, doing so is more of a stretch.[22] So it does seem that in addition to the justification of neural-representation-talk that stems from analogy with representation artifacts, there is also an analogy in play between neural behavior and goal-directed actions of whole animals. This analogy source, as with public representations,

---

their framework is equally well placed to model nonproximal interactions. Dynamicists in fact model neural systems in terms of relationships between variables that need not be causally contiguous, so they also have resources to depart from the proximality principle. However, the dynamical approach, which views cognitive systems through a perspective—a set of modeling techniques—developed in the study of physical systems would not be so able to account for these kind of robust and plastic responses that are never demonstrated in the world of physics. Consider the Watt governor and steam engine system, famously presented by van Gelder (1995) in his argument that cognitive systems could be dynamical rather than computational systems. What a physical artifact like the governor+engine lacks is robustness and plasticity to restore function after damage. The better analogy for the cognitive system is not the governor+engine alone, but governor+engine+engineer. If a weight falls off the spindle arm, disturbing the relationship between arm angle, valve opening, and engine speed, the engineer will come along with a new part to restore the relationship. It is this kind of counterfactual occurrence that is not accounted for in dynamical models.

22. For instance, the text message "Geet me somebing to droonk" can elicit the same response as "Get me something to drink," despite the perturbation.

is more perspicacious than the notion of original intentionality since it stems from common, publicly observable occurrences. The upshot is that there is a complex set of analogies behind talk of neural representations: on the one hand, neural responses are treated as constituting codes and maps, and on the other, they are analogized to the robust and distally sensitive behavior of intelligent creatures. What these two have in common is the indifference of the intentional or intending relationship of these artifacts and goal-directed actions, respectively, to the details of mediating factors between the representation or action and its target.

### 6.2.4 Interventionist Causes, Not Intentional Notions?

Interventionism is currently the most popular framework for explicating causal reasoning in science. Arguably, interventionism breaks with the proximality principle because assertion of a causal relationship between X and Y requires only that interventions on X, where these must fulfill some technical criteria, are accompanied by changes in Y (Woodward 2003, 59). It is a significant break with older accounts of causation in the philosophy of science, particularly *process theory* (Salmon 1984), which held contiguous transmission of a "mark" to be one of the conditions on causal relationships—this being a way of theorizing that captures the intuition behind the proximality principle. According to Woodward, a remote influence on Y, acting at a distance, is just as much a candidate for a cause as a proximal one, so long as the interventionist conditions are met. This account of causation does not motivate or support the proximality principle,[23] so it leads to an objec-

23. I quote the following long passage, in which Woodward seems to state that interventionism does not recommend that investigation into causes in the special sciences conform to the proximality principle, seeking discovery of "fine-grained" intermediate causes between the observed, noncontiguous "macro-causes":

> In common sense and the upper-level sciences, causal relata are often described as operating across spatiotemporal gaps or, alternatively, in a way that is non-specific about the spatiotemporal relationship between cause and effect. Recovery from a disease will typically occur some significant lapse of time after the administration of the drug that causes recovery. A slowdown in economic activity may be caused by the decision of the central bank to raise interest rates but it seems doubtful that there is any clear sense in which the latter event is spatiotemporally contiguous with the former. It is true that in many, but by no means all, cases involving macrocausality, there will exist (from a more fine-grained perspective) a spatio-temporally continuous process linking the cause to its effect. However, even when such processes do exist, upper level causal generalizations often do not specify them and the correctness and utility of the upper level generalizations do not rest on our actually having information about such processes.

tion to my account: given that the preeminent account of causal explanation in the special sciences departs from the proximality principle and accepts that demonstration of remote, noncontiguous causal relationships can be explanatory, not calling for research into intermediate causal relationships, the cognitive neuroscientist has no need to refer to intentional relationships to black-box the details of intermediate causal processes between distal object and neural activations—she need only refer to the interventionist standards shown by Woodward to be prevalent in various sciences such as economics and medicine. Intentional notions are indeed superfluous.

Consideration of this objection brings to light one last feature of the practice of positing intentional relationships in cognitive neuroscience, which is that they are fundamental to the explanation. By this, I mean that in stating what a neuron represents, the scientist is staking a claim on what is most essential to explaining the existence and operation of this neuron within the economy of the brain. A parallel here is with the status of computational descriptions at the top level of explanation for Marr (1982)—they purport to tell you what is most important to know about a cognitive or neural system while the causal details of implementation are secondary to them, both in order of investigation and explanatory significance.[24] More causal details can always be filled in, but *within the explanatory norms of cognitive neuroscience*, the computational or intentional explanations do not lack depth without them.[25]

On the question of whether explanations positing remote causal relationships between neural activations and extracranial objects could count as fundamental, Woodward would seem to deny this. I quote the following

> This feature is captured nicely by interventionist accounts which take the distinctive feature of causal relationships to be exploitability for purposes of manipulation, regardless of whether there is a spatiotemporal gap between cause and effect. (Woodward 2007, 82)

24. Nothing here assumes any autonomy (lack of direct constraint) between causal and intentional or computational descriptions of the systems. See Dennett (1995) for a helpful discussion of the level of computation or function and in what sense it is primary.

25. This marks my disagreement with Piccinini and Craver (2011), who assert that computational and intentional explanations are a kind of superficial causal-mechanistic explanation and therefore do not have a distinct and, as I say here, "fundamental" status. My view that within the explanatory context of cognitive neuroscience, such explanations are self-sufficient and not lacking in depth, aligns with points made by Egan (2017, 154) about the interest relativity of explanation.

passage, in which Woodward observes that assertions of interventionist causal relationships permit "deeper explanation," likely in more "fundamental" terms:

> I fully agree (who would deny this?) that if it is the case that some relationship R to the effect that interventions on X are associated with changes in Y holds (e.g., private school attendance boosts scholastic performance), then of course we should expect that there will be some deeper explanation, perhaps to be found in some other, more fundamental science, for why R holds in the stable way that it does. (Woodward 2014, 699)[26]

What this indicates is that positing even an interventionist causal relationship between two phenomena commits one to a framework in which the relationship between an effect and its spatiotemporally distant cause cannot be fundamental (explanatorily basic) and stands open (at least in principle) to deeper explanation in terms of more proximal processes or mechanisms, or in terms of some fundamental laws of physics. But in cognitive neuroscience, the distal relationship is just what is fundamental and primary in the accounting for neuronal behavior: in an inversion of the hierarchy that would be imposed by causal explanation, the causal-mechanical details of how the distal relationship obtains is secondary to this.

### 6.2.5 Summing Up

I have aimed to show that representation-talk in cognitive neuroscience is well motivated, even for phenomena that seem, on the face of it, to require deployment of less demanding relationships such as triggering and correlation. My argument began with the observation that scientists cannot and should not always follow the proximality principle. Cognitive neuroscientists' departure from the proximality principle, essential for the project of examining the neural basis of behavior, is aided by their drawing on an analogy between neural activity and certain artifacts used as representations, and to some extent on an analogy with goal-directed behavior as well. I noted that experimental and explanatory pluralism has grown up with subdisciplines of neuroscience focused on either distal or proximal relationships.

---

26. Compare also Woodward (2007, 103): "Given a true garden variety causal claim, there will be some associated in-principle physical explanation (or story or account, to use more neutral words) for its holding, and this will include, among other factors, appeal to fundamental laws."

One final point that I should emphasize is that by acknowledging that no more than a rough analogy may hold between neural representations and the representations with which we are most familiar (scripts, maps), we should expect plenty of disanalogies. It may well be that so-called neural representations do not have determinate content or well-defined vehicles and consumers. Indeed, the justification of positing neural representations should not depend on neural systems sharing all these properties in common with public representation. The expectation of there being more than a rough analogy is what causes problems for defenders of neural representations—the perceived need to solve the indeterminacy problem is a good example. Likewise, elmininativist objections such as Ramsey's attack on the receptor notion boil down to the observation of disanalogies between neural responses and the representational artifacts that set the gold standard for being representations. My view is that representation-talk in neuroscience is apt even if the neural representations do not have determinate content or well-defined vehicles and consumers. The scientific practice is justifiable even if the stringent "job description" for being a representation, which originated in the analysis of public representations, is not met.

This summary makes clear that my account satisfies the first two stated desiderata of charity and articulation of the epistemic benefit of representation posits. The third, metaphysical neutrality, is maintained by the view's making no commitment to a metaphysical theory of representation, content, intentionality, or causation. It merely assumes that some items are considered to be representations and says that neural activations are treated by analogy to them, while being agnostic on what makes the uncontroversial cases genuinely representational, and whether neural activations share those properties. Likewise, I have said that ordinary causal explanations are committed to following a proximality principle, and explanations positing intentional relationships to distal objects depart from this methodological principle, but I have remained noncommittal about the metaphysical picture surrounding the proximality principle—the question of whether nature fundamentally is or is not a densely connected "causal nexus." With this metaphysical neutrality, I can justify representation posits in neuroscience without tying their fortunes to the ability of philosophers to develop convincing theories about the fundamental nature of representation and causation, and indeed about the fundamental nature of nature.

## 6.3 Comparisons

Preceding philosophical treatments of neural representations divide into realist and antirealist camps. Realists affirm and antirealists deny that there are patterns of neural activation that satisfy the criteria for intentionality established elsewhere. There are many more published accounts than I can properly discuss in this chapter. Next, I take a representative sample of current realist and antirealist views and assess them with respect to the three desiderata that I set out for my own account: (1) charity, (2) articulation of the epistemic benefit of representation posits, and (3) metaphysical neutrality.

### 6.3.1 Realisms

The most comprehensive argument for realism about neural representations comes from Nicholas Shea.[27] Like me, Shea (2018, 29) places high value in having an account that can explain the successes of science by virtue of its positing neural representations. The difference is his assertion that the reality of neural representations explains the scientific successes, and this then feeds into a project of naturalizing intentionality. In contrast, my explanation of the success of the practice refers to its advantages as a simplifying strategy, and I do not venture into the project of naturalizing intentionality. Thus, both mine and Shea's accounts can satisfy the desiderata of charity and epistemic accountability, but his is not metaphysically neutral. Indeed, a selling point of my explanation of the success of the practice of positing representations is that it does not depend on the acceptance or endorsement of a naturalized theory of content, like Shea's *varitel semantics*, whereas Shea's realism puts the justification of the scientific practice at the mercy of the acceptance of varitel semantics or some such theory.

While Shea's overt explanation of the success of research positing neural representations rests on the truth of the claim that there are such entities, Shea and I converge on the idea that the indispensability of representation-talk lies with its ability to simplify matters. Shea concedes to antirealists like Ramsey and Egan that ordinary causal explanations of the distal relationships explored in sensory neuroscience are available, at least in principle (2018, 29). In answer to the question "Why not just talk about correlations,

27. See also Colombo (2014), Neander (2017), Thomson and Piccinini (2018), and Millikan (2020).

functions, etc., and drop the content talk?" Shea (2018, 205) writes, "The trouble with these views is their complexity. More complex properties are generally less good candidates for explanation." Furthermore, he holds that explanations limited to nonsemantic causal, relationships miss out on the "real patterns" of robust distal relationships occurring across different instances and converging on certain behaviorally significant items in the environment (Shea 2018, 202). In response, Egan (2020) argues that in this, Shea has conceded everything that matters, in that the ultimate justification for positing representations is *pragmatic*. In her own view, nonintentional causal relationships are the only ones that actually obtain, but it is pragmatically beneficial for cognitive neuroscientists to talk about their target systems and models of them having semantic properties like content (more on this in section 6.3.2). My own feeling is that the realist concedes too much in granting that the ordinary causal explanation is always available—even for when the explanandum is a singular event or individual cognizer, not a generality that holds across a range of events or individuals. A lesson from our exploration of the proximality principle was that there are certain systems and relationships in the world that conform to our intuitions about causal processes (such as causal import scaling up with proximity) and for which investigations adhering to the proximality principle are most fruitful; and there are other systems and relationships in nature, such as those examined in cognitive neuroscience, that are not like that. Thus, we should not assume, for the latter case, that ordinary causal explanations (by which we mean ones in alignment with the proximality principle) *are* available in cognitive neuroscience in principle. In fact, a priori belief in their availability amounts to a *metaphysical* view that fundamental reality is a causal nexus (see section 6.4).

The overt point of Bechtel's (2016) account of research on hippocampal place cells is to maintain an interpretation of neuroscientists' representation-talk as ontologically committed, not a mere gloss as supposed by Frances Egan. Implicitly, the paper endorses realism about neural representation. Moreover, identification of neural representations, as well as discovery of their properties, are presented as the culmination of decades of work on the role of the hippocampus in navigation in mammals. While I, like Bechtel, endorse neuroscientists' description of brain areas as having representational functions, I do not go along with the realist drift of his account. This is because it depends on the literalist acceptance of a comparison between brains and control systems that, in my opinion, is better given an analogical interpretation. Bechtel

accounts for the neuroscientists' attribution of representations to neural systems as being based on their assumption that the brain is a control system. In control theory, "a controller needs information about the plant that is being controlled or what the plant is interacting with," and hence it needs representations (Bechtel 2016, 1316).

Bechtel's argument that neuroscientists are right to uphold ontological commitment to neural representations rests on the assumption that the brain, literally, is a control system. Thus, his account does not satisfy the desideratum of metaphysical neutrality. In adhering to this literal interpretation, there is no acknowledgment that control theory has its origin in engineering and is essentially a framework for the analysis of human-built systems. As such, it is an instance of an engineering analogy being applied within neuroscience, and as in the case of the brain-computer analogy, I warn against overliteral interpretations. That the brain can usefully be treated as a control system does not entail that it literally is a "controller" interacting with a "plant." Again, the danger with ignoring the analogical nature of the neuroscientist's borrowing is that it leads to ignorance of disanalogies between brains and controllers; but once the analogical interpretation is in place, it blocks the realist inference that the terms employed in control theory, especially the notion of the representation of the plant, must actually have their counterpart in the brain.

These realisms were all of the robust variety, arguing that the success of the science entails ontological commitment to neural representations that satisfy some fairly stringent conditions on being representations (e.g., specification of content and normativity). This robust realism stands in contrast to the "deflationary realism" proposed by Coehlo Mollo (2021). Like me, Coehlo Mollo argues that scientific explanations with representation-posits are grounded in an analogy with representation artifacts ("public representations"), and therefore, we should expect disanalogies and not hold the scientists' posits to the definitional standards of public representations. For the cases discussed in that paper, such as the positing of neural or cognitive maps, a charitable and epistemologically satisfying account is offered. However, this does not provide the resources to underwrite the hard cases, where representations are attributed to sensory neurons. On the metaphysics, Coehlo Mollo is less committed than the usual realist because he does not stake his case on there being entities with full-blown semantic properties in the mind or brain. However, there is a metaphysical picture in play, one in

which the ultimate targets of explanatory models are "real causal patterns" (Potochnik 2017; and also see chapter 5 of this book). He writes: "Deflationary realism holds that the cognitive representational model is an idealised model: a partially distorted and simplified picture of the causal features that contribute to bringing about the characteristic patterns of behaviour that it aims to explain, namely representational patterns" (2021, 18). What this comes to is a claim that the relationships and patterns referred to in "cognitive representational models" are in fact nonsemantic, causal ones, but ones that are best brought to the fore by viewing them through an idealizing lens, whereby they are depicted as intentional relations.

### 6.3.2 Antirealisms

Since there are a range of views left open, once the reality of neural representations has been denied, the antirealists are a disparate bunch, ranging from eliminativists, to fictionalists, to polite deflationists. I will here offer some comments on each of these "-isms" in turn.

I take Ramsey (2007) as my representative eliminativist (at least regarding the receptor notion) since he gives the most sustained attack on the representational posits of sensory cognitive neuroscience.[28] His main argument is that such posits do not satisfy the "job description" of representations, properly speaking, and the sensory responses would be more accurately characterized as the end point of a nonintentional cause-and-effect chain, like a series of switches triggered in sequence.[29] As argued previously, this claim about the availability of causal explanations expresses a generalized metaphysical view about the nature of the relationships under discussion, and furthermore, it neglects to consider the actual applicability of the proximality principle to the systems investigated in cognitive neuroscience.[30] Needless to say, elimi-

28. See also Hutto and Myin (2014).

29. A further argument is that representation-talk is harmful, getting in the way of the conceptual development of sensory neuroscience (Ramsey 2007, 147). However, he does not take into account the cost of neuroscientists having to abandon the simplifying strategy provided by the positing of these representations.

30. Ramsey (2007, 142–143) nicely conveys how the causal-mechanical explanation gets stuck with the proximality principle:

> Sensory receptors are functionally similar to protein receptors on cell membranes. When the mechanics of cell membrane protein receptors are fully articulated, few people are inclined to claim that protein receptors actually serve as representations. Instead, they are seen as

nativism scores badly on my first desideratum since it is the most revisionary (and therefore uncharitable) interpretation of the scientific practice. It does not perceive any epistemic benefits in the positing of representations in sensory neuroscience, and it takes a strong metaphysical stance of denying all but nonintentional causal relationships in those systems.

The other versions of antirealism are not revisionary—they do not recommend that neuroscientists dispense with intentional talk, but they do insist that those terms not be understood realistically, as referring to actual representations in the brain; and it is because of the utility of the representation language that it is granted this stay of execution. Two neuroscientists, Kriegeskorte and Diedrichsen, nicely spell out the case for this kind of approach. They write: "We could avoid representational interpretations altogether and approach the brain as a dynamical system . . . The dynamical systems perspective is fundamental (in that it captures what the brain does at the level of physical mechanism) and complete (in that it should be able to account for all aspects of brain function)" (2019, 408).

However, they observe that that perspective would be impractical because of its complexity. Analogizing the brain to a computer, which like the brain is said to be fundamentally a dynamical system, they argue that a representational perspective is in practice indispensable:

> Consider the case of computers: They too can be understood as dynamical systems. However, interpreting the patterns of charges and currents as representations of data and instructions enables us to capture a computer's behavior more concisely in a high-level algorithmic description that reveals the dynamics in terms of the implemented functions. Like a computer, the brain is a dynamical system, and representational accounts can help us cope with its complexity. (2019, 409)[31]

The pragmatic justification for representation-talk, as a means of simplifying the brain, is in alignment with my account. But the approach takes

---

structures that reliably transport specific molecules (or other chemical or electrical phenomena) into the cell; they serve as a type of non-representational transducer. Similarly, when the mechanics of receptors in our sensory cognitive systems are properly understood, we see that they also play a relaying role, not a representational role.

The problem is that causal tracing of the mechanics of sensory receptors will never find their way out to the distal dependencies central to cognitive neuroscience.

31. See Roskies (2021) for an interesting philosophical discussion of the use of functional magnetic resonance imaging (fMRI) and data analysis to chart representations as promoted by these authors.

the further metaphysical step of endorsing realism with respect to the physics perspective on the brain (as a dynamical system or "physical mechanism") and antirealism regarding the intentional perspective. For my part, I opt to be agnostic on this point. This difference is apparent in the two philosophical articulations of the approach that I will examine.

Neural representation fictionalism (NRF)—articulated though not unequivocally endorsed by Mark Sprevak—proposes that

> neural-representation talk in cognitive science is perfectly in order and cannot, and should not, be eliminated or paraphrased away from serious fact-stating language. However, neural-representation talk does not bring with it any commitment to the existence of neural representations since it is understood as systematically false. Talking about neural representations is a useful device for cognitive science, but no more ontologically committing than talking about water as a *continuous incompressible fluid* is in fluid dynamics. (2013, 548)

The comparison with fluid dynamics brings our attention to the point that talk of neural representation is to be understood as a certain kind of scientific idealization. This would account for the epistemic benefits of this practice, and NRF is fairly charitable to current scientific practice since it does not charge it with a gross methodological failure (in the way that eliminativism does), although it leaves open the possibility that scientists who uphold ontological commitment to neural representations are mistaken about their objects of study.[32] As with the purpose of this chapter, one of Sprevak's motivations is to avoid the metaphysical burden taken up by the realist, that of having to naturalize intentionality. But NRF, as presented in the passage by Sprevak quoted here, is not metaphysically neutral since it denies the existence of neural representations and says that claims regarding them are, strictly speaking, false. That said, Sprevak (2013, 540) does mention that it is possible just to remain agnostic about the posits given a fictionalist treatment. That position would be quite similar to my account, in that it does no more than reject a literal interpretation of the science.

Egan's deflationary account (2020) is less straightforward to assign to our realist/antirealist classifications because it subscribes to realism about the

32. Ramsey (2020) argues that most scientists employing this framework take on this ontological commitment, unlike the noncommital stance of Kriegeskorte and Diedrichsen (2019), quoted previously.

representational vehicles and mathematical contents posited in neurocognitive models,[33] but not to the ordinary intentional content. The proposal is that "a cognitive model posits . . . representations just in case it identifies representational vehicles, via $f_R$ [the realization function], which play crucial causal roles in the exercise of the capacity, and assign these vehicles contents in $f_I$ [the interpretation function]" (2020, 40).

The critical feature is that the assignment of contents is pragmatic, which is to say that of the indeterminate suggestions for content that come from correlations or homomorphisms between external objects and neuronal activations, the researcher selects a determinate content most relevant to their explanatory project. This attribution of content by the researcher is no more than a heuristic gloss.

Since mind-independent facts do not by themselves fix intentional contents for neural representations, Egan's counts as an antirealist view. This antirealism leads her to the view that, ultimately, explanations positing neural representations are a species of causal explanation. This is because it is the states and structures of neural systems, identified as vehicles, that provide explanations of cognitive capacities; but only the causal properties of those states and structures are relevant to this. Ultimately, this leads to the position that there are not neural representations except in the eye of the researcher: "Representations are distinguished from *mere* causal relays by the fact that they are assigned contents by the interpretation function $f_I$, but since the content assignment is confined to the heuristic gloss, it might be argued that the phenomenon of interest—representation—has indeed disappeared" (2020, 43).

This is an implication that Egan does not disavow. Regarding my three aims, Egan, like Sprevak, scores well on the first two but not the third. Although the deflationary account is explicitly not a general metaphysical theory of representation (Egan 2020, 43), it does make the unequivocal metaphysical claim that relationships between neural activations and distal objects are *not* intentional ones, but they *are* causal ones.

33. See Egan (2017). Egan is a literalist about neural-computations, as mentioned in chapter 4.

## 6.4 Implications

If you have been keeping track of the scores, you will have noticed that there are rival views that I admit rate highly on the first two aims of charity and accounting for the epistemic benefits of representation-talk, but none that achieve the third, metaphysical neutrality. On this system of point scoring, my account is the winner; but this result, rather than convincing readers of the superiority of my proposal, may instead leave them with suspicions and doubts about the third aim. What does it mean to state that an interpretation of neuroscientific practice is metaphysically neutral, and why should neutrality be desired? In this final section, I say more about the nature of this third constraint. We will see that it is the adherence to a mainstream naturalistic picture, originating in the philosophy of mind, that prevents the rival theories achieving metaphysical neutrality; it turns out, furthermore, that this picture is not consistent with the naturalistic understanding of causation developed in the philosophy of science.

In operation, metaphysical neutrality means that in the philosophical interpretation of the neuroscientist's employment of controversial terms such as representation and causation, one foregrounds the epistemic and pragmatic features of these terms (their roles in empirical investigation and explanation), putting aside metaphysical worries about the grounds of these notions. A good model for this foregrounding of the epistemic and pragmatic, and the relegation of the question of metaphysical underpinnings, is provided by Woodward's (2014) *functional* approach to causation. His interventionism treats "causal cognition" as an "epistemic tool" and analyzes causal concepts and patterns of reasoning in terms of how well they serve the scientist's goals and purposes (2014, 694). It eschews the metaphysical project of showing how causal terms reduce to noncausal ones, thereby fitting causation into a more fundamental world picture. Similarly, my account of representation posits in neuroscience has focused on how they aid the scientist's project of simplifying the brain—the epistemic task that is the subject of this book. Similar to Woodward, I eschew the metaphysical project of showing how intentional terms are grounded in nonintentional ones. The same approach was taken in chapter 4, where I accounted for the epistemic benefits of positing neural computations while showing how this practice is not dependent on there being a philosophical consensus

on the tricky problem of implementation—how to state the objective conditions for a physical system being an implementation of a computation while avoiding pan-computationalism. As stated at the start of this chapter, metaphysical neutrality allows a more charitable interpretation of the scientific practice because it denies that the justification or motivation for the practice stands in need of support from a philosophical theory of the controversial terms, hitching the fortunes of the science to the outcomes of long-running debates in philosophy (cf. Rescorla 2013, 693).

The mainstream naturalistic picture that comes from the philosophy of mind and is inherited in most discussions of representation in the philosophy of neuroscience is very much one that puts metaphysics in the foreground. Its primary ontological assumption is that fundamental reality is constituted by physical entities, processes, and properties. These in turn make up a causal nexus, which is to say that reality has a "causal structure" revealed in true causal explanations.[34] The ontological status of intentional properties, entities, and relationships is taken to be questionable in comparison to the presumed realism with respect to physical and causal ones. To claim a place for the intentional within mind-independent reality, it must be shown how intentional properties, entities, or relationships are related to nonintentional ones by some respectable metaphysical dependency relationship such as supervenience. This is the problem of location (Jackson 1998, chapter 1) or placement (Price 2011, chapter 1). More specifically, the challenge, taken up many times, is to *naturalize* intentional relationships by showing how they are grounded in specific kinds of causal relationships.[35] Both the realists and the antirealists discussed in section 6.3 ask whether intentional content supervenes on causation. The realists answer yes and the antirealists answer no, while both subscribe to a metaphysical picture in which causal relationships are certainly present in mind-independent reality.

34. See Craver (2013, 144–145) on the notion of causal structure and nexus, drawn from Salmon (1984). While the talk of "causal nexus" is not ubiquitous (it occurs, for instance, in Egan 2020, 42: "content captures a salient part of the causal nexus in which the state is embedded"), it is just one way of stating the orthodox metaphysical idea of causal realism—"the doctrine that causation is a feature of objective or mind-independent reality" (Menzies 2007, 191).

35. See, for instance, Egan (2019) for a review of attempts in this area.

Not only is this mainstream picture problematic because it makes the respectability of scientific practice contingent on the success of the philosophers' project of naturalizing intentionality; its claims about the relationship between physical reality and causal relationships are dubious from a more broad naturalistic perspective that considers what contemporary physics has to say about these matters. As Price and Corry (2007, 2) write, the "issue of the place of causation in the constitution of the kind of reality revealed to us by physics remains both highly problematic and highly important." Since Russell (1913), causal realism has been condemned for its alleged incompatibility with modern physics.[36] Yet some kind of causal realism is required for the notion of there being a set of causal relationships that are always more basic than intentional relationships. Maybe the idea is that this causal nexus is found not at the level of fundamental physics. But the naturalistic standing of such proposals, which are in any case very vague—more of an intuition or a worldview—can also be queried.

In this chapter, we have seen that a key metaphysical assumption accepted by realists like Shea and antirealists like Ramsey and Egan is that a causal explanation of the relationship between a distal object and particular neural activation is always available in principle as an alternative to one referring to intentional notions. But given that the complexity of the actual neurophysiological situations means that the replacement causal explanations are not practically achievable, the thought must be that there are always *ontic explanations* available; that is, explanations existing in nature independent of human activity and articulation (Craver 2014). This is another strong metaphysical commitment that my account dispenses with. Within philosophy of science, the dominant account of causation is interventionism, due to its good fit to scientific practice, especially in the special sciences. Given his emphasis on human cognition and the practice of giving causal explanation (Woodward 2014, 693), Woodward's interventionism, arguably, does not allow the supposition of (human-independent) ontic causal explanations. Hence, it would not support the view that explanations positing intentional contents can always, in principle, be replaced by ordinary causal explanations. A more general, and perhaps more important, point is

36. See, for instance, Norton (2007).

that Woodward's account leaves it open that there are modes of inquiry into classes of phenomena for which causal notions are inapplicable.[37] It may be that the robust, distally directed relationships that cognitive neuroscientists treat as intentional ones, and whose characterization as such is taken to be explanatorily fundamental, are one such class of phenomena, at least within that mode of inquiry. On this alternative to mainstream naturalism, there is no pressure to naturalize the intentional relationships by grounding them in causal ones because the metaphysical assumptions that motivate this kind of project are absent.

To properly illustrate where we have arrived at the end of this chapter, it is helpful to contrast it with Dennett's picture of the three stances that can be taken with regard to any target of investigation—the physical, design, and intentional stance.[38] The reader may have been reminded of Dennett's picture, in my account of explanatory pluralism, whereby the molecular neurobiologist provides proximate causal explanations of an FFA neuron firing, the cognitive neuroscientist refers to an intentional relationship with a distal object, and an evolutionary biologist could refer to adaptive pressures. The major point of difference is that Dennett starts from a metaphysical assumption about what minds fundamentally are—that they are complicated physical systems (Dennett 1988, 495)—and then aims to show under what conditions the positing of representations can add predictive and explanatory value. In contrast, my approach does not begin with a prejudice against the intentional stance as being less fundamental than the physical one, as this would be incompatible with my constraint of metaphysical neutrality. Dennett's three stances form a hierarchy, whereas my perspectives are on a level footing. The brain is a highly complex organ that, like other material objects, can be treated as an ordinary physical object open to causal explanation, but it can also be treated as an intentional system and made subject to explanations positing neural representations.

37. Woodward (2014, 702) writes that "from the point of view of a functional approach to causation, it is entirely possible that there may be some contexts or domains of inquiry in which causal thinking and representation, or at least the kind of causal thinking associated with interventionism, are not useful or functional."

38. See Dennett (1981/1997), and also Lee and Dewhurst (2021), who introduce a "mechanistic stance."

Rather than Dennett's stances, a closer precedent to the pluralism arrived at with my account is in the proposal of Wilhelm Dilthey (1894/2010) that explanation in psychology either can be conducted in line with the pattern of causal explanation exemplified by the physical sciences or take a sui generis approach. Dilthey himself was a neo-Kantian (broadly speaking), and it is not coincidental that the egalitarian pluralism that I propose requires the rejection of causal realism—this rejection being a feature of the Kantian tradition of philosophy of science (Price and Corry 2007, 9). Metaphysical neutrality also requires that I not stake a claim on whether there really *are* neural representations, independently of scientists modeling the brain in this way. An arresting implication of metaphysical neutrality is that no body of neuroscientific research can inform us of how the brain is "in itself." This is the conclusion to be explored in chapter 7, in which we consider the pluralism of modeling perspectives within neuroscience of the motor cortex.

# 7 The Heraclitean Brain

> Die Natur ist nur einmal da. Nur unser schematisches Nachbilden erzeugt gleiche Fälle.
>
> —Ernst Mach (1910, 230)[1]

> Socrates: Then how can that be a real thing which is never in the same state? . . . Nor yet can they be known by anyone, for at the moment that the observer approaches, then they become other and of another nature, so that you cannot get any further in knowing their nature or state, for you cannot know that which has no state.
>
> —Plato (1961)

Change is the ever-present, constant thing in the appearance of the natural world. The seasons, the tides, the clouds, maturation and decay—all incessantly cycling through in patterns that never exactly repeat. The world that we inhabit is a flowing world in which nothing is permanent, time erodes everything, and gives birth to all new things. Yet, in the *Cratylus*, we have a dismissal of the Heraclitean philosophy of flux, opposing the real and permanent to the changeable and merely apparent. What is more, changeable things without any fixed state could not be known. Knowledge, properly speaking, is of the forms, which are beyond the reaches of time, more like the celestial bodies that shine more constantly and revolve with more perfect periodicity than anything in the sublunary realm.

The epistemological projects of modern, exact science, no less than those of the Platonic philosophy, need fixed targets. Mathematical physics (not

1. From the 1882 lecture, "The Economical Nature of Physical Enquiry": "Nature is but *once* there. Only our reflexion produces equal cases." (Translation from Banks 2004, 29; emphasis in original.)

coincidentally a discipline that originated in astronomy) sets itself against the changeableness of worldly events through the inscription of eternal laws that predict those occurrences. The reality claimed to be discovered by means of such inquiries is knowable by virtue of its stability. With these tools in hand, nature is rendered predictable, and hence controllable, at least in principle. Attendant to this approach is a lack of intererst in the particularity of things, a mindset that Cassirer (1929/1957, 409) contrasts with that of the historian, for whom all events are essentially unique:

> Even where the physicist describes a single event, confined to a definite situation in space and moment in time, he is not concerned with the particular as such, but considers it under the aspect of its repeatability.

A message of this chapter will be that mathematical neuroscience, just as much as physics, has this drive toward the framing of events as repeatable, and hence as the reflection of objects of knowledge that are essentially stable. This is yet another aspect of the need to make things simple which, as this book argues, has profoundly shaped neuroscience. Temporal heterogeneity, as was noted in chapter 1, is one way in which the brain is exceedingly complex; abstracting away from its instability is one way to simplify the brain.

The importance of this methodology can be appreciated without commitment to the ontological picture of real stability underlying apparent changeableness. Mach, for instance, was skeptical that exactly repeating events ever occurred in nature, but this did not undermine the power and utility of physical laws as schematic, economical representations of that flow of events (Banks 2004, 24–25). For Henri Bergson, there was a general point here about the scientific intellect, which is that in its representations, especially of living beings (we might say "models" where he says "signs"), we have freeze-frames of a mobile actuality. The ultimate purpose of this substitution is pragmatic because the application of this fixative brings objects more easily under control: "Signs are made to dispense with this effort by substituting, for the moving continuity of things, an artificial reconstruction which is its equivalent in practice and has the advantage of being easily handled. . . . What is the essential object of science? It is to enlarge our influence over things" (Bergson 1907/1944, 358).

In this chapter, I will argue that the brain is Heraclitean in its never-exactly-repeating richness. Although the Heraclitean character of the brain is there to be observed, it is not theorized as such, since theoretical and modeling

approaches are founded on assumptions of there being some level at which stability is present in the brain, there to be uncovered. Section 7.1 will review recent observations of how neural structures and properties once thought to be stable—including ones underlying long-term memories—turn out to be quite labile. The case study presented in section 7.2 describes a long running controversy over the motor cortex. I will show how approaches to meeting the challenge of this temporal complexity have structured mathematical frameworks for modeling the functionality of motor cortex. Embedded in each modeling perspective are various assumptions about how best to simplify the brain, and in section 7.3, I present some more general implications for understanding perspectives in the philosophy of science as strategies for simplification. Given that no perspective can encompass the brain in its full Heraclitean complexity, does it follow that there are nonperspectival truths about the motor cortex (and by extension, the primate brain) that are off-limits to neuroscience? In section 7.4, I voice support for this conclusion and advocate for a Kant-inspired (as opposed to a Platonic) assessment of the role of mathematical abstraction in science.

## 7.1 The Ever-Changing Brain

An article published recently in the *Atlantic* has a title that sounds sarcastic (though not, I think, intended as such): "Neuroscientists Have Discovered a Phenomenon That They Can't Explain" (Yong 2021). The piece is all about observations of *representational drift*, the tendency of the neural activations correlated with task-related stimuli, actions, and cognitive variables (Rule, O'Leary, and Harvey 2019, 141)—which we saw in chapter 6 are ordinarily conceived as neural representations—to change their distal targets in the course of time. For example, the centerpiece of the *Atlantic* article is a publication by Schoonover et al. (2021) on the mouse olfactory cortex. The group recorded the activity of well-isolated neurons over a period of thirty-two days. During the experimental sessions, mice were presented with a panel of four or eight different odor stimuli, seven times a day, every eight days. While the recordings taken within one day showed neurons to have a consistent response to repeated presentations of the same odorant, measurement eight days later would reveal a departure from the stimulus-evoked response that had occurred previously. Since a strong response to an odorant is interpreted as the neuron representing that particular smell, the finding that neurons'

stimulus preferences change in the course of days and weeks is taken as an indication that the representations of these odors are drifting around, lacking stable associations between certain neurons and particular odorants.

Representational drift has also been reported in mouse visual cortex (Deitch, Rubin, and Ziv 2021), in mouse posterior parietal cortex, where neurons represent stimulus-action pairs during a spatial navigation task (Driscoll et al. 2017), and with the place cells representing the animal's location in mouse hippocampus (Ziv et al. 2013).[2] Although such results are not particularly surprising for many neuroscientists, they do contrast with the picture of sensory cortex neurons being "fixed filters," having receptive fields and trigger features that are stable following the critical periods of early life. An instance of that idea is the standard model of the primary visual cortex discussed in section 5.1 of chapter 5. Likewise, areas of the brain involved with navigation and motor coordination have also been assumed to be fairly fixed in terms of the neural activations associated with distal landmarks and targets.[3]

One reason to think that the brain would be quite stable in adulthood is that the population of neurons, unlike the cells of other organs such as the liver, does not turn over, and we rely on the same set of neurons throughout life.[4] Neuroplasticity, of course, is associated with learning, and on the picture of a relatively stable brain, this would be due just to changes in connection strength between neurons, the upregulation and downregulation of

2. In this case, the place cells retained their preferences for particular locations, but only 15–25 percent of the cells involved with place representation on one day would be active in the next recording session. See Rule et al. (2019) for a review of studies on representational drift. Liberti et al. (2022) and Sadeh and Clopath (2022) argue that much of the drift in responses can be accounted for as modulation due to changing behavioral patterns and attentional states in the animals.

3. There is an interesting comparison to be made with modern electronic computers, where, for reasons of efficiency, the physical tokens associated with a symbol type are constantly being updated (Sprevak 2019, section 3); arguably, the same efficiency gains from representation remapping are to be expected in neural systems. However, the increased difficulty of reverse engineering a free-floating representational scheme may be what pushed neuroscientists toward the optimistic working assumption that representations are fixed.

4. See Chambers and Rumpel (2017). Of course, some neurons do die and adult neurogenesis does occur in certain brain areas—something once thought not to happen at all (Snyder 2019). Still, the point holds that there is not the wholesale replacement of neurons, equivalent to the turnover in cells in other parts of the body.

synapses through *long-term potentiation* (LTP) and *long-term depression* (LTD).[5] Memories would therefore persist through the persistence of a cell assembly, a set of neurons connected together in a particular way. This more static picture stands against an emerging view of the brain as dynamic and constantly undergoing reconstitution and reorganization at subcellular, synaptic, and population levels. Surprisingly enough, the synaptic connections between neurons, including ones underlying memories that can last almost a lifetime, are constituted by dendritic spines and axonal boutons that come in and out of existence, and change in size, over various timescales ranging from days to years (for a review, see Chambers and Rumpel 2017, 173–175). While turnover rates for these structures vary depending on brain area and cell type, it cannot be the case that memories have the persistence that they do because of the changelessness of the physical medium in which they are inscribed. Our fading memories are less like etchings in stone, ground down after years of slow weathering, than transient patterns that produce copies of themselves, with fidelity eventually lost as distance increases from the original.

The trend toward more attention to the instability of the brain is probably in large part due to the more recent availability of methods for long-term monitoring of subcellular structures and for recording activity of the same individual neurons across extended time periods, which is itself a significant technical challenge. When neurons are viewed through one small temporal window, it is easy enough to assume that this snapshot represents a steady state of the system. But when observations occur across a greater stretch of time, the evidence for reorganization is available and undeniable. It is interesting that these findings of instability are often offset by those of robustness in other neurophysiological parameters,[6] and an overall stability

5. As Trachtenberg et al. (2002, 788) observe, additional ways for connectivity to change are through the generation and elimination of synapses (i.e., sprouting and removal of dendritic spines), and through new growth of axonal and dendritic processes. In this study, they observed the former, but not the latter, in adult mouse somatosensory cortex.

6. The point here is that there can be consistency of function (e.g., the role played by a neuron within a particular circuit), in spite of underlying fluctuations in cell components and their activities (Marder, O'Leary, and Shruti 2014). And see Chirimuuta (2017a) on how there is a need for different accounts of robustness in organisms, as opposed to machines, given these dynamic characteristics.

of behavior that belies this underlying neural instability. As will be discussed in section 7.2.2, there are various findings of population profiles being more stable than single neuron properties, which may account for the behavioral consistency (Montijn et al. 2016).

An implication of these various observations is that the brain is an organ that maintains itself and orchestrates behavior only by undergoing continual change. That is, it is more of a process than a static thing—a Heraclitean object. John Dupré (2012) has argued that all living organisms should be characterized in this way, as processes rather than entities. Indeed, the changeability of the brain is a direct consequence of being made of living tissue. Neurons, like other cells, are always changing their makeup as they metabolize: "Each neuron is constantly rebuilding itself from its constituent proteins, using all of the molecular and biochemical machinery of the cell" (Marder and Goaillard 2006, 563). In addition, Peter Godfrey-Smith (2016a) proposes that the Heraclitean nature of biological cells has important implications for how we understand cognition. The essential plasticity of neural tissue was an important difference between brains and computers, as discussed in section 4.4.2 in chapter 4. The idea is that the inherent changeability of biological tissue was leveraged during the evolution of the nervous system as a means for learning and coping with the challenge of staying alive in an unstable world, and remains an important factor behind the functionality of the brain. This, Godfrey-Smith argues, puts limits on the functional equivalences between brains and computers.

Not only is changeability a pervasive feature of living cells, but many of these changes can be considered to be adaptations to external circumstances. It is characteristic of living cells, including single-celled organisms, for past occurrences to alter activity going forward, and these forms of plasticity are related to epigenetics, modulation of the cell's genetic readout due to events in the organism's lifetime. Bateson and Gluckman (2011, 43) write, "The central elements underlying many forms of plasticity are epigenetic processes, and plasticity operating at different levels of organization often represents different descriptions of the same process. Underlying behavioral plasticity is neural plasticity and underlying that is the molecular plasticity involving epigenetic mechanisms."

Without getting into the semantics of whether a single-celled organism can have "memory," the important point is that memory, as we understand the concept from human and animal psychology, is at the very least

continuous with and dependent on the existence of such basic capacities for adaptation.[7]

I will conclude here by reiterating the point that exact science needs fixed targets. Given the observations reviewed in this section, it would seem that making a fixed target of the ever-changing brain would be like pinning jelly to a wall—except that scientists have ways of devising fixatives. Chapter 5 described concrete experimental methods for making neural responses less variable than they otherwise would be. With modeling, things can be crystalized even further. Such strategies will now be discussed in relation to research on the motor cortex. This field of neuroscience is a particularly apt because there has been a long-standing and often heated dispute over what the function of the motor cortex actually is, leading the neuroscientists themselves to be explicit in stating and arguing for their different theoretical perspectives. Furthermore, neurophysiologists recording from this area have long been aware of the instability of the single neural correlates of bodily movements, and this has been a source of the difficulty on reaching consensus on the functional interpretation of motor cortex (Rokni et al. 2007; Scott 2008).

## 7.2 Perspectives on the Motor Cortex

Motor cortex holds a prominent position in the history of neurophysiology, being the first area from which it was found possible to elicit behavioral effects—muscle twitches and fragments of recognizable bodily movements—from the application of electrical current to the surface of the brain.[8] It is also

7. Ginsburg and Jablonka (2019, 462) write:

> Neurobiologists have realized that they need to consider not only synaptic memory but also the epigenetic memory embedded in the cell nucleus and the transmission of 'memory molecules' between neurons (and other cells); these memory molecules include small RNAs, which can alter both the nuclear epigenetic memory within neurons and the synaptic connections between neurons. Thus, there are additional biochemical memory systems in the nervous system, not just the well-studied synaptic one, and although these mechanisms interact, they can be semi-independent, with nonlinear interactions. If we want to understand learning and memory in neural organisms, we need to consider all these mechanisms. Crucially, the epigenetic mechanisms of cell memory are very ancient; they preceded the evolution of neurons and are found in all living organisms.

Also, see Gershman et al. (2021) on the evidence for Pavlovian conditioning in the single-celled paramecium.

8. See Fritsch and Hitzig (1870) and Ferrier (1873); also see Young (1990).

one of the first regions to be theorized in terms of neural representations and cortical maps (Chirimuuta 2019). Yet in spite of it being one of the earliest brain areas to be subject to electrophysiological investigation, and to receive a recognizably modern theory of its function in terms of it representing the body and its movements, the motor cortex is also one of the most controversial parts of the brain, subject to continuous debate over the most basic questions about its functional role as a center crucial for the execution of movements.[9]

Control of muscles, and therefore movement, was perhaps the original driver of the evolution of nervous systems (Keijzer 2015). The body's motion occurs in environments that are themselves moving and reconfiguring, which means that motor control must be flexible and adaptive. The brain's signaling to the muscular system cannot be a rigid set of commands; it must incorporate some Heraclitean characteristics of the surroundings in which the body operates, and will be highly integrated with the sensing of the body and of the effects of the body's movements on the things around it. And yet theoretical neuroscience, in work that aims at mathematical models of the workings of the motor cortex, has been driven to abstract away from the inherent dynamism of this brain-body system. In this section, we will examine two traditions of research on motor cortex (one older, the other more recent), which posit some stable representations or command patterns, either at the single neuron or at the population level. Even though the newer approach uses the tools of dynamical systems theory (DST) and would seem by this fact to be attempting to represent the dynamism of the brain, it should still be interpreted as the attempt to find some stability, and hence simplicity, underlying the changing appearances.

Cunningham and Yu (2014, 1507) compare these two perspectives in terms of their strategies for simplifying the brain:

> One of the major pursuits of science is to explain complex phenomena in simple terms. Systems neuroscience is no exception, and decades of research have attempted to find simplicity at the level of individual neurons. Standard analysis procedures

9. In what follows, most of the findings concern *primary motor cortex* (M1) of the primate. Note that the motor cortex contains additional regions (*supplementary motor cortex* and *premotor cortex*), and that the role of the motor cortex is thought to be different in other mammalian groups, such as rodents.

include constructing simple parametric tuning curves and response fields, analyzing only a select subset of the recorded neurons, and creating population averages by averaging across neurons and trials. Recently, studies have begun to embrace single-neuron heterogeneity and seek simplicity at the level of the population as enabled by dimensionality reduction.

I will now say more about these contrasting approaches and methods.

### 7.2.1 The Neuron Stability Perspective

The first perspective on the motor cortex is comparable to the fixed-filters model of the visual system.[10] The core assumption is that individual neurons represent or encode some parameters relevant to movements in specific body parts—these may be individual muscle contractions, sequences of muscle activations, or higher-order parameters such as the velocity of an arm movement. Neurophysiological experiments seek correlations between neural activations and the precise movement parameters, although in practice different kinds of parameters (e.g., direction and velocity) tend to be correlated with one another. Therefore, one of the major difficulties for this approach has been that neurophysiological recordings have yielded partial evidence for various hypotheses about what the motor cortex encodes, leading to a proliferation of versions of this perspective, without consensus.[11]

Within this tradition, trial-to-trial variability in neuronal responses is classified as noise rather than as variance to be modeled and explained. This is in part because of the assumption that the neurons' tuning properties are fixed, so variability in responses is not coding anything, and also because the trial-to-trial variability is beyond the scope of the experimenter's interests and ability to seek explanation. However, the fact that there is a wide variety of activity patterns across neurons, as well as instability of movement parameter encoding of a single neuron across experimental conditions, has been a continual bugbear of this approach (Gallego et al. 2018, 2). The individual

10. That is, the idea that individual neurons in the visual system selectively respond to a particular kind of stimulus or "trigger feature" (e.g., a small edge of a certain width and orientation), and these tunings are stable across time—independent of task and stimulus context (see section 5.1.3 in chapter 5).

11. "Most theories of M1 function over the past 50 years have focused on different time-independent movement parameters such as force, direction, velocity, position, speed, acceleration, or combinations of these that might be encoded in the firing rates of individual M1 neurons" (Omrani et al. 2017, 1832).

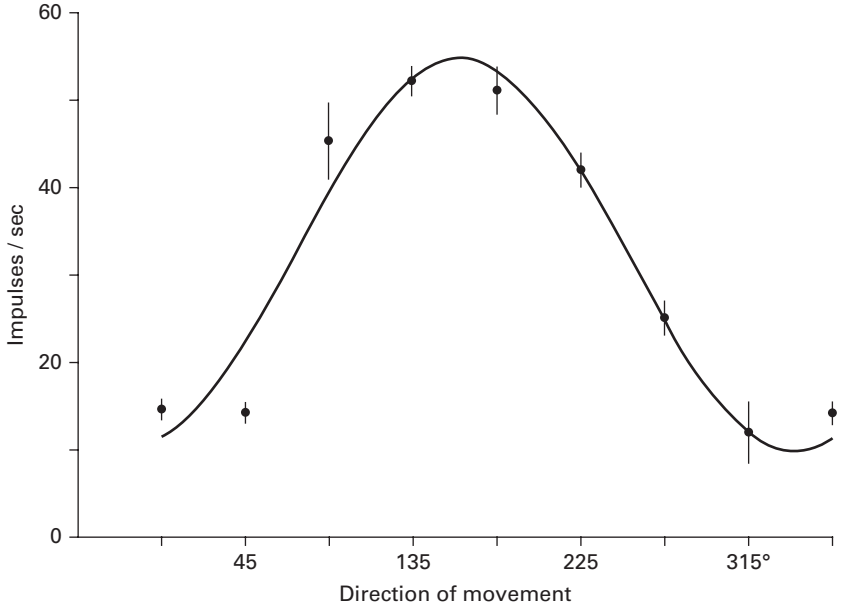


**Figure 7.1**
The cosine tuning model treats each motor cortical neuron controlling arm movement as firing maximally at its preferred direction of movement, with firing rate dropping away as a cosine function for nonpreferred directions. See Georgopoulos et al. (1982) and So et al. (2012).

neuronal responses are just not as stable and homogenous as demanded by the hypothesis that they encode specific movement parameters.

It is important to appreciate that there are both single-neuron coding and population code versions of the neuron stability perspective. For example, Evarts (1968) recorded the activity of *pyramidal tract neurons* of the motor cortex of head restrained monkeys performing voluntary hand movements for a juice reward. The responses of individual neurons were interpreted as encoding force exerted by muscles in the wrist rather than the displacement of the hand through space. An influential version of the population approach came with the *population vector* model of Georgopoulos et al. (1986). The idea is that each single neuron has a preferred direction of movement for a particular body part, which can be discovered through neurophysiological observation of the firing rate associated with, say, an arm moved laterally at a range of angles (see figure 7.1). At the same time, the actual movement undertaken

by the arm is determined by the combined firing rates of a population of neurons. The population vector, which gives the angle of movement to be made by the arm, is the sum of each neuron's preferred direction, weighted by the strength of its firing rate.

The population vector model, with its assumption that motor cortical neurons code for direction of intended movements, has found a practical application in brain-computer interface (BCI) technologies, that record from about 100 neurons in this brain region and employ decoding algorithms on the data to derive signals for controlling a robotic limb or cursor on a screen (Velliste et al. 2008). However, it does not follow from the fact that direction of movement can be decoded from M1 data, using the population vector model, that the assumptions of this model are true. Certain assumptions made by the decoding algorithms have been shown to be false with respect to neurophysiology of the motor cortex, but during the BCI experiments, the brain adapts to biases introduced by the models (Koyama et al. 2010; Chirimuuta 2013).

While there are both single-neuron and population-code versions of the neuron stability perspective, the core assumption that covers them all is that individual neuronal activity says something meaningful about the movement that is being or is about to be performed. In other words, by itself, it represents some property of the movement undertaken—the part of the body involved, its angle, force, or speed. The newer, alternative perspective drops this assumption and insists that single-neuron activations are not inherently meaningful. Instead, it is only in populations of cells within motor cortex that control signals and movement representations will be found. I call this the *population pattern approach* because its working hypothesis is that stability and regularity—regarding the relationship between neural activity and bodily movement—will be discovered only at the population level. This is how Gallego et al. (2018, 2) describe the difference between the old and newer approaches:

> An intriguing alternative [to the traditional approach] is that the computations involved in generating movement are not based on the independent modulation of single neurons, but rather performed at the population level by networks of interconnected neurons whose coordinated activity commands the muscles that cause the behavior. In this view, correlates between single neuron activity and behavior are epiphenomenal, and yield only a limited and distorted view of the causal relation between M1 activity and behavior.

With this, they dismiss the viability of the theories and models of M1 based on single-neuron recordings—these observations could give only a distorted view, and the theoretical houses would only be built on sand. Thus, it is a feature of the newer perspective that single neurons lose their privileged status. Even though activations of individual neurons will still be recorded, the difference is that no attempt will be made to interpret individual neuronal responses in the way shown, for example, in figure 7.1. Indeed, it is a prediction of the population pattern perspective that the firing patterns of many of the individual neurons will not be interpretable in terms of external parameters (Omrani et al. 2017, 1835).

### 7.2.2 The Population Pattern Perspective

The shift to the population pattern perspective on M1, as with the big data ethological approach to primary visual cortex (V1; see section 5.2 in chapter 5), has been fostered by the invention of methods for large-scale population recordings and statistical techniques for analyzing the massive data sets produced by such experiments. Whereas the empirical support for the neuron stability perspective comes in the form of single-neuron tuning curves for movement parameters, the new view has relied on neural population data, processed to show low-dimensional structure. These kinds of data analyses have become common elsewhere in neuroscience with the increase in number of neurons simultaneously recorded. If the activity of 100 neurons is observed during one experimental trial (e.g., an arm reach), the resulting data set has 100 dimensions (one neuron per dimension). But given the correlations between neurons' activation patterns, dimensionality reduction techniques such as principal component analysis (PCA) or factor analysis can typically fit the data into about a ten-dimensional space.[12]

In addition to dimensionality reduction methods, the framework of DST is brought in for further analysis of the role of population activity in movement coordination. DST originated with a study of celestial mechanics by physicist Henri Poincaré (1890). In its short history, it has been applied very widely to describe the evolution of complex systems, also known as *chaotic systems* (Mitchell 2009a, chapter 2). An application within biology, well known to

12. Cunningham and Yu (2014) is a clear exposition and review of such methods.

philosophers of science (e.g., Weisberg 2013), is the Lotka-Volterra model of the coupled, cyclical growth and decline of predator and prey populations.

Even though DST is the theory of changing systems, ones undergoing some alteration (depicted as a journey through their state space), the whole point of the analysis in a certain sense is to reduce change to stasis, for the trajectory that the dynamical system takes through its state space is attributed to there being some fixed parameters of the system and a changeless set of laws governing it. This is a point made by Walsh (2015, 212), regarding what he calls "object theories":

> In an object theory, the domain of interest is a set of objects. The goal of the theory is to describe and explain the dynamics of these objects. So, we set out a space of possible alternatives for those objects—a state space—and we look for principles that might account for various possible trajectories through the state space. The objects in the domain are subject to forces, laws and initial conditions. . . . [W]e describe the dynamics of a system by answering two simple questions: "(i) what are the possible configurations of the system? and (ii) What are the forces that the system is subject to in each configuration?". . . . The principles that govern the dynamics of the objects in the theory's domain are not part of the domain itself. They do not evolve as the system under study does. The laws of nature, and the space of possibilities through which the objects move remain constant as the objects change. In this way, we can explain the change state of the system under study by appeal to the unchanging laws.

As with the orbit of a planet, the dynamics of motor cortex are represented by plotting the activity of the neural population as a trajectory through a low-dimensional state space. To the extent that the equations describing the evolution of the systems are taken to be fixed, what this analysis achieves is the transformation of the restless motion of the planets, or the motor cortex, into a snapshot, a formalism, that contains this movement all at once.[13]

Some presentations of the dynamical approach have insisted on it being antirepresentational, in line with dynamical theories of cognition that seek to dispense with the positing of mental representations (e.g., Chemero

13. Note that the equations describing the systems need not be assumed to be fixed. There could be versions of the dynamical approach to motor cortex that allow for their evolution as well. In that case, the reduction of dynamics to changelessness would not go through. I thank David Robbe for this point. See also Barack (2019, 2020) on neurodynamical explanations.

2011).[14] In such cases, one way to summarize the difference between the older, neuron stability perspective and the new one is to say that the relationship of *causation* (between neurons and bodily movements) has replaced an representational (*intentional*) relationship. While all agree that motor cortical activity is causally upstream of movement, proponents of the antirepresentational view do not give the population activity an additional gloss as representing some aspect of the intended movement; instead, they treat the cortex and muscles as coupled oscillatory systems and ask how the cortex orchestrates its sequence of oscillations (of neural population firing) such that they eventually cause an intended sequence of muscle contractions. A basic intuition here is that the oscillations in populations of cortical neurons, at different frequencies and phases, are analogous to a Fourier basis set of sine waves, from which any irregular waveform can be generated. Likewise, firing patterns in the motor cortex constitute a basis set which, when appropriately deployed, leads to the execution of the range of bodily movements.

However, as various philosophers and cognitive scientists have discussed, there is not an inherent incompatibility between DST and the positing of cognitive or neural representations.[15] So it is unsurprising that versions of the population pattern approach have emerged, also using DST, that posit that there are features of neuronal population activity that encode the parameters needed to control bodily movement. Barack and Krakauer (2021) give a strongly representational interpretation of the new perspective, which they label "Hopfieldian," in contrast with the older "Sherringtonian," single neuron–based one. They argue that the population trajectories thought to explain motor planning are "representational in our more full-blooded sense" (Barack and Krakauer 2021, 11). That is, beyond just evoking a correlation between an external movement parameter and a neural activation, these representations are said to have accuracy conditions and to be involved in offline motor planning. On Barack and Krakauer's

14. See in particular Shenoy, Sahani, and Churchland (2013) and the discussion of Kaufman in Omrani et al. (2017). The antirepresentational one is the only version of the population pattern perspective that I discussed in Chirimuuta (2020b). Since then, the other versions have become more prominent. Commentators who emphasize the antirepresentational current are Favela (2021) and Lindsay (2021, chapter 8).
15. See Clark and Toribio (1994, 422), Beer and Williams (2015), and Weinberger and Allen (2022).

interpretation, the evolution of a trajectory in the low-dimensional state space is the transformation of a representation, which is to say that it is a computation according to their preferred definition of computation. This stands in clear contrast with interpretations of dynamical models in cognitive science, which present them as an alternative to the computational theory of cognition (van Gelder 1995).

Another computational and representationalist interpretation of dynamical models of motor cortex comes from Lee Miller's laboratory. A feature of their discussion is the concept of the *neural manifold*. This is the low-dimensional portion of the full 100-dimensional neural space that M1 activity is found to reside in, even across a range of motor tasks (Sadtler et al. 2014). The basis vectors of the neural manifold are known as *neural modes,* which are "patterns of neural covariance thought to arise from the network connectivity" (Gallego et al. 2018, 2). The crucial part of their interpretation is that these neural modes capture population activity patterns that are thought to be the "fundamental computational units" of the motor cortex. As Gallego et al. (2017, 978) also write, "It is the activation of these neural modes, rather than the activity of single neurons, that provides the basic building blocks of neural dynamics and function."

What we see is that the search is still on for the elementary units of brain operations. In previous chapters, we have encountered earlier contenders like the simple sensorimotor reflex and Barlow's single neurons in visual physiology. A difference is that the newly proposed elementary neural modes can be defined only at the population level and require some sophisticated data analysis for their presence to be determined by the investigator. This is an indication that neural modes, no less than the other examples given in this book, are *ideal patterns* whose existence depends as much on the investigator and their choices in how to analyze experimental data, as on the brain activity by itself. But it is interesting that when discussing neural modes, these researchers equate the products of the simplifying procedures that they themselves have performed with hypothetical simplifications that the brain itself is carrying out. For instance, Gallego et al. (2017) report, "These studies also suggest that the constraints embodied by the neural manifolds *simplify movement generation* by providing a small number of signals that are independently controlled to achieve a desired behavior" (980, emphasis added). This is asserted even though it is everywhere acknowledged that scientists employ dimensionality-reduction methods because it is impossible for people to

visualize a 100-plus-dimensional space, and these scientists need a lower-dimensional representation of their data to reach some intuitions about the relationship between movements and neural activations.[16]

The question of whether neural modes are largely the product of scientists' efforts to simplify, or indicative of an inherent simplicity in nature (or both), relates to an old debate about the interpretation of latent variables (the products of dimensionality reduction) in data sets generated by psychometric testing. Stephen J. Gould was one scientist who warned against the reification of latent variables, arguing that the simplified structures found after processing data in this way are, by themselves, no more than mathematical abstractions and should not be taken to be causes acting in the world, absent additional evidence in support of a more loaded interpretation.[17]

A concern with reification comes to the fore if we appreciate that the dimensionality reductions involve loss of information present in the raw data. Normally, this is a price worth paying for a representation of the data that is far more interpretable in its simplified format.[18] But whatever is left out of the simplified representation should not be assumed to be insignificant. An everyday comparison from Lindsay (2021, 239–240) helps make the

16. It could be pointed out that in addition to these pragmatic reasons for performing dimensionality reduction, because of human cognitive limitations, there are some more general epistemic ones, indicated by the fact that many statistical methods, including machine learning, require dimensionality reduction. (I thank Mark Sprevak for this point.) My view on this is that there is not a firm distinction between the pragmatic and the epistemic. If nature is infinitely complex—which is the implication of Mach's vision of a nature that never exactly repeats—then any knowledge of it had by a finite system (be it a human or a computer performing statistical analyses) must involve abstractions such as dimensionality reduction. This knowledge serves the practical tasks of the finite knower, and should not be mistaken for knowledge of nature as it is in itself, in its infinite complexity.

17. Gould (1981) says, "The first principal component is a mathematical abstraction that can be calculated for any matrix of correlation coefficients; it is not a 'thing' with physical reality. Factorists have often fallen prey to a temptation for *reification*—for awarding *physical meaning* to all strong principal components" (250; emphasis in original).

18. Note that in the M1 population studies, the dimensions found using principal components analysis will account for a high proportion, but not all, of the variance in the original data set. For example, Gallego et al. (2018, 3) report that a 12-dimensional representation accounts for about 75 percent of the variance. It is unknown how much of the remaining variance is pure noise (e.g., experimenter introduced, instrument noise) and how much is task relevant.

point. Common personality tests reduce the spectrum of human characteristics to a few dimensions—the "Big Five" latent factors are agreeableness, neuroticism, extraversion, conscientiousness, and openness. Lindsay explains that these are derived from finding correlations between more fine-grained personality traits that normally occur together in one person, such as cleverness and quick-wittedness, and collapsing them into one trait, intelligence. The point is that the coarse-grained schema will not be able to represent the individuals for whom correlations among the fine-grained traits diverge from the usual pattern—those who are clever without being quick-witted and vice versa. However, Lindsay asserts that the loss of information with dimensionality reduction is insignificant, and the fine-grained set of concepts used to describe human personality in everyday life is an "overrepresentation." Likewise, Lindsay surmises, neuroscientists should have no compunctions about concluding that their reduced representations of neural activity are capturing all the essential information about the population.[19] In other words, Lindsay is very quick to reify the latent variables, taking them to be inherent, fundamental features of the system itself. Yet, as we find with the individuals we know, being quick-witted is different from being clever—it does not just amount to one and the same factor, intelligence—and these differences matter to us as human beings. Our more nuanced lexicon does not overrepresent human personality. It is just that in our day-to-day dealings with people, we do care about individuality. In contrast, psychologists examining personality traits en masse are likely to be indifferent to the details of individuals and will find it more convenient to ignore the unusual cases for the purposes of their research projects. Similarly, the details around the edges that get left behind after dimensionality reduction of neuronal population data may not matter at all to the neuroscientist, but this does not mean that they are irrelevant in an absolute sense. It could well be that those details matter in subtle ways to the functioning of the brain itself. This is the same basic point made

19. "Just as our folk notions of personality over-represent its dimensionality, there are many reasons why the 'true' dimensionality of a neural population is likely to be less than the number of neurons in it" (Lindsay 2021, 240–241). There is indeed good reason to think that some of the information in the full data set is redundant due to the fact that redundancy has to be built into the system so the death of one individual neuron does not lead to catastrophic failure. But from this, it does not follow that reduction to the number of dimensions interpretable to a human scientist will be one that captures the system's fundamental behavior in its entirety.

in relation to the older perspective (section 7.2.1), that neuroscientists may opt to classify unexplained variance as noise, and thus irrelevant to the cognitive task; but really, some of it is task-relevant, just beyond the scope of the neuroscientist's interests and ability to handle the system's complexity.

## 7.3 Perspectives as Modes of Simplification

> Here we are asked to think an eye where the active and interpretative powers are to be suppressed, absent, but through which seeing still becomes a seeing-something, so it is an absurdity and non-concept of eye that is demanded. There is *only* a perspective seeing, *only* a perspective "knowing."
>
> —Nietzsche (1887/1994, iii.12, 92)

In the preceding discussion of research on motor cortex, I have been using the term "perspective," pretty much interchangeably with "approach" without saying what I mean by a scientific perspective or *perspectivism*—the attendant philosophical view about the nature of scientific knowledge. There are various characterizations available in the recent literature.[20] A shared feature of the different versions of perspectivism is, as Massimi Massimi (2018c, 166) puts it, "the acknowledgement of the human vantage point (as opposed to the God's eye view) from which only knowledge of nature becomes possible for us." The prominent feature of my account of perspectivism is that I insist on simplification being the process through which knowledge of nature becomes available to scientists. It is, if you like, the scientist's especially refined sensory modality.[21] It follows that perspectives, as research traditions, should be understood as coherent collections of experimental and modeling practices, theoretical and conceptual frameworks, all working together to present an object of investigation in an appropriately simplified, and hence knowable, way. In the cases presented in this chapter, the old perspective cohered around the assumption that individual neurons encode motor parameters in a stable and interpretable way; the newer

20. See Giere (2006a, 2006b), Massimi and McCoy (2020), and Massimi (2022).
21. I do not mean to imply that other modes of knowledge do not also employ simplifications, just that this characteristic of all propositional knowledge, even simple linguistic descriptions of states of affairs where there must be abstraction from the concrete particulars, is particularly exaggerated in science. Science stands apart for the technical ingenuity and intellectual energy that it puts into its simplifications.

perspective—still somewhat under construction—makes a comparable assumption regarding patterns of population activity. Likewise, we saw in chapter 3 that the perspective of reflexology was defined by the assumption of simple reflexes being the stable, elementary units of the nervous system, and in chapter 4, that the computational approach itself offers a simplified view on the brain—one in which spike patterns are postulated to be the only events relevant to cognitive functions and are examined in isolation from the complicated background of all the other electrical and chemical activity that takes place in the brain.

It is not at all controversial that science thrives when complicated things can be made to seem simple. Various authors have made the case that complex systems, especially those studied in the biological and behavioral sciences, afford modeling from a variety of perspectives because no one set of theoretical handles or experimental practices gives the scientist access to all the relevant phenomena in the domain of interest.[22] I build on this work by emphasizing not only the way that scientific perspectives passively filter out details not relevant to their own theory and practice, but also the way that they actively impose simplifying assumptions onto the target system. The active side of this process is what *constructivist* notions of scientific knowledge often allude to, and it comes to the fore if our understanding of perspectivism begins with consideration of the distance between the complexity of the world and the simplicity needed so that scientific representations can be useful. This is how Giere (1999) puts it:

> Rather than thinking of the world as packaged in sets of objects sharing definite properties, think of it as indefinitely complex, exhibiting many qualities that at least appear to vary continuously. One might then construct maps that depict this world from various perspectives. . . . Here we have a way of combining what is valuable in both constructivism and realism. . . . We can agree that scientific representations are socially constructed, but then we must also agree that some socially constructed representations can be discovered to provide a good picture of aspects of the world. (26, quoted in Plutynski 2020, 161)

The key question here is what counts as a "good picture" of some aspect of the world. The answer offered by *haptic realism*, which I couple with perspectival pluralism (see section 2.1 in chapter 2) is that we should avoid the

22. See for instance, Mitchell (2003), Mitchell and Gronenborn (2017), Longino (2006, 2013).

notion of correspondence (between the scientific representation and mind-independent nature) and its ideal of pictorial accuracy. Instead, the suitability of a map is determined both by constraints coming from the domain it represents and by the demands placed on it by its eventual users. Scientific representations are the result of an interaction between human mapmakers and the territory they seek to navigate. Both participants in this interaction leave an ineliminable impression on the representation produced. Hence, science can never rise above the ground-level humanity of its makers. It can never achieve a God's-eye view.

In short, my preferred way to characterize scientific perspectives is as bets on how best to achieve productive simplifications. In this chapter, I have focused on the kind of complexity that comes with changeability—the temporal heterogeneity of the brain, the fact that neural systems are never in the same state twice. The correlative notion of simplicity is therefore stability. Neuroscientists have simplified the motor cortex through different idealizing assumptions about which properties of neurons and neuronal populations remain fixed. There are further questions that versions of perspectivism are normally presented with, questions about the relationship between perspectives: *Are they competitors or complementary to one another? Could one perspective ultimately subsume the others? Is perspectivism just a version of relativism? What is truth, according to perspectivism?* This is not the place to delve into all these matters, but I will briefly indicate how attention to the Heraclitean complexity of biological objects points toward some interesting answers here.

Expectations for some kind of convergence of perspectives are associated with versions of perspectivism that are closer to traditional scientific realism (Massimi 2018a). Indeed, some traditional scientific realists have charged perspectivism with not being a distinctive enough position, precisely because it seems easy enough to take different perspectives merely as partial views on one underlying reality that could in principle be unified once the partiality of each viewpoint is better understood, as in the parable of the blind men and the elephant. One maneuver that aids the traditional realist here is to assert that the basic properties of the entities of mind-independent reality are dispositional ones, which manifest in various ways depending on how the scientist choses to probe her object of inquiry (Chakravartty 2010). It follows that the dispositional properties of the one, mind-independent object of inquiry get manifest in different ways within

the different scientific perspectives, but in principle, there could be convergence, and truth could amount to achievement of an adequate correspondence with these mind-independent dispositional properties.

However, Chakravartty (2017, chapter 4) now argues that the positing of one set of dispositional properties underlying all the perspectives and affording in principle unification is a purely metaphysical one and cannot be in any way confirmed or disconfirmed empirically. I emphasize, in addition, that this move stands in denial of the view presented in this chapter that the brain, like other living systems, is a thing that is inherently changeable. The supposition of a substratum of stable dispositional properties underlying the variable observations is a concession to the thought voiced at the outset of this chapter that fundamental reality must be changeless, in spite of how things appear, in order to be and to be knowable.

Nietzsche, who along with Kant is often put forward as an instigator of perspectivism, is an oddly relevant interlocutor at this juncture. For him, the notion of truth dominant in the history of philosophy from Plato onward was one that set its sights on a mind independent, changeless reality, holding a supreme position over the flux of appearances. For Nietzsche, this "will to truth" was the core of an ascetic ideal that was inherently life-denying. As Poellner (2000, 7) puts it, the "other" that this ascetic ideal negates or devalues is, "life—which means, among other things, that realm of continuously changing, sensible particulars which confronts us in everyday experience." The thing we might conclude here is that in taking the changeableness that neural and other living systems present under observation to be basic, not a degraded mode of existence, not a mere becoming that we distinguish from true being, we must also give up on the venerable ideal of *eternal* truths, toward which the best science approaches.

This is not the place to give alternative positive proposals for a new notion of perspectival truth. I will just note, briefly, that the main position to be unsettled here is the idea of reality already carved into determinate structures, essences, and forms, which can be fully apprehended in a passive way by the knower.[23] In its full complexity, the world overflows all the conceptual forms that would make sense to a finite cognizer. The activity of knowing

23. What I'm describing here are some tenets of standard metaphysical realism. See Anderson (1998) on how Nietzsche and Putnam, both in their own way, reject them, leading to perspectivist positions that have a common origin in Kant's philosophy.

is that of bringing determinateness to bear on this excess of possible formations. This is why pluralism is inevitable: there must be more than one way to determine pattern in the flood appearances. But while absolute knowledge of eternal truths is ruled out, this does not mean acceptance of a relativism in which any system of determination—any body of knowledge—is as good as any other. As we have seen in this chapter, scientific perspectives can lose their viability and be superseded when not sensitive enough to the appearances that they seek to regularize.

## 7.4 Two Philosophical Perspectives on Mathematical Abstraction

In section 2.1 of chapter 2, I called this position *formal idealism*, referring to Kant's rejection of *formal realism*, Aristotle's vision of a natural world already delineated into a taxonomic scheme of beings with their various essences. To close this chapter, I will discuss how the contrast between formal idealism and formal realism applies to the distinction between Kantian and Platonist positions on the status of mathematical abstractions in natural science. It is interesting that the word "abstraction" bears two different meanings in contemporary philosophy—one lofty, the other mundane. In the lofty sense, an abstraction is an abstract entity, not spatially and temporally located, and as such possibly residing in Plato's heaven. Speaking mundanely (the use more common among philosophers of science), abstraction is synonymous with simplification and paired with idealization—an important model-building strategy employed by earthbound scientists. These two conceptions of abstraction animate two very different explanations for the "unreasonable effectiveness of mathematics"—that is, they provide two different answers to the question of why quantification is such a useful tool in science. The lofty explanation is that the underlying reality of nature consists in mathematical structures, and the task of the exact sciences is to discover these. The mundane one is that science progresses when humans find ingenious ways to simplify complex domains, and mathematics is the preeminent means for doing this.

Adherents to the lofty way of thinking about abstraction and the role of mathematics in science are in good company—not only Plato, but Galileo (stating that the book of nature is written in the language of geometry), Descartes (and other rationalist philosophers), and contemporary ontic

structural realists have intellectual kinship here. However, this Platonic tradition meets difficulty with the existence of pluralities of different kinds of mathematical representation of natural occurrences, examples much discussed in the literature on perspectivism in science.[24] If the construction of a predictively powerful model of a target system is also, in some sense, a revelation of the mathematical laws that underlie it, then how can it be that the book of nature seems to be written by multiple authors?

Kant is a figurehead for the mundane approach. Instead of taking abstract mathematical representations to be revelations of a mind-independent reality, we hold them to be a set of structures employed by human minds to regularize the many observations of natural occurrences.[25] When confronted with the varying, multifaceted, and ambiguous appearances, mathematics offers a useful set of structures for imposing representational order on them, especially by omission of details—the process of abstraction. It is not jarring or surprising to the Kantian view that there are multiple ways to achieve order and abstract; hence, there may be a plurality of kinds of mathematical models for the same target of research.[26]

This way of thinking about abstraction and the scientific method looks back to ideas commonplace in philosophy of science a century ago. Even

24. See Weisberg (2007) on "multiple models idealization"; and Morrison (2011) on the multiple conflicting models of the atomic nucleus. Further, Massimi (2018b) and Rice (2020) give responses to the problem of inconsistent models that they claim to be compatible with scientific realism.

25. Note that my characterization of the Kantian tradition does not come with any commitment to Kantianism regarding the ontology of numbers or epistemology of mathematics. For instance, one could think that mathematics is learned by the mind's apprehension of Platonic forms but still be a Kantian in the sense relevant here—that is, by denying that mathematical structures are the truer reality underlying the appearances in nature and asserting that the utility of math in science comes from the mind's ability to "project" certain simple structures onto nature. However, there is a connection between the Platonic tradition I characterize here and Platonism regarding the ontology of numbers, in that the indispensability arguments for the existence of numbers presuppose the lofty explanation for the success of mathematical science. I thank Alastair Wilson for this point.

26. There is much more that deserves to be said about how the question of the effectiveness of mathematics gets addressed on this viewpoint. My thoughts on this have been shaped by Cassirer (1910/1923, 1929/1957), except that I do not think that his ideas about scientific progress should be applied to neuroscience.

if neglected now, I believe they offer significant benefits for thinking about perspectival pluralism.[27] Not only does it welcome the existence of multiple perspectives, it also permits a relaxed response to the possibility of there being facts about reality that are in principle unknowable, which is a sticking point for Massimi (2018a, 353). In our case, we can say that the brain-in-itself is not knowable in its endless, Heraclitean complexity because no finite knowers would be able to theorize it completely and accurately as such. The brain-in-itself is not mysteriously remote, removed from observation; it is just excessively complicated. At the same time, the Heraclitean brain provides constraints on what counts as an acceptable representation of it, and this means that neuroscientific knowledge claims are not merely fictitious or relativistic.

It might be objected that the Heraclitean brain, insofar as its instability is observed through advanced scientific instruments and protocols, *is* known scientifically. However, collections of observations do not constitute a body of scientific knowledge; there has to be systematization and theorization. It is with the systematization—which is the result of a theoretical approach and modeling perspective—that the full changeability of the brain is taken out of the picture. The Heraclitean brain is replaced with a more stable counterpart. In the words of Bergson (1907/1944, 366), "Real time, regarded as a flux, or, in other words, as the very mobility of being, escapes the hold of scientific knowledge."[28]

---

27. I am thinking here of Duhem (1906/1954), Cassirer (1929/1957), Husserl (1970), and Whitehead (1925/1967).

28. What is shown is that there are limits to neuroscientific knowledge attained through the simplifying perspectives of exact science because it must abstract from the Heraclitean quality of brain processes. This does not preclude forms of knowing or investigating that in some way acknowledge the flow of neural or mental life. In the history of psychology, this seems to be the point of William James's treatment of the "stream of thought," and Dilthey's (1894/2010) insistence on a "descriptive" form of psychology that begins with experience of one's own life, complementary to psychology that borrows the "explanative" approaches of the physical sciences. It is interesting that James (1890) makes a neurophysiological argument for the claim that felt sensations never exactly repeat: "Every sensation corresponds to some cerebral action. For an identical sensation to recur it would have to occur the second time *in an unmodified brain*. But as this, strictly speaking, is a physiological impossibility, so is an unmodified feeling an impossibility; for every brain-modification, however small, must correspond a change of equal amount in the feeling which the brain subserves" (232–233; emphasis in original).

# Part III

The purpose of this part of the book is to expand upon the philosophical lessons suggested by the case studies of part II. Chapter 8 is about what we can learn regarding that most basic question, *What is science?*, as well as *How is technology changing science?* The final two chapters delve into some classic problems in philosophy of mind—the explanatory gap for consciousness and mind-body dualism. The payoff of the fixation on abstraction, exemplified in this book, is that we can help ourselves to new diagnoses of what is at stake in these old debates.

# 8 Prediction, Comprehension, and the Limits of Science

> Science should be confident that its powers to know are not limited owing to its track record of starting at a point no one can question, and going on successfully to unify an incredibly diverse range of phenomena under one explanatory "roof." That success has had a never-ending pay-off in technological application. It's possible that there are limits to science. But it would be foolish to bet there are.
>
> —Alex Rosenberg (2014, 40)

## 8.1 On Limits

Neuroscience is a frontier science. Interred with this imperialistic dead metaphor is the notion that the brain is an expanse of undiscovered country, little by little succumbing to the advancing forces of scientific exposition. One notable pioneer of neuroscience, Emil du Bois-Reymond, started a controversy, the *Ignorabimusstreit*, when he voiced the opinion that the explanation of consciousness—how the material brain conditions "mental facts"—posed an impassable barrier to scientific progress: "Between these limits the man of science is lord and master; he dismembers and builds up and no-one durst say wherein his knowledge and his power are circumscribed. Beyond these limits he cannot now, nor can ever go" (1872/1874, 29).

The question of limits is not here to be treated as a test for who is a "booster" or "naysayer," "optimist" or "pessimist," about science and modernity, which is normally what happens.[1] Instead, I wish to prompt scrutiny into our concept of science, for to set boundaries or limits is to demarcate and therefore to

1. See Beiser (2014, chapter 3) on the *Ignorabimusstreit*, and Churchland (1994) for invective against the naysayers.

define. The demarcation question familiar to most philosophers of science is the one that asks where the boundary between science and pseudoscience is to be drawn. I think it is more helpful to examine the border between science and engineering.[2] The occasion for this inquiry is the recent encroachment, as one might call it, of a particular kind of engineering—machine learning—into the territory of neuroscience. Not only is the engineering of artificial neural networks (ANNs) becoming unified with neuroscience, but the success of machine learning on challenges not met by more entrenched scientific methods has caused many researchers to rethink the aim and scope of their discipline. All this will be discussed in section 8.3. Before then, I will pick up a line of thought dropped at the end of chapter 5, where it was concluded that the slogan "What I have not made, I do not understand" best summarizes the role of artifice in neuroscience. Section 8.2 shows how this conclusion is reinforced once we consider the interrelationship between science and technology, especially as it came about in the early modern precursors to current science. Section 8.3 is about the impact of machine learning on neuroscience today. Finally, section 8.4 expands the argument that the ineliminability of artifice in neuroscience is itself a barrier or limit to neuroscientific understanding of the brain in all its native complexity. The technology of machine learning, as powerful as it is, does not break through this limit.

## 8.2 *Verum Esse Ipsum Factum*[3]

This book has placed simplification at the center of its account of what neuroscience is up to. This occurs not only in the domain of models and other scientific representations, but also, as the studies of chapters 3 and 5 showed,

2. This demarcation has been usefully discussed within philosophy of biology; for instance, see Kastenhofer (2013) on the relationship between systems biology and synthetic biology.

3. "*Verum esse ipsum factum*" (Vico 1710/1988, 46) means "The intelligible is precisely what is made." In the introduction to that text, Palmer (18), Vico's translator, explains:

> The formula states an analytical relation, namely, that only as made is something interchangeable, or convertible, with the true—and thus intelligible to its maker or doer. But it can also have some degree of intelligibility to any being who could make or do it. The formula is not a case of the correspondence theory of truth. *Verum* does not mean "true" as the function of a proposition (whose opposite is "false"); rather, it means true as intelligible. (Also, cf. Fisch 1969, 407–408).

in the material arena of experimentation, where we find the production of simpler model systems as clearings in the otherwise impenetrable forest of neural complexity. To simplify, to make a simpler thing, is an inherently constructive and also destructive process. It should not, as I have in various places emphasized, be misinterpreted as the discovery of a hidden simplicity which was there all along. That the brain houses a latent simplicity, within reach of discovery, is probably an idea as fabulous as El Dorado or the mines of King Solomon. The benefits of simplification, some discussed in part II of this book, are in turns pragmatic and epistemic. Simpler systems are certainly easier to manipulate and understand, and I concur with Potochnik (2017) that understanding—the felt sense that the processes of nature have been made comprehensible (see section 2.2.3 of chapter 2)–is a primary epistemic aim of science. By examining the material as well as representational simplifications in the case studies in part II, one sees that these strategies make things intelligible to the scientist by making things that *are* intelligible to the scientist.

The end point of chapter 5, that an amendment of Feynman's saying gives a more honest reflection of the operation of neuroscience ("What I have not made, I do not understand"), might have seemed too far-reaching a conclusion from too small an inductive base. The task is now to show that the proposal is reinforced once related to work in the history and philosophy of science that has given prominence to the interdependency of science and technology. The *verum factum* is due to Giambattista Vico, who thematized the connection between doing, producing, and knowing as against what Dewey later called the "spectator" theory of knowledge.[4]

---

Dupuy (2009, 137) writes of cybernetics giving "final expression to the principle of *Verum et factum convertuntur*."

4. See Fisch (1969) for a comparison between Vico's philosophy and American pragmatism. See also O'Malley (2009), who takes Feynman to be expressing the same sentiment as Vico. Feynman is normally presumed by neuroscientists and synthetic biologists to be saying that the ability to create a phenomenon is a prerequisite for the claim to understanding it—for instance, "Until we can assemble a form of life in vitro from defined, functionally understood macromolecules and small-molecule substrates, how can we say that we understand the secret of life?" (Forster and Church 2007, 5, quoted in van den Belt 2009, 258). The lesson that I am drawing from Vico, in contrast, is that in these disciplines, a thing's *having been created* is necessary for the understanding of it. According to Landeweerd (2021, 62), this is closer to Feynman's original intent since he was referring to the re-creation of the steps of a mathematical argument toward a

Vico's own application of his maxim was to the point that both geometry and the human-made social world are more fully comprehensible than the divinely made natural world; mine is that neuroscience is in its own business of producing objects of comprehension as proxies for those natural objects held by Vico to be transcendent. These model systems are, like domesticated animals, both natural and artifactual at the same time.

I do not intend to take the *verum factum* just on the authority of Vico (as he held it on the authority of the ancient Italians whose wisdom he detected in the etymology of these words). While it may be that there are cases from across the sciences for which this epistemology of understanding-via-making is not apt, it is consistent with an argument made compelling by various twentieth-century historians of science that artifacts, not natural systems, are the primary and original objects of rigorous theorization, indicating that the reliance on stand-ins for the natural or wild-type system is by no means a peculiarity of contemporary neuroscience. This was the thesis of Boris Hessen and Henryk Grossmann, who both asserted that science, in its most nascent stages, came into being through the study of technological objects (Freudenthal and McLaughlin 2009, 10). Here is Grossmann, who stated:

> It is evident that man, in all these technological upheavals, acquired new, important material for observing and contemplating the actions of forces. In the machines, in the turning of the water wheels of a mill or of an iron mine, in the movement of the arms of a bellows, in the lifting of the stamps of an iron works, we see the simplest mechanical operations, those simple quantitative relations between the homogeneous power of water-driven machines and their output, viz. those relations from which modern mechanics derived its basic concepts. (1935/2009, 128)[5]

Similarly, Paolo Rossi (1962/2002, 56) writes of "the importance that many practical problems (e.g. the speed of ships, construction of canals, ballistics, manufacture of pumps, ventilation of mines, etc.) came to assume with respect to the birth and progress of investigations of a theoretical character

---

theoretical result. It is also interesting to compare this with Kant's saying that, "reason has insight only into what it itself produces according to its own plan" (*Critique of Pure Reason*, quoted in van den Belt 2009, 258).

5. Grossmann also outlines the economic imperatives behind these technological activities in the context of an emerging capitalist system of production. Cf. Merton (1938/1970, chapter VII). The economic imperatives behind advances in machine learning are not to be discounted, though there is not space to discuss them here.

(hydrostatics and hydrodynamics, astronomy, chronometry, dynamics)." He also makes the connection between the rise of machine models for explanation in the physical universe, as we see in the image of God the watchmaker, and the thesis that human knowledge is properly of the humanly constructed (1962/2002, 23).

It is significant for Rossi that many of these technologies long predated their theorization. For example, lenses were an invention of the twelfth or thirteenth century, but only later did they receive attention as theoretical objects in the sixteenth century and beyond in works such as Della Porta's 1589 *Magia naturalis* and Kepler's 1604 *Paralipomena* (Rossi 1962/2002, 55). The point is that technology and invention can easily exist independently of a body of knowledge that explains their workings conceptually. Canguilhem's essay "Descartes et la technique" highlights Descartes's interest in the acquisition of knowledge about how devices work. It is noted that the philosopher "despises art [i.e., technology] without explication, inventors without method" (Canguilhem 1937/1982, 114). The pursuit of truth is to be guided by the rules of the Cartesian method such that the practice of invention is systematic, not haphazard. Knowledge is to be sought not least for its practical effects, as we find in the famous "masters and possessors" passage from the sixth *Discourse* of 1637. Descartes writes of how the learning of some general notions in physics made him aware of their potential

> to secure the general welfare of mankind. For they opened my eyes to the possibility of gaining knowledge which would be very useful in life, and of discovering a practical philosophy which might replace the speculative philosophy taught in the schools. Through this philosophy we could know the power and action of fire, water, air, the stars, the heavens and all the other bodies in our environment, as distinctly as we know the various crafts of our artisans; and we could use this knowledge—as the artisans use theirs—for all the purposes for which it is appropriate, and thus make ourselves, as it were, the lords and masters of nature. (1985, 142–143)[6]

Of the many possible things to say here, we will notice in particular how knowledge of crafts, once well articulated, is put forward as the model for knowledge of nature, and how mastery of the world beyond the workshop will stem from this successor to the "speculative philosophy."

6. See Rossi (1956, 142) for a discussion of Descartes that is in a similar vein as Canguilhem's.

This mid-twentieth-century historiographical tradition leaves its mark on Peter Dear's more recent answer to the question of how to define science for the purposes of demarcating an area of study for the history of science. In his view, Western science[7] is the product of a unification that occurred in the early modern period, bringing together natural philosophies and enterprises aiming at the disinterested understanding of nature, with materially directed practices of technology. The idea is that "this period saw the establishment of a new enterprise, one that took the old 'natural philosophy' and rearticulated it in the new terms of instrumentality: the engagement with the world that, in the nineteenth century, produced modern science was thus born of a discursive hybrid of these analytically unrelated endeavors" (Dear 2005, 397).[8]

The point about the two premodern enterprises being "analytically unrelated" stems from the frequently made observation that the natural philosophies of the Greco-Roman and medieval Christian worlds were contemplative activities whose "effects," so to speak, were to be found within the practitioners—the elevation of their minds.[9] The stability of the old view,

---

7. Of course, this is a contestable term, but it suffices to mark the tradition to which contemporary neuroscience belongs.

8. A similar conception may be familiar to philosophers of science from the work of Ian Hacking:

> Reality as intervention does not even begin to mesh with reality as representation until modern science. Natural science since the seventeenth century has been the adventure of the interlocking of representing and intervening. (1983, 146)

In addition, consider the term *technoscience*, which foregrounds the dual nature of this activity (Kastenhofer and Schmidt 2011, 127). Technoscience is explicitly defined by Nordmann through the association of knowing with the making of phenomena:

> *In technoscientific research, the business of theoretical representation cannot be dissociated, even in principle, from the material conditions of knowledge production* and thus from the interventions that are required to make and stabilize the phenomena. In other words, technoscience knows only one way of gaining new knowledge and that is by first making a new world (2006, 8, emphasis in original).

Nordmann, unlike Dear, is keen to assert a division between "classical modern science"—in which the "in principle" dissociation could occur—and contemporary technoscience. Note that Nordmann's (2006, 23) subsequent characterization of technoscience is applicable to the situation that will be described in section 8.3.1, where models and tools have lost their intelligibility so that the epistemic task of understanding becomes moot.

9. See Hadot (1995, chapter 3) on the study of physics as a "spiritual exercise." For Dewey, Aristotle is the exemplar of the attitude whereby "pure intellectual inquiry";

in which the desire to achieve understanding of the processes of the natural world was not associated with the aim of turning those processes to material, mundane purposes such as medicine or agriculture, shows that there is no inherent link between these projects.[10] According to Dear, the marriage of these distinct tasks was definitive of science. Nowadays the connection appears inextricable—to claim an innermost *knowledge* of the workings of nature but no kind of power over it is, to the descendants of Francis Bacon, to admit that your knowledge is counterfeit or empty. Yet the commonly held view, contested by Dear's conception of science, is that science, properly speaking, has nothing *necessarily* to do with technology; it is knowledge for knowledge's sake, which just so happens to produce the results that make technological progress possible (Dear 2005, 401–402).[11] This view is the "cascade" model, in which pure science is the source from which applied science flows (Carrier 2004).[12] The incompatibility with Dear's picture lies in its assumption that there is such a thing as science in pure form—purified

---

that is, "knowledge as something not to be put to any use" is most highly prized (1929, 75). He relates this to a denigration of practical skills: "The mechanical arts dealt with things which were merely means; the liberal arts dealt with affairs that were ends, things having a final and intrinsic worth. The obviousness of the distinction was reënforced by social causes. Mechanics were concerned with mechanical arts; they were lower in the social scale" (Dewey 1929, 74).

See Schuhl (1938, 11–12) on this attitude in the writings of Aristotle and Plato.

10. The stability of this view must be seen in the light of the rigidity of social orders in which a learned elite set itself as far apart as possible from the class of slaves, serfs, or menial laborers, those who made things and dirtied themselves. Zilsel (1942) argues that the story of science is bound up with the history of class relations, specifically, the rise of the bourgeoisie—a middle class between laborers and the learned elite. Coincident with the first development of science comes a value system that does not exalt disinterested contemplation and eschew worldy engagement, and hence is willing to put knowledge to work toward the mastery of nature (Schuhl 1938, 30–31). R. K. Merton finds this ethos strongly (but not exclusively) expressed in Puritan communities. For instance, he writes of the "Protestant ethic" that "this scheme of orientation embraced an undisguised emphasis upon utility as well as control of self and the external world, which in turn involved a preference for the visual, manual and concretely manageable rather than the purely logical and verbal" (Merton 1938/1970, 115).

11. See, for example, the quotation from Rosenberg (2014) at the beginning of this chapter. Dear does argue, moreover, that by presenting itself as knowledge for knowledge's sake, science acquired the prestige of natural philosophy.

12. The cascade model is also known as the "linear model" of Vannevar Bush.

of its instrumental aims, with no bearing to material application, however remote or latent.

The point I wish to draw out is that the coupling of the task of understanding nature to that of controlling nature brings with it a reconfiguration of epistemological terms to accommodate both of the participants in this conjoined endeavor. The way of knowing of science must be a way of knowing that facilitates or at least accommodates practices of manipulation and control. The premium placed in science on simplification and communicability of knowledge (as opposed, for example, to ways of knowing that do not attempt to prune away complications or deny the ineffable) is not unrelated to the circumstance that science, this mode of understanding nature, is also an instrumentally minded one. Simplification—as opposed to resignation at the complexity of things as they stand—facilitates real-world effectiveness and communicability; and without communicability, the accrual of expertise across generations, critical for technological progress, is impeded.[13] We can also now appreciate the connection between Dear's historical definition of science as a hybrid of natural philosophy and instrumentality and the initial prompt for this chapter, the thesis that artificial systems are the primary (or at least original) targets of theoretical understanding in science. On the one hand, artifice supplies an array of simpler model systems that help bootstrap the investigator into dealing with somewhat more complex ones: artificial systems aid simplification, which is itself crucial for instrumental capability. On the other hand, this process of hybridization comes with a destruction of the former epistemic and metaphysical division between the natural and artificial. Without the

13. See Rossi (1956, 140–141) on communicability and the communal process of science, especially in relation to Francis Bacon. The inherent instrumentality of scientific thought is thematized by at least two very different twentieth-century philosophers—John Dewey and Theodor Adorno (Snir 2020, chapter 2). While Dewey gives it a positive spin, the message from Adorno is that scientific understanding of things is tied to a logic of domination, crucial for capitalism but antithetical to thought, properly speaking. This connection is to be found in the Enlightenment ideal of science as the inquisitor of myth and metaphysics, where "anything which does not conform to the standard of calculability and utility must be viewed with suspicion" (Horkheimer and Adorno 1947/2002, 3). Adorno does not elevate nonpragmatic theory over praxis, but he does call for a different integration of the two, as expressed in the essay "Marginalia to Theory and Praxis" (Adorno 2005); see also Horkheimer (1947/2013) on the confrontation with American pragmatism.

removal of this barrier, the scientist's reliance on artifice, in the process of acquiring knowledge, could not be justified.

The old division, of course, is strongly associated with Aristotle's teleological natural philosophy.[14] A commonplace of intellectual history of the seventeenth century is that mechanism replaced teleology in the cosmic picture. Dear concurs, with reference to Robert Hooke's *Micrographia* of 1665, that the mechanical philosophy of nature was demanded by the newly conceived practice of natural philosophy aiming at control of nature: "The logical incommensurability between natural philosophy and utility is short-circuited by having natural philosophy speak only in the terms of mechanical tools" (Dear 2005, 397). Similarly, an implication of the Hessen-Grossmann thesis that science founded itself on the analysis of technologies is that with the transference of the framework honed on machines to nontechnological objects, such as planetary motions, came a reconceptualization of the "natural" (Freudenthal and McLaughlin 2009, 10). At minimum, the natural world would have to be conceived as something not fundamentally different in its operating principles from the artificial.[15] Along with the production of hybrid natural-artifactual objects for experimental science, there is the intellectual construction of systems not made by humans as still machine-like.

We can find in the *Principles* of 1644 (part IV, section 203), Descartes once again expresses the new ethos, where discoveries are made through what we would now call a reverse engineering of nature:

> In this matter I was greatly helped by considering artefacts. For I do not recognize any difference between artefacts and natural bodies except that the operations of artefacts are for the most part performed by mechanisms which are large enough to be easily perceivable by the senses. . . . Moreover, mechanics is a division or special case of physics, and all the explanations belonging to the former also belong to the latter. . . . Men who are experienced in dealing with machinery can take a particular machine whose function they know and, by looking at some of its parts, easily form a conjecture about the design of the other parts, which they cannot see. In the

14. However, see Newman (2004, chapter 1) for refinements of this picture of Aristotle.

15. This shift is discussed by Rossi (1956, 142) and Canguilhem (1965/2008c) in relation to Descartes, and by Schuhl (1938, 33) in relation to Francis Bacon. Of course, the very notion of the "natural world" is fraught and contestable. Here, I mean only to refer to things not brought into being by people's activity in the world around them. See Hadot (2006) for one study of the idea of nature in the history of natural philosophy and early science.

> same way I have attempted to consider the observable effects and parts of natural bodies and track down the imperceptible causes and particles which produce them. (Descartes 1985, 288–289)

Among so many things, the reader should note the succinct justification given for the investigation of nature via artifact analogies: granted no fundamental difference between the natural and the artifactual, the process by which an engineer would learn of the workings of machinery is directly transferable to the workings of nature, and knowledge of macroscopic mechanisms is projectable down to the microscopic mechanisms said to be at play in nature.

I will shortly say more about neuroscience in the light of the definition of science presented here. One last point, for now, is to offset the concern that the primacy of theorization of artifice is a peculiarity of science at its supposed inception in the sixteenth and seventeenth centuries. Two of the big theoretical achievements of the nineteenth century, which incidentally is the century in which science took the institutionalized form that we know today, had their first footings in the realm of technology. I am referring here to the theory of thermodynamics in physics and the theory of natural selection in biology. Carrier (2004) presents thermodynamics as an important example of "application innovation," his term for the phenomenon of theoretical innovation originating in the context of technological application, in contrast to the cascade model, in which theory is the precursor to applied science. The important detail here is that the invention of steam engines occurred without a scientific theory, but once these machines had been created, thermodynamics grew out of an examination of their workings with a view to their improvement, the founding document being Sadi Carnot's 1824 treatise on the motive power of heat. With the full development of thermodynamics, it came to be a theory of energetic relations in all things, not only engines. As Rabinbach (1990, 46) relates, "Thermodynamics conceived of nature as a vast machine capable of producing mechanical work or, as von Helmholtz called it, 'labor power.' Initially a measurement of the force of machines, 'labor power' became after the discovery of energy conservation the basis of all matter and motion in the physical world."

Darwin's own account of the origins of the theory of natural selection gives preeminence to his study of the practice of breeding by artificial selection, as well as the discovery that an analogy could hold between this and the process of speciation. While there is something of a scholarly controversy

over whether Darwin's self-report should be credited, the technology of artificial selection certainly cannot be ignored in the background to the *Origin of Species* (Largent 2009). Writing of the significance of the analogy between natural and artificial selection, Secord (1981, 164) explains, "While his basic orientation—in both social and intellectual terms—always remained that of a naturalist, Darwin became one of the few to study the productions of man with the scientific care usually reserved for the productions of wild nature."

With all this in view, it should now be clear that the reliance of neuroscience on the computer model, as well as its failure to deal with brain "in its own terms," without such systems of comparison, are not symptoms of the underdeveloped state of neuroscience, as Daugman (2001) and Eliasmith (2003) would have it. Rather, they are signs of neuroscience's continuity with the rest of science.

### 8.2.1 Neuroscience So Considered

> The intention and the result of a scientific inquiry is to obtain an understanding and a control of some part of the universe.
>
> —Rosenblueth and Wiener (1945, 316)

I have just discussed a historiographical tradition that supports my proposal from chapter 5, that the making of an artificial object of investigation is a prerequisite for understanding in neuroscience. The key idea is Dear's characterization of science as a "discursive hybrid" of projects aiming to understand nature, on the one hand, and to instrumentalize it, on the other. Beyond the point that the intelligible is the made, that the neuroscientific aim of understanding the brain is served by the construction of experimental and representational objects, it is worth briefly considering some other aspects of neuroscience in light of these historical perspectives, as they yield insight into a number of common assumptions and research practices.

First, it is easy to find expressions of the view that the deliverables of neuroscience will be both understanding of the brain, and effective means of intervening in brain disorders. To give one example, a group of high-profile researchers state the rationale for the federally supported Brain Research through Advancing Innovative Neurotechnologies (BRAIN) initiative as follows: "The envisioned long-term pay-off of the BRAIN Initiative is a more comprehensive understanding of how the brain produces complex thoughts

and behaviours that will provide an essential guide to progress in diagnosing, treating and potentially curing neurological and psychiatric diseases and disorders that devastate so many lives" (Jorgenson et al. 2015, 2).

Notice the adherence to the cascade conception, whereby basic research aimed at understanding will subsequently guide applied research aimed at intervention. We saw earlier that this conception underestimates the way that supposedly pure or basic research is intrinsically connected with instrumentality in its construction and conceptualization of its objects, even when not overtly set out on an application. Instead of the cascade conception, with its notion of pure science, we should picture a situation in which science, founded on model systems—either abstract mathematical models, or concrete ones—presents two alternative faces. By virtue of their simplicity, models afford understanding; at the same time, because of their simplicity, they serve instrumental purposes, aiding prediction and manipulation. To reiterate, the model is not a representation (in the case of a mathematical model) or exemplar (in the case of a concrete model system like the cerebral organoid)[16] of an inherent simplicity in nature. The model is simple by virtue of its being an artifact designed to render things more simple than they actually are.

The Grossmann-Hessen thesis makes sense of the observation that invention of new technologies preceded theorization as neuroscience developed in the mid-twentieth century. Computationalism could not have become the dominant theoretical framework for understanding neurophysiological phenomena if it had not been for the prior invention of electronic computers, and the mathematical tools attending to them, as reported in many histories of cybernetics.[17] More specific examples can also be found. Information theory was invented after World War II as a statistical framework for signal engineers. It was then picked up by researchers such as Attneave (1954) and Barlow (1961) as a framework for theorizing responses in the visual system. Movshon (2021, 188) writes of Barlow's enthusiasm for information theory as surpassing anything that psychology or neuroscience had developed by itself, and that he considered the paper by Shannon (1951) to be "one of the

16. These are small clumps of neural tissue grown from stem cells in petri dishes. See, for instance, Nowogrodzki (2018).

17. This point is discussed at length in chapter 4, and also, for instance, in Dupuy (2009), Kline (2015), Abraham (2016), and Husbands and Holland (2008).

greatest ever in the whole of psychology and neurology, even though Shannon was a mathematically inclined electrical engineer with no training either in psychology or neuroscience." The deployment of engineering concepts is a feature of Barlow's output as a theoretician, and this tradition continues in more recent texts such as *Principles of Neural Design* by Sterling and Laughlin (2015). Predictive processing models, which have generated much interest within neuroscience and philosophy of cognition (Hohwy 2014, Clark 2016) first came to light in the context of television engineering.[18] Nowadays, ANNs are the inventions that are shaping neuroscientists' conception of "biological neural networks," and these will be considered in section 8.2.

We saw that already in Descartes's *Principles,* there is the conception of natural scientific research as a kind of reverse engineering. This strategy is common in theoretical neuroscience, nicely explained by Dennett (1995) in relation to Marr's theory of vision, and which I have discussed elsewhere in the case of theorizing the robust properties of nervous systems (Chirimuuta 2017a). The strategy, as Descartes made explicit, requires the denial of a fundamental or essential difference between natural and artificial objects. By noticing that this methodology demands commitment to a philosophical view about the ontological equivalence of organs and artifacts, we can make sense of the dominance within the scientific community, as noted in chapter 4, of the literal interpretation of computational models of the brain. The default opinion among researchers is that a neural system is identical, at some level of abstraction, with a computational model representing it, even though the dictum of statistician George Box—that "all models are false, but some are useful"—is so often heard on neuroscientists' lips (e.g., Lindsay 2021, 15). However, the historical perspectives lend support to my analogical interpretation of computational models—as against a literal one, even though the literal one has the authority of the scientists behind it—for they reveal why practitioners themselves would tend toward literalism, and this tendency is based not on principle or discovery, but rather it is conditioned by the methodological framework.

18. The source of the model is mentioned by some early adopters in neuroscience, Srinivasan, Laughlin, and Dubs (1982, 428), and by Sterling and Laughlin (2015, 249): "Predictive coding, an image compression algorithm invented by engineers almost 60 years ago to code TV signals efficiently, is implemented in animals by a basic sensory interaction."

The historical view also helps to meet a potential objection that the analogical interpretation collapses into either a literal or metaphorical construal of these models. Lande (2019) observes that the choice is normally a dichotomy: the claim that the brain is a computer is either a literally descriptive one or a mere metaphor, without descriptive ambitions. My analogical interpretation presented itself as a distinct alternative to these. An obvious objection is that either I mean by "analogy" a similarity, a sharing of a subset of properties, which would collapse the view in literalism, or I mean it as a figure of speech, a simile, which would collapse the view into the idea that brains are computers, just metaphorically. As already mentioned in chapter 4, we may rule out analogy as direct similarity or sharing of properties by pointing out that what the computational model represents is an *ideal pattern*, a set of features suggested by neural activity but underdetermined by it. The work in chapter 5, on the creation of ideal patterns and the constructive work that goes on in experimental neuroscience, reinforces this point.

Finally, the historical perspective lends itself to a more general way of considering the distinctive place of analogies in science. As Hesse (1955, 353) says, analogies allow the unfamiliar to be described in terms of the familiar. One can think of this along the lines of domestication. *Domestication* is the process of making a wild thing familiar or "homely," but it is at the same time a process of making a new kind of creature, one both natural and artificial, amenable to somebody's purposes. The point is that scientific analogies work not by simply fixating on similarities between the familiar and unfamiliar, but by the construction of objects that are intermediary between what is artificial and comprehensible and what seems wild and inscrutable. This is all consistent with the *verum factum*, the idea that intelligibility comes about with the process of making, and with the view that historically, theorization begins with machines and devices and then circles out to encompass other objects. In neuroscience, the expansion of theory is to systems that are organic, but these should not be thought of as purely natural: they are at the same time still artificial. Classic computational models have some claim to be literally representative, albeit in an approximate manner, of ideal patterns gleaned from the simplified behavior of neural systems under artificial conditions. Thus, computation is not a mere metaphor, but it is also not literally representative of neural activity beyond the enclosure of the laboratory. Its further relevance, its representational adequacy vis à vis the wild type, is indeterminate.

I emphasize that even with respect to artificially constrained neural systems, computational models can be taken as literally representative only at a high level of abstraction. There are countless material dissimilarities. A view expressed by computational neuroscientists is that these dissimilarities are inessential to processes by which the brain enables cognition (Lindsay 2021, 15).[19] But given the number of unanswered questions about what details matter in neural systems (Shenoy 2015, 83), this is too hasty a view to maintain. At the same time, this is the view that has the weight of the Cartesian tradition behind it, invested in the denial of essential differences between organic and mechanical objects. The philosophical significance of disanalogies between brains and computer models will be the subject of chapter 9. I will conclude here by reiterating the lesson of chapter 5, that reliance on artifice—the construction of hybrid systems, and the deployment of artifacts like computers as prisms through which to glance at nervous tissue—constitutes a limit to scientific understanding of the brain. Neuroscience cannot deliver knowledge of the brain unadulterated by instrument analogies, and indeed, by the instrumental aims bound into the scientific enterprise.[20] Therefore, the brain-in-itself is not a possible object of scientific comprehension. One might wonder, though, that the construction of more brainlike machines could deliver some means to surpass this limit (Bongard and Levin 2021). This possibility will be considered in section 8.4 following an examination of these new technologies.

---

19. However, we saw in chapter 5 that experimental neuroscientists are more likely to emphasize such differences.

20. I do not mean to suggest that this limitation is a peculiarity of neuroscience. That the sciences cannot envisage their objects other than as a "manipulandum" is a point argued by Merleau-Ponty (1961/2001, 288):

> Science manipulates things and gives up living in them. It makes its own limited models of things; operating upon these indices or variables to effect whatever transformations are permitted by their definition, it comes face to face with the real world only at rare intervals. Science is and always has been that admirably active, ingenious, and bold way of thinking whose fundamental bias is to treat everything as though it were an object-in-general—as though it meant nothing to us and yet was predestined for our own use.

A thing that is distinctive about this limitation when it comes to the mind-brain sciences is that failure to appreciate the gap between model and target has troubling ramifications for human self-understanding. This is a point that Merleau-Ponty insists on, as discussed further in Chirimuuta (2020d).

## 8.3 Renegotiating the Relationship of Understanding and Control

> Since the dominant ideology of modern science is inherently unstable, what counts as science constantly requires reestablishing and remaking.
>
> —Peter Dear (2005, 405)

One important caveat is that I do not mean to give the impression that human-designed systems are all perfectly intelligible. There are limitations even of the maker's knowledge, a point well put by Horace Barlow.[21] Still, my view is that the neuroscientist's understanding is based on there being artificial systems that are at least to some extent intelligible, and these model-based practices are consistent with a scientific tradition in which for centuries the projects of understanding and control of nature have been intertwined. But there is no guarantee that model systems will always be intelligible enough to support this union. What happens when the models themselves become so complex that the scientist does not really see how they work? This was a situation envisaged by John von Neumann, one of the originators of cybernetics:

> At the Hixon Symposium, finding himself taxed by the neurophysiologists (including McCulloch) for not stressing enough the difference between natural and artificial automata, he replied that this distinction would grow weaker over time. Soon, he prophesied, the builders of automata would find themselves as helpless before their creations as we ourselves feel in the presence of complex natural phenomena. (Dupuy 2009, 142)

And this is, arguably, the situation in which computational neuroscience finds itself today. The state-of-the-art models of many neural systems are deep (i.e., many-layered) ANNs with millions of parameters. The debate over the question of the intelligibility of these models is fraught and ongoing. What is uncontroversial is that they are far less easy to understand than the simpler, older classes of models, and their success in neuroscience as predictive devices comes with challenges to established theories. As Saxe, Nelli, and Summerfield (2020, 57) write, "At worst, the deep learning framework seems to face neuroscience with an existential challenge. . . . it seems to propose sweeping away existing knowledge about how specific

21. "The person who understands most about a machine is its designer, and no designer of a complex machine would claim that everything about it was perfectly understood" (Barlow 1990, 16).

classes of computation underpin behaviour, merging the goals of theoretical neuroscience with those of contemporary AI research."

Some commentators, such as *Wired* magazine's Chris Anderson, have taken the rise of machine learning methods in science to presage an "end of theory"–the obsolescence of the scientific method as known until today, where clearly articulated hypotheses and models were tested against human-scale data sets.[22] The task of this section is to examine what happens to neuroscience as it departs from its traditional modus operandi of building models that are simple enough to be, for the most part, understood by their creators. Since it is always too early to predict future history, I will describe three possible scenarios: first, that we are witnessing a radical break from neuroscience as we know it; second, that there is no significant breach from past practice, just a change in the set of theoretical questions deemed tractable; and third, the proposal that machine learning–based science is an intensification of certain tendencies established in the prior history of science.

### 8.3.1 Divorce of Prediction and Understanding

An implication of Dear's conception of science, applied to current model-based research, is that the scope of science is limited to those areas for which the models are simple enough to be understood. On this conception, understanding is one of the two essential faces of science, and without it, a practice would revert to being instrumentality (technology) alone. The point here is that science is demarcated as a mode of activity in which work productive of understanding cooperates and comingles with practices generating instrumental control. A major component of instrumental success is the development of accurate means of prediction. So we can take prediction as a key instrumentalist goal, standing alongside the goal of understanding. If the two goals cannot be harmoniously satisfied—if practices resulting in understanding inhibit the achievement of instrumental control, and vice versa—then these two forms of activity, technology and natural philosophy, whose union is definitive of science, will have to go their separate ways. Arguably, this is the situation facing neuroscience today, for a thing made apparent by the entry of machine learning into neuroscience is that there is a trade-off between understanding and the instrumentalist

22. See Anderson (2008), and for a critical discussion, see Boon (2020).

aim of prediction. With complex enough systems, like the brain, the most predictively accurate models are too complex to yield much understanding, and the most theoretically illuminating models are too simple to be good at prediction.[23]

An instance of this trade-off—"between a model's simplicity and its ability to accurately predict neural responses" (Butts 2019, 458)—has already appeared in chapter 5. There, it was noted that the classic linear-nonlinear models of primary visual cortex (V1) cells were fairly accurate in their prediction of responses to artificial visual stimuli, but they failed when stimulus conditions became more naturalistic. However, the models were highly transparent and afforded a computational explanation of the operation of these cells, as proposed by Barlow and others. The deep convolutional neural networks (DCNNs), now employed for modeling in visual neuroscience, yield accurate predictions, even for responses to natural images and movies, but such models are far less interpretable and may never lead to a more advanced understanding of the cells' response profiles.[24]

Discussion of the trade-off between prediction and understanding has not been confined to neuroscience. It appears in the literature on statistics and psychology, disciplines that are now like neuroscience witnessing the power of machine learning to make inductions on the basis of gigantic data sets, without enforcing the assumptions of linearity that made the mathematics of old-fashioned models tractable and interpretable.[25] It is interest-

23. The case for the trade-off is put forward at length in Chirimuuta (2020c); and in Chirimuuta (2023a), I discuss why models that fail predictive accuracy still count as delivering understanding of the nonfactive sort. Here, I focus on scientists' discussion on the trade-off.

24. With such models, Butts (2019, 463) reports, "what a given neuron is selective to and how such selectivity is generated are essentially inscrutable." The argument of Emily Sullivan (2019), that the opacity of ANNs presents no obstacle to their providing scientific understanding, seems to ignore the issues being raised by the scientists themselves and to neglect the differences between machine learning and traditional modeling techniques.

25. See for instance, Breiman (2001), Shmueli (2010), and Yarkoni and Westfall (2017). Of particular interest for our study is Breiman's observation that the success of ANNs for prediction unsettles the traditional value placed in simplicity: "Occam's Razor, long admired, is usually interpreted to mean that simpler is better. Unfortunately, in prediction, accuracy and simplicity (interpretability) are in conflict. For instance, linear regression gives a fairly interpretable picture of the y, x relation. But its accuracy is usually less than that of the less interpretable neural nets." (2001, 206)

ing to see some of these scientists commenting on the historical fact of the interconnection between the tasks of making behavior intelligible by explaining it and making it controllable through prediction:

> The goal of scientific psychology is to understand human behavior. Historically this has meant being able both to *explain* behavior—that is, to accurately describe its causal underpinnings—and to *predict* behavior—that is, to accurately forecast behaviors that have not yet been observed. In practice, however, these two goals are rarely distinguished. The understanding seems to be that the two are so deeply intertwined that there would be little point in distinguishing them. (Yarkoni and Westfall 2017, 1100)

Among neuroscientists, we find Hasson, Nastase, and Goldstein (2020) noting the trade-off[26] and a radical lesson for reform of experimental methodology. They observe the same trend as discussed in chapter 5, from highly controlled, nonnaturalistic experiments in which a relatively small amount of data was collected over just one of the neuron's possible response profiles, to more naturalistic paradigms allowing the neural population to drift through a large range of its response states, and with recording techniques capable of producing many orders of magnitude more data. They describe how the older methodology went in tandem with a conception of the task of science as that of devising a simple predictive model describing the process by which the data were generated. Given the controlled conditions and paucity of data, scientists were aware that they were undersampling from a vast distribution, but the hope was that with the right theoretical insight, they

---

See also Napoletani, Panza, and Struppa (2011) for a discussion of the trade-off in various other disciplines, including molecular biology and meteorology. I do not expect that many of the issues discussed in section 8.3 are peculiar to neuroscience. I should also emphasize that predictive success within neuroscience of ANNs does not imply that these models are themselves brainlike. This is evidenced by the fact that ANNs achieve predictive success in these other disciplines in which claims of similarity to their targets would not make sense. ANNs are predictively powerful because of their huge number of parameters and access to massive amounts of data, combined with clever engineering techniques to avoid overfitting to nonprojectible patterns in the data (i.e., noise).

26. "As with any scientific model, neuroscientific models are often judged based on their interpretability (i.e., providing an explicit, formulaic description of the underlying causes) and generalization (i.e., the capacity for prediction over broad, novel contexts). However, in practice, interpretability and generalization are often at odds: interpretable models may have considerable explanatory appeal but poor predictive power, whereas high-performing predictive models may be difficult to interpret." (Hasson et al. 2020, 417, references omitted).

would hit upon the model that extrapolated accurately beyond the sample, predicting and explaining behavior in a range of nonexperimental conditions.[27] However, Hasson et al. (2020) argue that the failure of the older models to predict data from a broad range of conditions is an indication that this expectation was flawed. They advocate for a new methodology for the big data era, in which data are densely sampled over a wide range of conditions, and the ambition to devise simple, theoretically motivated models that extrapolate beyond the sampled regime is no more. Instead, machine learning is used to generate direct fit models that, with an excess of parameters, can mold themselves to the contours of any large data set without explosive overfitting of noise, thus yielding accurate predictions within a broader sampled regime. There is no attempt to build a general, explanatory theory as this new process "does not rely on explicit modeling of the overarching generative principles" (418).

With this trend of predictive power losing its moorings from understanding comes reason to wonder if the flagship successes of science seen in the last few centuries have been due to something of a coincidence: that there exists a delimited range of systems for which the predictively powerful theories and models are also intelligible to human scientists. Given these successes, it was assumed that the natural world just was that way: simultaneously intelligible to scientific reason and amenable to technological control. But it could be that there is nothing in the fabric of the world ensuring that understanding and predictive success should accompany one another. Instead, this is only to be expected for the few things in nature that are quite simple, or rather, simplifiable without a disturbing loss of predictive accuracy. Those quarries of things that are both understandable and predictable have been mined intensively for some time. The divergence of models for prediction and understanding is perhaps the first indication that a resource, always taken to be unlimited, is depletable after all. And if this raw material for science—things subjectable simultaneously to scientific understanding and manipulation—is limited and nonrenewable, it means that science itself has a boundary. It will not be capable of taking in all the natural world in one almighty gaze—as the cosmic iconography of science has always depicted it.

27. See Hasson et al. (2020, figure 1).

To set the limits of science at the outer edges of the intelligibility of nature, as du Bois-Reymond (1872/1874) did, is not to set any bounds on technology by itself. Neurotechnology, powered by gigantic deep networks running on supercomputers and guzzling data, seems to be quite capable of maintaining a steady, if not sprinting, pace of progress by itself, at least for some narrowly defined applications. Still, it is prudent to wonder about the negative implications of a decoupling of understanding and instrumentality. This is in fact something that the biologist Carl Woese raises as a deep concern: "A society that permits biology to become an engineering discipline, that allows that science to slip into the role of changing the living world without trying to understand it, is a danger to itself" (quoted by Callebaut 2012, 71).

However, it is tempting to brush away these worries, and the "end of science" forecasts, with a few objections to the argument presented here. For one thing, it is an exaggeration to say that advanced machine learning models like large ANNs are completely uninterpretable or will never lead to theoretical insights. The ways that ANNs, fully lodged into neuroscientific practice, are expected to afford understanding will be discussed shortly. Another thing is to point out that the discovery and use of means of intervening, without understanding, are not new in the neurosciences. All the major psychiatric drug classes were discovered serendipitously (Nutt and Need 2014; Berk and Nierenberg 2015). The important point here is that such inventions are often the inspiration for theory-driven research, such as the line of investigation into the dopamine hypothesis for schizophrenia. Indeed, this cycle of technological growth followed by theoretical development through reverse engineering is highly characteristic of traditional science, as discussed in section 8.2. Thus, next to the argument that machine learning methods presage the end of neuroscience, we find that there are equal grounds for an argument that neuroscience, facing no ineluctable barrier or epochal change to the terms of its existence, is to be expected to carry on more or less as before.

### 8.3.2 Understanding Redirected

There are high expectations for ANNs to generate progress in many domains of cognitive neuroscience, not only vison but also spatial navigation (Bermudez-Contreras, Clark, and Wilber 2020), and language (Lappin 2021), among other topics. Yet the state-of-the-art models do not yield understanding of the target system in the same way that previous generations of models did. This is largely due to the complexity of the ANNs—their very many

layers and connection weights—and the fact that the mathematical functions they use to fit experimental data are not hand-coded or selected by the modeler, but are arrived at by adjusting virtual connection weights over many iterations of training, remaining implicit in the connection weights of the trained network. The concern raised just now was that the predictive success of these models would force neuroscientists to abandon their traditional goal of understanding the brain, or at least that the practices conducive to understanding would have to be disentangled from those employing machine learning, leaving us with two sets of activities not recognizable by themselves as science.

However, in this section, I will discuss examples of research in neuroscience that fully incorporate the use of ANNs but do not give up on the goal of understanding. Instead, we will see that understanding is redirected toward challenges taken to be more feasible than the traditional one of charting the computations underlying cognitive performances and giving them a theoretical rationale, such as efficient coding. In these cases, scientists do not renounce the aim of finding simple—and hence comprehensible—principles in operation in the brain, as the end-of-theory prognosis would have it. There is a recurrence of some of the traditional theory-driven norms and practices, showing us that the old scientific pursuit of simplicity is ongoing in the era of deep learning–based modeling.

In this approach, an artificial proxy—an ANN—is still being used to gain understanding of the brain. But rather than seeking explicit representations of computations supposed to be shared between the artificial and biological neural network (BNN), for this version of the analogical strategy, the epistemic targets are *principles*, *design specifications*, or *constraints*, which are said to determine the operation of both classes of systems. Such principles would be simple enough to be understood, even if the details of processing within the ANN and BNN are not. For example, Hasson et al. (2020), whose negative argument was discussed in section 8.3.1, make the positive case that researchers should instead direct their efforts to the discovery of "design specifications" common to both artificial and organic neural networks, thus focusing their theoretical efforts on the understanding of network architectures, learning rules, and objective functions (423). In my two detailed examples, as we will see, the authors are clear on their rejection of the old ways that neuroscientists have sought simplicity, and through that, understanding of the brain. And, unlike some (e.g., Rudin and Radin 2019),

these authors are not denying that the lack of transparency due to the complexity of ANNs, compared with other classes of models, is significant.

First, the proposal of Lillicrap and Kording (2019) is a response to the entrenched idea that simplicity in the brain will be found through descriptions at Marr's level of computational explanation, with its abstraction both from the messy details of neurobiological implementation and the knotty strata of algorithm and representational formats. Computation-level descriptions would describe in elegant mathematical formulas the function transforming sensory inputs into commands generating intelligent, perceptually guided behavior. The validation of these theories would occur through the building of artificial systems, also implementing these computations, showing lifelike perceptual capabilities. This approach was exemplified in the classic experimental and theoretical work on the visual system, discussed in chapter 5, and was more recently cast as the search for *canonical neural computations* (Carandini and Heeger 2011; Chirimuuta 2014).

Machine learning has been the undoing of this approach not because ANNs like AlexNet and ResNet have "solved vision" (Serre 2019), but because they have dramatically outclassed the predictive accuracy of the simple, theoretically motivated models, leaving the idea that what the visual system is doing, in principle, could be represented by a handful of models of a few parameters each, looking naive if not foolish. Lillicrap and Kording pose the question "What does it mean to understand a neural network?" and their answer is intended to apply equivalently to organic and artificial networks. They write that "compactness is necessary for what we would call a meaningful understanding" (3). In effect, they are making simplicity of representation of a network—having a "compact description" (2) of it—a necessary condition for its intelligibility. Since the function learned by an ANN trained to classify photographed objects to humanlike performance levels will be embedded in the millions of parameters of the model, Lillicrap and Kording argue that unless this full description of the model could be compressed into a much more parsimonious description, understanding of how the network classifies an image cannot be obtained.[28] There is no reason to think that

28. Lillicrap and Kording (2019) do not specify how compact the description would have to be to make the network intelligible, although they do state that the number of parameters must be reduced by quite a few orders of magnitude since a network compressed to around 100,000 free parameters would still not be intelligible by their

an ANN with few enough parameters to be interpretable would give high performance with real-world data sets.

Although a compact description of the computations occurring in ANNs cannot be obtained, Lillicrap and Kording emphasize that the learning rules of the network are fully understood, the architectures are known, and in fact, the program for building a high-performing network like ResNet takes fewer than 100 lines of code. From this, they draw a lesson for neuroscience: for both ANNs and BNNs, simplicity is to be found in the learning rules and architectural principles. Those are feasible targets of understanding, as opposed to the hopelessly complicated and distributed patterns of information processing. Systems neuroscience, they argue, should give up on the long-standing ambition of building interpretable models that explain cognitive performances, and focus instead on theories targeting anatomy and plasticity rules.

Their argument for abandoning the former agenda of seeking computational-level explanations is worth discussing in detail. They point out that belief in the feasibility of this program is supported by an analogy from the success of statistical physics in finding compact, midlevel descriptions of systems containing countless particles.[29] Here, physicists employ simplifying methods such as the renormalization group (see Batterman 2018) to arrive at descriptions that are both interpretable and useful for prediction and control. Although bolstered by some successes of similarly coarse-grained neural models for narrowly defined processes, like calculation of eye movement direction, Lillicrap and Kording argue that the analogy between statistical physics and neuroscience soon breaks down:

> In the gas case, all atoms are the same, are exchangeable, and have short memory while in brains each cell may be unique and have a memory that effectively goes back to the birth of the animal. Moreover, the argument we made here suggests that such a compact mid-level model of computation can not have the property

lights (3). They also argue that standard methods for reverse engineering ANNs, such as visualization of the response profiles of some of the nodes, are insufficient for grasping how they classify an image (3). It is worth comparing their definition of intelligibility of models to that of Gao and Ganguli (2015), who, like de Regt (2009), follow Richard Feynman and Werner Heisenberg in stating that a model is intelligible if one can make qualitative predictions of its outputs without actually running through the calculations.

29. For instance, see Carandini (2012, 507).

> of actually working in the domains of brain performance where the environment can not be compactly communicated. Thus the analogy to physics may be misleading in the context of neuroscience. (2019, 5)

What is interesting here is the point that the physics analogy breaks down because in physics, the low-level, molecular details really do not matter for predicting the collective behavior, so it is safe to abstract away from them. This assumption does not hold in neuroscience since it is likely that the brain's functionality is due to its heterogeneity (which is to say, its complexity) at the neuronal and subneuronal levels. We may note in passing that ANNs themselves abstract away from that complexity, modeling a homogenous population of nodes lacking the physiological and biochemical characteristics of actual neuronal cells.

Cao and Yamins (2021b) share the notion of intelligibility set out by Lillicrap and Kording,[30] and like them concede that the absence of compact descriptions of the processing within neural networks sets up a barrier to intelligibility: "In the actual case of neural networks (whether biological or artificial), it may turn out that no efficient encapsulations of the relevant dependencies are available—either of how the system's behavior depends on inputs, or how its behavior changes in response to perturbations of the mechanism. We think that it is primarily this apparent characteristic of NN models that provokes critics to say that they are unintelligible" (3–4).

However, the point of their paper is to argue that there is a second form of intelligibility achievable for these systems, akin to optimality explanations in evolutionary biology. This comes about through discovery of the constraints that govern learning and optimization for these networks. They emphasize that both artificial and biological neural networks are kinds of functional systems—"a system that is the way it is for a purpose or a reason" (4)—and critically, that it is shaped by the demands on it to perform a particular task such as object classification. Since there are dependencies between form and function, with certain architectural features and representations being demanded for the most exacting tasks, Cao and Yamins argue that functional considerations will yield explanations of the form of these networks (11). Moreover, they make the case that a task such as object

30. "One well-recognized way in which a system can be made intelligible is through the discovery of a concise mathematical description of its performance" (Cao and Yamins 2021b, 15).

classification is so difficult that solutions will be highly constrained—there will *not* be more than one way to crack the egg—meaning that knowledge of the form taken by an ANN that solves the problem will give insights transferrable to the case of the primate visual system (16).[31]

It is interesting, finally, to note that like Hasson et al. (2020), Cao and Yamins are critical of the old experimental methods of making neural systems intelligible by using highly controlled, "reduced" behavioral tasks:

> It does not make sense from the optimization perspective to choose the most reduced version of a given task domain and then seek to thoroughly understand the mechanisms that solve the reduced task before attempting to address more realistic versions of the task. In fact, this sort of highly reductive simplifying approach is likely to lead to confusing results, precisely because the reduced task may admit many spurious solutions. (2021b, 19)

Their point is that by using simplified tasks, one makes a task easy to perform, so that the range of network forms that can solve the task is not highly constrained. Without sufficient constraint, transference of knowledge of an ANN solution to a BNN is not valid, so the research program premised on the assumption that both classes of networks converge on the same solution (the only peak in the fitness landscape—see Cao and Yamins 2021b, figure 3), and hence take the same form, cannot be carried out. Thus, they endorse the recent trend, discussed in chapter 5, of making tasks in experimental neuroscience lifelike and challenging.[32]

---

31. But see Sinz et al. (2019, 971) for the contrary opinion:

> There is probably a very large group of networks, our visual system included, that can solve single tasks such as ImageNet, but they might use vastly different solution strategies and exhibit quite different robustness and generalization properties. This implies that our current datasets, even though they contain millions of examples, simply do not provide enough constraints to direct us toward a solution that is similar enough to our visual system to exhibit its desirable robustness and generalization properties.

32. One should object here that it is too quick to assume that simplified tasks are easier, and therefore less constrained. Simplified tasks of classical neurophysiology are often not easy in the sense that they do not come naturally to the animal doing them. The fact that they are nonethological means that they are hard for the animal to learn to perform. It can be surmised that the notion of easy/difficult being invoked by Cao and Yamins is one tracking the size and complexity of the data set that the network is being trained on. The worry here is that this does not necessarily map onto the division between reduced and ethological experimental paradigms.

We may conclude from these examples that inclusion of machine learning methods within neuroscientific research does not force scientists to abandon the aim of understanding. Even when conceding that old targets may be permanently out of reach, the understanding of neural networks in computers and in brains is pursued in different modes. This serves as a counterargument to the prognosis in section 8.3.1 that neuro*science* is coming to an end because machine learning is driving off from researchers' agenda the goal of understanding the brain. At the same time, the material in this section does not undermine the previous argument that there is a limit to neuroscience, in that there is a limit to what may be understood about the brain. Those proposals discussed here have all conceded that because of the complexity of the brain, there is a barrier to understanding the neural processes underlying intelligent behavior. This limit has been made apparent because even ANNs, though nowhere near the complexity of actual brains, in order to become sophisticated enough to reach performance levels comparable with organic systems, have had to grow in size and intricacy beyond the range of what is humanly interpretable.

### 8.3.3 Understanding Redefined

The third of my forecasts about the reshaping of neuroscience through the incorporation of machine learning focuses on the way that the technology will change what is meant by understanding the brain. One way to describe the reorganization of terms that occurred at the origin of science, according to Dear's account, is to say that the natural, philosophical, and contemplative notion of understanding was adjusted to incorporate a new technically driven spirit. Whereas previously the notions of knowing and understanding nature were not linked with instrumental control, definitions changed such that the ability to predict and manipulate things was taken as a sign of one's having understood them. The impact of this shift was profound. The most common argument today for scientific realism, the "no miracles argument," rests on the claim that without science having grasped some deep truth of nature, the fact of predictive and technological success would be miraculous (Psillos 1999, chapter 4; Dear 2005, 404). Here, I will discuss a case of ANNs being employed in neuroscience as a form of automated science and show how it suggests a yet more radical redefinition of understanding along instrumentalist lines—one consistent with the ideals of a purely empiricist epistemology of science.

Bashivan, Kar, and DiCarlo (2019) report an experiment in which an ANN is built to perform the task of receptive field mapping for neurons in area V4 of the primate ventral stream. Of the studies discussed in this chapter so far, this is the best example of machine learning as automated science because the ANN is doing work normally carried out by a human experimenter and theorist. The purpose of the study was to build encoding models of V4 neurons' response profiles that would predict their activation to any arbitrary stimulus, and then to use these models to generate novel stimuli that maximally activate the neurons, constituting an effective "control handle" for neural activations.[33] Whereas in science without automation, the experimenter would select stimulus classes and the theorist would write down or program the encoding model, this study shows how these tasks can be offloaded to an ANN trained on a large set of labeled natural images. The impressive thing about the study is that the novel images synthesized by the network allow a much greater degree of control of neural activity in V4 than by any former methods of stimulus selection. The downside is that this control is not accompanied by a gain in understanding of V4 neurons' response profiles, in any traditional sense.

One of the purposes of the study was actually to address criticisms that their previous ANN models of the ventral stream did not generate understanding (Bashivan et al. 2019, 1). In response, their intention was that "at least one ANN model can be used to noninvasively control the brain—a practical test of useful, causal 'understanding'" (11). This is a very interesting remark because it in effect lays down an operationalist redefinition of scientific understanding. From a cognitive meaning—a sense of comprehension or ability to grasp the principles of a thing—we have an entirely noncognitive sufficient condition: *a neural system is understood if it can be controlled*. This is not merely a verbal ruse since it is quite consistent with a long-standing empiricist conception of science. Carl Hempel (1965, 337), for example, defined understanding a phenomenon as the ability to predict it: "The [deductive-nomological] argument shows that, given the particular

33. A comparable study by Walker et al. (2019), mentioned in chapter 5, uses ANNs to find the maximally excitatory stimuli for mouse V1 neurons. In that study, the ANN is initially trained on prediction of neuronal responses rather than on classification of images.

circumstances and the laws in question, the occurrence of the phenomenon *was to be expected;* and it is in this sense that the explanation enables us to *understand why* the phenomenon occurred" (italics in original, quoted in de Regt 2009, 23–24).

Notions of scientific explanation and understanding that referred to people's feelings and cognitive states instead of hard metrics like predictive success were treated with suspicion by twentieth-century empiricists because they threatened to taint science with subjectivity (de Regt 2017, chapter 2). The instrumentalist emphasis of Bashivan et al.'s proposal is also reminiscent of this tradition of philosophy of science. Instrumentalism is, of course, normally opposed to scientific realism since it asserts that the aim of science is not the discovery of theories that represent nature as truthfully as possible, but rather the discovery of theories that allow prediction and control. Mieke Boon (2020, 52), likewise, has made the observation that the data-driven principles of automated science are empiricist ones and, conversely, that if one adheres to an empiricist epistemology of science, one has no grounds to hold that data-hungry machines will not eventually be better scientists than human beings.[34]

Given the historical theme of this chapter, the point I would like to emphasize is that the rise of machine learning in neuroscience is consistent with a notion of the development and ideal trajectory of science that can be derived from the work of Ernst Mach, one of the founders of empiricist philosophy of science.[35] For Mach, the point of scientific concepts was prediction, and thereby the ability to achieve material results, with maximal economy of thought (Mach 1882/1895).[36] Thus, the goal of science is specified by Mach to be an instrumental one. Importantly, Mach recognized that science had its origins in natural philosophy, but he held that the progress of science consisted of its gradually purging itself of its metaphysical legacy, in particular

34. See Hooker and Hooker (2018) for the argument that the achievements of automated science make the case for empiricism over scientific realism, and Buckner (2018, 2023) on the link between deep learning and empiricist philosophy of mind.
35. Mach was also an inspirational figure for twentieth-century logical empiricists like Hempel (Stadler 2021).
36. "As it is usually understood, that doctrine [of the economy of thought] holds that scientific laws and abstract class terms are tools for compiling and organizing experience by means of the fewest possible concepts, a mastery that is useful for the prediction and control of events." (Banks 2004, 23)

the notion of material substance.[37] Mach rejected the question of ontological commitment to the terms of physical theories, such as "force" and "atom," which he regarded instead as conceptual tools of science (Mach 1883/1919, chapter IV). This results in a picture, in opposition to Dear's one, in which *only* instrumentality properly belongs to science, and natural philosophy has no place. To the extent that the incorporation of machine learning into science leads to an attenuation of nonempirical concerns with understanding, as normally defined, the development is a progressive one.[38]

It is to be noted that Mach at times defined science in such a way that makes it amenable to automation through machine learning, writing that "science itself, therefore, may be regarded as a minimal problem, consisting of the completest possible presentment of facts with the least possible expenditure of thought" (Mach 1883/1919, 490).[39] The connection here is that in Mach's view, the world consists only of elements, which we might analogize to data points, and the point of science is to represent them as compactly as possible so predictions can be made about the outcome of events—the unfolding of new data. In Mach's day, the best means for that was the numerical representation of data observed via the senses and the formulation of quantitative theories (laws of nature) describing the relationships among data, thus affording predictions. But more generally, the task is to discover the best methods for data processing—best in the sense of most efficiently representing the data, for purposes of prediction and control. The automated production of classifiers and predictors for large data sets, through machine learning, does exactly

37. See Skidelsky (2003, 367): "The development of modern science, writes Mach, 'consists in the fact that the original, naïve concepts of substance (*Stoffvorstellungen*) are recognised to be unnecessary, that we acknowledge real constancy and substantiality to lie in discovered quantitative relations, expressed in the fulfilment of equations, and do not seek some 'lump' outside of thought.'" However, I grant that the question of what way Mach was opposed to metaphysics is itself rather complicated (Guzzardi 2021).

38. Mach (1872/1911, 55–56) equates understanding with "analysis alone"; that is, reduction to fundamental facts (quoted in Guzzardi 2021, 176). This stands against the notion of making nature intelligible, employed in this chapter (Dear 2006), and against the *meaning-making* notion of understanding that I will develop in section 9.5 of chapter 9.

39. However, see Patton (2021, 152, 155). Taken out of context of Mach's other writings, focus on this passage can lead to an overly narrow view of his philosophy of science. The ultraempiricism that I am aligning here with automated science is one strand of Mach, but more of a caricature than a true portrait. I thank Richard Staley for pressing me on this point.

that. Indeed, the fact that the task of processing all that data can be offloaded from the human mind to the machine allows even greater efficiency. Mach observed that mathematics, even when done with pencil and paper, is a way to release scientists from mental labor, and he looked forward to the possibility of this being fully accomplished with mechanical computers:

> A total disburdening of the mind can be effected in mathematical operations. This happens where operations of counting hitherto performed are symbolised by mechanical operations with signs, and our brain energy, instead of being wasted on the repetition of old operations, is spared for more important tasks. . . . The drudgery of computation may even be relegated to a machine. Several different types of calculating machines are actually in practical use. The earliest of these (of any complexity) was the difference-engine of Babbage, who was familiar with the ideas here presented. (Mach 1883/1919, 488)

The paradox here is that with progress, science becomes more and more thoughtless. *Thinking* about nature, attempting to understand it, is the lumbering weight of the metaphysical legacy of science. The promise of machines is not that they will do our thinking for us, but that they will speed up the process of eliminating thought from routine science.

Yet, the very thoughtlessness of ANNs, revealed in their being pressed into making misclassifications indicating, for example, that *they really have no idea what a cat is*, is what for many prohibits them from being reliability used as autonomous scientific agents.[40] The significance of these limitations of ANNs, as clues about something different going on in the lives of our minds, will be discussed in chapter 9. For now, I will conclude that this form of AI presents us with a pure embodiment of empiricist principles in philosophy of science. To the extent that we find them wanting, that we are dissatisfied with their capacity to do science by optimizing for prediction and control while at the same time redefining or removing understanding from the agenda of science, we are also committed to critique of those empiricist principles.

We have now worked through three lines of analysis concerning the implications of machine learning for our conception of neuroscience, and

40. I am alluding here to the vulnerability of deep convolutional neural networks (DCNNs) to adversarial attacks (misclassifications induced by spurious changes to images), which will be discussed further in section 9.5 of chapter 9. See Buckner (2020) on the problem of adversarial vulnerability for automated science. Of course, deep learning is a fast moving field, and there is currently an unresolved debate over whether a different class of ANNs, *transformers*, does demonstrate understanding.

to some extent for science more generally. Each of them has something to recommend it. The first makes clear a limit of science in the complexity of the brain. The second shows how the relationship between instrumentality and understanding can evolve, and neuroscience can continue more or less as before in spite of acknowledgment of this limit. The third option reveals the deep connection between automated science and the ideals and tenets of empiricist philosophy of science. One word of warning about the second option is that the proposals for redirecting understanding are very programmatic, and it remains to be seen if these new explanatory projects can succeed. What we may conclude is that, as a matter of research interest, neuroscientists have not given up on trying to understand the brain. But we cannot presume at this point in time that brain will be understood in these different ways. It may be that the limitations on neuroscientitific understanding of the brain, as described under the first scenario, are more complete than these investigators have hoped.

## 8.4 The Fixed Net of the Machine

> *Our intellect, when it follows its natural bent, proceeds on the one hand by solid perceptions, and on the other by stable conceptions*. It starts from the immobile and only conceives and expresses movement as a function of immobility. It takes up its position in ready-made concepts, and endeavors to catch in them, as in a net, something of the reality which passes. This is certainly not done in order to obtain an internal and metaphysical knowledge of the real, but simply in order to utilize the real, each concept (as also each sensation) being a *practical question* which our activity puts to reality and to which reality replies, as must be done in business, by a Yes or a No. But in doing that, it lets escape that which is the very essence of the real.
>
> —Henri Bergson (1903/1912, 66–67, emphasis in original, translation modified)

This chapter opened with an examination of the leading role of technology in science through the making of devices that serve as model systems and offer an understanding of the forces and processes, presumed to be at work equally in artifactual and natural things. This was followed by an exploration of the consequences that may arise in neuroscience once the machines take on some of the inscrutability of things not made. We saw that there are limitations to scientific understanding revealed by the complexity of deep learning models, but neuroscientists hope that the task of understanding can be successfully redirected to new theoretical questions. Another hope that

might be entertained about the use of such models is that, with their lifelike complexity, they will overcome any scruples about the gulf between organic and machinic forms of information processing (Bongard and Levin 2021). In other words, through the creation and reverse engineering of brainlike machines, scientists may surpass the limit marked at the end of section 8.2, that of the brain itself being inaccessible, only approximately grasped through the medium of artificial analogs that are in too many ways unlike it.

The thought is that with the building of such complicated models, as different as they are with respect to material substrate, the neuroscientist has at least ceased to rely on the simplifications imposed by the traditional search for models that expressed interpretable theories of neural processing. Even though the ANN is still an artificial proxy for the actual brain, it most importantly lacks the idealizations of previous models needed to keep their mathematics workable, such as assumptions about linearity. Moreover, its design is somewhat brain-inspired (see section 5.1.3 in chapter 5), and like a real neural network, it is self-organizing, not tailor-made to do a task, discovering solutions to classification problems through its own iterative learning procedures. Although learning algorithms such as backpropagation are not biomimetic, at least the principle of an adaptive system is. ANNs have the attraction of, on the one hand, attaining human-level performance on some tasks, such as rapid object and face recognition in natural scenes; and, on the other hand, being far more accessible than the actual brain. The architecture and training process (learning rules and developmental history) of the ANN are fully known to the researcher, so the hope is that the many things that may be discovered about its operating principles could reasonably transfer to neural networks in the brain.

According to the view, argued by Cao and Yamins (2021a,b), that ANNs and BNNs hit upon interchangeable solutions to well-constrained, difficult tasks, the ANN appears not as a model of a brain area—with the gap or distance between representation or target that this entails—but as another member of the same class of information-processing systems. What is more, the focus on these proposed similarities encourages the thought that the brain itself is a kind of adaptive model: a model of the sensory data delivered to it in the course of life. This is the view of Hasson et al. (2020, 423), who write that in spite of the substantial differences between them, ANNs and BNNs "belong to the same family of direct-fit models," whose design specifications are simple and therefore intelligible. That said, I have since chapter 4 of this

book argued against temptations of this sort, to neglect the dissimilarities between models and things modeled and to bank on the irrelevance of the details over which they differ. This argument, therefore, needs to be restated.

I will begin with a quotation from a group of neuroscientists referred to in chapter 5 on the utility of ANNs as "model organisms" in experiments using complex ethological tasks. They give some words of caution: "One large difference is that ANNs are trained only once during an optimization process and the connection weights are not subsequently modified, while animals continually update and refine their behavior. This discrepancy seems fine for understanding 'instantaneous snapshots' of animal behavior but is highly problematic for understanding how animals learn or how their neural representations evolve over time" (Musall, Urai, et al. 2019, 235)."[41]

In essence, what is different about the brain is its tendency for continual adaptation and learning from new experience. This should be taken in relation to chapter 7's depiction of the Heraclitean complexity of the brain. The brain is an organ in a living body, made up of metabolizing cells, such that neural tissue is processual and never fixed; it must always be changing itself just to maintain whatever functionality it has. Plasticity is ever-present, not restricted to a developmental phase. Thus, the continual refinement of behavior observed in animals, the finding that learning is continual and that there is no set demarcation between training phase and performance, is probably due to the nervous system's inherent tendency for material modification, and hence plasticity. Machines, including digital computers running ANNs, have a different nature. They are made of nonliving components that are not inherently self-modifying. Moreover, machines—even ones like deep learning classifiers that have some self-organizing characteristics—are built to do a task and to keep doing it in the same way once built. That is why the division between learning and task operation is clearly demarcated for these devices. They are fixed nets, yet with them, the neuroscientist tries to capture, conceptually, the fundamentally changeable networks of the brain.

41. Cf. Kell and McDermott (2019, 128): "The most fundamental difference between current DNNs [deep neural networks] and human perceptual systems may lie in the relative inflexibility of artificial networks—a trained network is typically limited to performing the tasks on which it is trained. Representations learned for one task can transfer to others, but usually require training a new classifier with many new training examples."

Attempts are being made to engineer ANNs to be more brainlike in this respect—to achieve what's known as "lifelong machine learning,"[42] which is an ironic label since these devices are never alive and lack the life trajectories without which the developmental phases of animal cognition make no sense. As successful as they are at tasks for which they are trained, deep neural networks do not learn from new data without this disrupting performance on the previously trained task. They are liable to "catastrophic interference," which in the worst case involves the complete overwriting of previously learned classifications (Parisi et al. 2019, 55).[43] Needless to say, engineers are putting effort into the challenge of making more truly adaptive ANNs, with various solutions reviewed by Parisi and coauthors, such as the supplementation of new resources for the network when a new distribution is to be learned and regularization techniques constraining how the weights between nodes will be updated, not to mention the brain-inspired method of "generative-replay"—the reactivation of memory patterns (van de Ven, Siegelmann, and Tolias 2020). The potential for these fixes to increase machine adaptability, however, does not undermine the more fundamental point that the difference between actual neural networks and ANNs lies with the fact that in the brain, the processes associated with learning and with doing are not cleanly separable. Musall et al. (2019, 235) note that "biological brains implement both the computation underlying behavior as well as the system that enables learning of novel behaviors," whereas "ANNs . . . use externally available cost functions and optimization routines, typically written as auxiliary software, which are discarded after training."

42. Parisi et al. (2019, 55) define this property as "the ability to continually learn over time by accommodating new knowledge while retaining previously learned experiences is referred to as continual or life-long learning. Such a continuous learning task has represented a long-standing challenge for machine learning and neural networks and, consequently, for the development of artificial intelligence (AI) systems." Lifelong learning is a well-recognized, important capability of humans and other animals.

43. Another problem is that the training regime of current deep neural networks assumes a fixed data distribution, with training data sampled from it. The ANN cannot gracefully adapt to changes in the data distribution. In a more "ecological" training regime, in which new kinds of samples become available over time, it is found that performance on previously learned classifications decreases as the new ones are learned (Parisi et al. 2019, 55).

My disagreement with them over computational literalism—their description of the brain as implementing computations—does not undermine the point I want to draw from them. For it is certainly a truism about animal development that learning and performance have to take place concurrently. It is only in the controlled conditions of the psychology laboratory that the two tasks become demarcated, and these artificial scenarios are the ideal setting for machines in which workings of training and doing are modularized. We saw earlier in this chapter that Lillicrap and Kording (2019) propose that learning rules should replace canonical computations as a priority for theoretical neuroscience (section 8.3.2). Their reasoning is based on the finding that in machines, learning procedures are transparent (because explicitly coded), and therefore far more intelligible than the classification operations in the trained network. But given the relative lack of separation between processes for learning and performing behaviors in the brain, why should one bet that learning in brains will be more intelligible than other operations? Lillicrap and Kording's proposal is symptomatic of a failure to attend to what is different between these two kinds of networks.

Thus, we can conclude that a limit to the understanding of the brain by way of building brainlike machines is still in place. Adaptability, the ability to learn, is the basis of intelligence in humans and other animals. The form that animal learning takes is conditioned by the inherent plasticity of neural tissue, which is in turn derived from its being made of living cells. What is more, the form that animal learning takes is shaped by the learner being an organism, something self-driven, with its own life trajectory, needs, and motivations. It is doubtful that this mode of learning, which supports flexible behavior in uncontrolled circumstances, can be captured in nonliving machines even if machine learning can match animal learning for a variety of prespecified tasks. It is true that learning in the artificial network is easier to understand than its trained classification procedures, but it would be a mistake to assume this to be the case in living systems in which adaptation is not a discrete stage in an engineer's flowchart, but a mode of being.

# 9 Revisiting the Fallacy of Misplaced Concreteness

> We believe that men and other animals are like machines from the scientific standpoint because we believe that the only fruitful methods for the study of human and animal behavior are the methods applicable to the behavior of mechanical objects as well. Thus, our main reason for selecting the terms in question was to emphasize that, as objects of scientific enquiry, humans do not differ from machines.
>
> —Arturo Rosenblueth and Norbert Wiener (1950, 326)

> I am open to the idea that a worm with 302 neurons is conscious, so I am open to the idea that GPT-3 with 175 billion parameters is conscious too.
>
> —David Chalmers (2020)

## 9.1 Revisiting the Literal Interpretation

In the preceding chapters, I have presented neuroscience as a discipline, like any other in science, that is reliant on certain modes of dealing with its objects, which project things onto a higher plane of simplicity. Attendant to this, I have urged caution in the interpretation of the simplified products of neuroscientific research. To say that the cortex is literally computing the function stated in a model is, I argued in chapter 4, to identify two fundamentally different kinds of things. It is like pouring a gallon of water over an egg cup and convincing yourself that the overflow is irrelevant to the makeup of the gallon. And yet the literal interpretation of computational models is the dominant one within theoretical neuroscience and among technologists and naturalistically inclined philosophers. In chapter 8, we saw that this quasi-consensus makes sense, given that the centuries-old methodology of using artificial workings as exemplars for natural processes has tended to elide the

distinction between the products of technology and objects not made by human hand.

This chapter is about the negative consequences, for philosophy, of the literal interpretation of computational models. The *fallacy of misplaced concreteness* is the mistake of taking the abstractions of science for concrete reality, confusing the model with the target, the map with the territory (Whitehead 1925/1967, 51–55). The literal interpretation is guilty of this, but the error within routine science is tolerable. There is an efficiency gain in assuming that the research object is as the streamlined model, not the concrete, ungainly thing. A cost comes from the presence of unknown unknowns and in keeping the door closed to possible discoveries. Still, the costs are offset by the pluralism of science and the likelihood that another researcher, with a different kind of model, will open a new door. Naturalistic philosophers of mind look to science to tell them about the natures of those objects of research—what *is* memory, vision, understanding, and consciousness? Their concern is not with how things are with the model—which they take too often to be a transparent, undistorting medium—but in our brains and cognitive processes. If the fallacy of misplaced concreteness occurs here—if an abstraction of a model, imposed by practical necessity, is mistaken for a discovery about how the brain *is*—then this brings about an irredeemable flaw in the logic of the inquiry.

In this chapter, I will argue that popular philosophical views about the potential of machine intelligence are the product of the fallacy that comes with the literal interpretation. The current, dominant paradigm of AI research builds models that are loosely "brain inspired," with the goal of engineering humanlike cognitive performance on predefined tasks such as object recognition, language production, or game play. Many of these expert systems, which are artificial neural networks (ANNs) trained to excel at one such task, have achieved superhuman abilities. A noted example that brought deep reinforcement learning to the attention of the world a few years ago is AlphaGo, a creation of DeepMind, which unexpectedly beat the world grandmaster 4–1 in its first tournament outing. This program was outdone by the even more powerful AlphaGo Zero (Silver et al. 2017), with both versions playing Go beyond a level that a human could ever now hope to attain. At the same time, ANNs have demonstrated a range of surprising failures seeming to stem from their expert nature, which is to say their lack of general intelligence, basic common sense. The question is whether the current, silicon-based machine

learning technology, when scaled up to some sufficient network size, will deliver humanlike general intelligence, with the characteristics now lacking in current AIs.

Expectations of artificial general intelligence (AGI) have been high, but as I will argue, ill founded. The failures of ANNs to obtain the capacities associated with general intelligence, such as sentience and the ability to apply learned knowledge to fundamentally novel situations, are not surprising if one considers the analogical interpretation of computational models proposed in chapter 4, which I will summarize again here. My view is that computational models of neural systems should not be interpreted as approximately true representations of computations actually performed by the brain. Even though the brain is not, literally, an evolved computer, such models earn their keep by providing a framework for a drastic abstraction away from the complexity of neural systems. The computational framework excuses an otherwise unjustified separation between the entities and processes within the brain said to carry out information processing, with the remaining systems classified as metabolic support.

ANNs employed for machine learning today are founded on the work of much earlier theorists. For example, McCulloch and Pitts (1943) brought a crucial abstraction to neuroscience by positing that a neuron, qua information processor, is simply an input-output device that takes a weighted sum of impulses coming in from its dendrites and sends a message to next-layer neurons through its axon. Rosenblatt (1958) combined the ideal of the artificial neuron, with simplified models of learning inherited, via the cell assembly theory of Donald Hebb (1949, 1960), from behaviorism. At a very high level of abstraction, the nodes in ANNs are somewhat like actual neurons, and neural plasticity is in part somewhat like the adjustment of connection weights within an ANN. Whereas the literal interpretation proposes that the abstracted commonalities are computational structures and rules implemented both in the brain and artificial hardware of the model, I decline to infer that these exist in the brain independently of its representation by the modeler as a computing system. Instead, I hold that an *ideal pattern*—massaged out of the neural system through the methods of experimentation and data handling, abstraction, and idealization—is what is actually represented and instantiated in neurocomputational models.

In other words, the analogical interpretation declines to think of the abstracted structure depicted in the model as a real, inherent feature of

the concrete system. With this, our attention is drawn to the disanalogies between brains and computers, and we consider the computational model to be no more than a lens, not transparent and somewhat distorting, through which neural data are systematized and made intelligible to the researcher. The obvious possibility that then comes to mind is that it is those characteristics of animals and their nervous systems not shared with computers, that are responsible for those capacities of general intelligence, capacities not replicable in machines according to indications available so far.[1] This chapter will focus on two of these features—phenomenal consciousness and understanding.[2]

Section 9.2 will discuss, using the example of visual object recognition, how computational explanation leads to an explanatory gap: ANNs offer explanations of object recognition, but not the visual awareness that goes with visual classification in a range of animals under normal viewing conditions. In section 9.3, I explain why we should not expect ANNs to scale up to consciousness. Next, in section 9.4, I show that philosophical arguments for the in principle possibility of consciousness in organic machines are grounded in a fallacy of misplaced concreteness, in their assumption of the equivalence of brains and hypothetical silicon-based "isomorphs." In effect, mine is a new argument for the position known as "biological naturalism," the view that inorganic machines cannot become conscious.[3] Then, in section 9.5, I move

1. That is my assessment of the current situation. Attention-grabbing claims for progress toward AGI (e.g. Bubeck et al. 2023) rest on notions of ANNs showing emergent capabilities, a notion that is problematic (Schaeffer et al 2023).

2. By "phenomenal consciousness," I mean qualitative experience—that there is something that it is like to be an animal with phenomenal consciousness. I note that self-consciousness, awareness of oneself as a being who experiences things, is another important characteristic of general intelligence, but it is not under discussion here. Understanding has been far less widely treated in the philosophy of mind, although the article "Minds, Brains, and Programs" by Searle (1980) is one instance. In cases like Searle's, understanding can be taken as the capacity not only to manipulate symbols, but also to appreciate that those symbols have meanings and are part of a nexus of meanings with referents beyond the subject. Furthermore, the capacity of understanding is indispensable for our sense of having a coherent world around us. Although "understand" is a success verb, I treat understanding as the task of continually trying to make sense of the world, often not achieved (section 9.5). In my view, systems that do not even attempt to understand should not be considered intelligent, properly speaking.

3. The term is due to Searle (1992, 1), who defines "biological naturalism" as the view that "mental phenomena are caused by neurophysiological processes in the brain and

to the topic of understanding, a feature of human intelligence exemplified in scientific attempts to make sense of objects and processes. Continuing the line of thought started in chapter 8 (section 8.4.3), I hold that discussions over the potential of automated science must acknowledge the obstacle posed by the lack of understanding in ANNs. And with a picture of the difference between human and machine intelligence in place, we have a better basis to say why understanding is indispensable and why machines may be expected to lack it. Finally, section 9.6 presents the conclusions of the chapter.

## 9.2 Locating an Explanatory Gap

> The questions which can be asked concerning this phenomenon in a theoretical brain model (where we are not free to assume any intrinsic similarity of processes to those in the human brain) are questions of what can be discriminated, "seen," "attended to," or "remembered" under specified conditions. All that we can say, in the last analysis, is that the system acts as if it were conscious, leaving the question of the actual existence of consciousness in the system for metaphysicists to consider.
>
> —Frank Rosenblatt (1962, 66)

When I flick through a photo album, looking at the people I know, I can put a name to the faces, and in doing that, I have a visual experience of those faces—there are colors and configurations and textures that show up in my awareness. This experience is what I will refer to as *visual phenomenal consciousness*.[4] All our sensory modalities come with phenomenal consciousness, though that is not to say that we are conscious of all that we perceptually discriminate or that we do not sense anything when in a nonconscious state, such as deep sleep. There are now numerous computational models of perceptual systems offering explanations of the detection and discriminatory capacities of humans and other animals. But they are silent about the phenomenal consciousness that accompanies perception

---

are themselves features of the brain." However, as the term is used in the debate over AI consciousness (see section 9.4), it simply means the view that consciousness cannot be achieved by a nonliving artifact, such as an electronic computer. Hence, my use of the term does not subscribe to the concrete claims that Searle puts forward here about the brain—by itself—being the *cause* of mental phenomena.

4. This is also known as the "what it's like" of visual perception. I use the word "sentient" to mean the kind of creature that can have phenomenal consciousness, regardless of whether there is self-consciousness.

in humans and, we may presume, many other animals. There are now numerous machines that can make certain visual classifications equivalent to a human's, correctly attaching names to faces. And yet they have no awareness accompanying those discriminations.[5] This section is about the significance of this silence and this difference.

In philosophy, the computational theory of mind asserts that cognitive processes (including perception, motor control, and affectivity) are essentially computational processes (Sprevak and Colombo 2019).[6] To be a cognitive creature is to have a brain that implements computations honed by natural selection to support intelligent behavior. This theory is a popular way to naturalize the mind—to show how mind is the result of ordinary physical occurrences—by positing that just as a manufactured computer is a lump of matter orchestrated in such a way as to perform cognitive feats like logic and arithmetic, so the brain supports cognition through the assemblage of its material parts into a computational system. Animal intelligence can thus be accounted for as no more mysterious than the workings of any elaborate machine. It is assumed that consciousness is the result of some of the particular kinds of computations occurring in the brains of those creatures that have consciousness, and if the computations for consciousness were discovered, they could in principle be implemented in a machine, resulting in a system with the same form of awareness as the animal.

The computational theory of mind pairs with the literal interpretation of neuroscientific models: it is asserted that the brain is literally performing computations and that these account for an animal's cognitive capacities. A task for neuroscience, therefore, is to find out what those computations are by presenting them, to as close an approximation as possible, in a model. Such discoveries yield not only computational explanations of cognition, but also the possibility of replicating cognition in machines. In chapter 4, I contrasted the literal interpretation of neurocomputational models with

5. This is the most reasonable assumption to make about these systems, although some may dispute it.

6. Piccinini (2020, chapter 14) makes a distinction between the computational theory of *cognition* and the computational theory of *mind*—the second includes consciousness within its explanatory scope, and the first does not. On his view, only the nonconscious capabilities of the mind are fully explicable through computational models, whereas the conscious features may be medium dependent and therefore not open to computational explanation—a view that I am sympathetic to.

my preferred analogical one. I attended to the fact that neurocomputational models, like all scientific models, are abstractions whose value to the researcher lies in their suppression of details present in the concrete system but not relevant to the scientific task at hand. Thus the brain, under experimental conditions, with data analyzed in certain ways, yields patterns of activity like the dependency relationships presented in the model (the *ideal patterns*), and one can say for this reason that the brain is somewhat like a computer, but we cannot say that it *is* an evolved computer.

Abstraction is the process of leaving out details in a representation of the target under investigation. Occasionally, the very details left out of the abstract representation may be critical to another task or relevant to the explanation of a different phenomenon. This is the problem with consciousness, so I will argue. In chapter 4, I accounted for the wide uptake of computational models in neuroscience as being due to the computational framework licensing a powerful abstraction away from neurobiological particularities not shared between computers and brains. This permits neuroscientists to ignore countless anatomical, biochemical, and physiological details when offering computational explanations of cognitive processes. Computational models of processes like visual perception are abstractions in the sense that they purposefully disregard the complexity that is present in the nervous system stemming from the fact that it is made of living tissue. The computational approach comes with a commitment to establishing a correspondence between visual processing in nonsentient machines (i.e., digital computers)[7] and sentient animals.[8] Therefore, the explanatory framework must abstract away from the consciousness that an animal has and the computer lacks.[9]

7. That is, actual machines, not hypothetical future machines that some have argued might become conscious (see section 9.4). The point is that the analogy source is the machines familiar to scientists today, which are clearly not conscious.

8. It is taken as read, for the purposes of this argument, that the cognitive processes being modeled are ones accompanied by consciousness in the animal, and that all the animals in question are sentient, which is indisputable for models of perception in humans and other mammals. This is not to say that there are no models of the cognitive processes not accompanied by consciousness, or of ones that occur in animals that may be nonsentient.

9. Bridewell and Isaac (2021) make a similar observation, but draw out a positive lesson. They propose that this dissimilarity between animals and computers can be leveraged to increase scientists' knowledge of where consciousness is and is not relevant to cognition, which they call an "apophatic" methodology for consciousness

This, I contend, leads to an explanatory gap for the computational approach that is puzzling under the literal interpretation but completely predictable under the analogical interpretation. A helpful consequence of the analogical interpretation is that it allows us to diagnose why this explanatory gap appears.

I will now state the argument and then provide more details of the example. The first point is that computational models of the brain are powerful abstractions, permitting neuroscientists to ignore countless biological details. With the analogical interpretation of neurocomputational models, one expects as many disanalogies as analogies between brains and computing systems. An important disanalogy between brains and computers is that brains are an organ of sentient animals,[10] whereas no computing systems are conscious. On the literal interpretation, however, computation is the essence of the brain's cognitive capacities. One ignores disanalogies, the features not shared between brains and computers, as irrelevant to the explanation of cognition. For example, deep convolutional neural networks (DCNNs) can achieve human-equivalent performance in the task of *core object recognition* (COR),[11] but without the visual phenomenal consciousness found in primates. This yields a computational explanation of COR, but on the literal interpretation of the model, there should be puzzlement that the model cannot explain, and has no potential to explain, why or how there is visual phenomenal consciousness in animals performing this task. In contrast, from the analogical interpretation, it is clear that it is a precondition of the explanatory framework that consciousness is irrelevant for the explanation of COR: consciousness must be assumed irrelevant because the computational framework can encompass only characteristics posited to be shared

science. In contrast, I emphasize the distorting effect of attempting always to explain the cognition of sentient animals through the medium of nonsentient computers.

10. In most philosophical accounts of consciousness, it is acceptable to say that it is the brain that is conscious (e.g., Prinz 2012), but current ignorance about the basis of consciousness, I prefer not to commit myself on the question of whether the term "consciousness" is best employed as a characteristic of the whole animal, its mind, or an organ within it (the brain).

11. A cautionary note from Cantwell Smith (2019, 50) is that the ANNs are not doing object recognition in any rich sense. Concretely, what it does is learn mappings between pixel strings of image input and letter-string output—the words that humans use to refer to objects in the images.

between brains and computers. Thus, consciousness presents an explanatory gap for the computational perspective, but it is not a gap just *there* in nature. Rather, it is a built-in lacuna of the explanatory framework.

It is worth saying more about the example of core visual object recognition, also known as "preattentive recognition" (Serre 2019, 416). This is the object recognition that humans and other primates perform quickly (within 100 milliseconds), independently of attention or top-down influence of scene interpretation (Cadieu et al. 2014; Kheradpisheh et al. 2016). It has been associated with activity in the ventral stream of the primate visual cortex, running from primary visual cortex (V1) to inferotemporal cortex (IT). The ventral stream was hypothesized by Milner and Goodale (1995) to be responsible for visual consciousness and object recognition, in contrast with the dorsal stream, serving nonconscious visual guidance for action. In my argument, I am not claiming that objects are never recognized unconsciously, nor that the ventral stream is the neural correlate of visual phenomenal consciousness. My point is that in the behavioral paradigm of COR, consciousness undoubtedly accompanies object recognition for humans, and it would not be reasonable to doubt that other primates have similar experiences. Although introspective reports of awareness may be deemed unreliable in some threshold cases (e.g., when stimuli have very short presentation times of less than 10 milliseconds), there are no grounds to be skeptical that when people report visual awareness of a stimulus presented in the viewing conditions used in these object recognition experiments, they actually have those experiences.

Very shortly after the breakthrough performances of DCNNs on visual object classification, neuroscientists began to look for similarities between the architecture (i.e., the many-layered hierarchy) and "representations" (including receptive field structure) between DCNNs and the ventral stream.[12] Such models were found to offer the best predictions to date of the responses of neurons in areas of the ventral visual system such as V4 and IT that had traditionally been hard to model (Yamins et al. 2014). Various researchers,

12. A significant result here is AlexNet, by Krizhevsky, Sutskever, and Hinton (2012). ResNet, by He et al. (2016), is the first DCNN considered to have surpassed human-level accuracy at object classification. The story of the impact of these developments on vision science is recounted by VanRullen (2017), Serre (2019), and Yuille and Liu (2021).

such as Khaligh-Razavi and Kriegeskorte (2014), have argued, on the basis of observed similarities, that the DCNN provides an explanation of the neural activity underlying primate object classification, a view also echoed in the philosophical literature (Cao and Yamins 2021a).

It must be appreciated that the DCNN model is in no way presented as being in the business of explaining visual phenomenal consciousness. Instead, the researchers argue for its performance equivalence in the behavioral task, which is then accounted for through the demonstration of similarities between the DCNN and the ventral stream regarding layered, hierarchical organization, representations, and computational operations such as convolution. The key point for us is that there is not even a whiff of a hint at how such a model could explain those conscious experiences that occur in the primate but not in the computer. And to the literalist about neurocomputational models, this should be puzzling because the computational model is supposed to tell us what COR essentially is; yet it has no scope to explain how or why consciousness accompanies object recognition in animals. Thus, the computational account faces an explanatory gap that is not surprising on the analogical interpretation, for it is clear that it is a *precondition* of the explanatory framework that visual consciousness is irrelevant for the explanation of COR. This is because the program of computational explanation depends on the prior assumption that what is relevant to the explanation of a cognitive capacity, like core object recognition, is the proposed set of features in common between brains and computers. Consciousness cannot be among the commonalities since animals have it, but no known machine does. I diagnose this as the source of an explanatory gap: the fact that the computational framework must restrict itself to explaining cognitive capacities thought to be shared between animals and computers, in terms of processes and features also thought to be common to both classes, and since consciousness is not one of them, it can be no part of such explanatory programs.

To commit the fallacy of misplaced concreteness is to fall into the temptation to reify the abstractions that make scientific modeling successful. It appears that the literal interpretation of neurocomputational models, as well as the computational theory of mind that goes hand in hand with it, are guilty of this substitution of the brain, with all its concrete details, for a mathematically precise, simplified version of some of its processes. In succumbing to the fallacy of misplaced concreteness, one forgets not only that the model is an abstraction, but also that in the decision to make certain simplifications, one

may be disallowing room in the explanatory framework for the very features for which an explanation may later be sought. To reiterate, it is an assumption of the modeling framework, grounded in selective machine-organism analogies, that the differences between the analogy source and analogy target can be ignored for a circumscribed set of predictive and explanatory purposes. There is obviously no guarantee that that assumption will hold when the framework is extended to attempt to explain additional features of the target beyond the initial scope of the analogy. But this obvious point is lost once the computational brain has replaced the concrete brain.

## 9.3 Scaling Up and Multitasking toward Machine Consciousness?

In this chapter, I am working toward an argument as to why, even in principle, inorganic computing machines could not be conscious. This argument will not be completed until section 9.4, where the in principle positions of some philosophers will be discussed. At this stage, it should be appreciated that the observations of section 9.1 do not make much progress in the stated direction. This is because the claim that machines can or cannot be conscious is about the kinds of things that have sentience, whereas the identification of the explanatory gap is an epistemic matter—it just presents a conundrum for those whose faith in the computational theory of mind leads them to think that it will be possible to explain consciousness in computational terms. But a believer in the viability of machine consciousness can easily sidestep such concerns: even if a computational explanation of consciousness cannot be achieved, so the response goes, it can still be the case that brains (or their owners)[13] are conscious because brains are very large, densely connected neural network computers. So, the idea goes, if you keep scaling up the ANNs, eventually you might build a conscious machine.

However, it turns out that there are important lessons to be gleaned from my criticisms of the literal interpretation, which apply equally to this scaling-up idea. There is an argument to be made about why we should not expect expert system ANNs, of the sort built so far, to grow into an AI that is more humanlike, to become AGI with sentience. The relevant issue is that there is only an applicable similarity between the ANN and the organic

13. This is to be neutral on the question of whether it is animals or their neural organs that are sentient, properly speaking.

system when one considers the narrowly specified task that the ANN has learned to perform. The scope of the analogy has built-in limitations that are neglected by the literal interpretation. However, the literal interpretation gets implicated in the view that the ANN has tapped into something essential to organic cognition—namely, a kind of computation first realized in the brain—and that this coinstantiation of a computation is the basis of animal-equivalent performance in a learned task. In this view, it then makes sense to think of the trained ANN, as narrow in its expertise as it is, as having nudged its foot over the threshold of having some set of characteristics that mark the difference between cognizing and mindless systems in nature, and therefore being at the early stages of the path toward AGI.[14] And with this, there is imagining that each giant leap of an expert AI is also a small step toward AGI.[15] The problem is that the literal interpretation fosters the idea that ANNs could scale up to sentience, and this expectation is due to its failure to recognize the narrow scope of the similarities between machines and organisms.

To see what is treacherous about this failure to acknowledge differences, let us imagine we are back in the nineteenth century, when steam locomotives had recently demonstrated their superior strength and speed at dragging heavy loads along flat surfaces—better than even the best-bred shire horses. Some Victorian technooptimists might start speculating that in the next few generations of steam technology, we will have locomotives capable of leaping fences and doing the other elaborate moves that horses

14. AGI is the stated long-term goal of research at AI company Google DeepMind. See https://www.deepmind.com/about.

15. This has long been a theme in press reports on expert system advances, such as the following report in *Wired* magazine on AlphaGo Zero: "The new DeepMind research has been published in the journal *Nature* and is another significant step towards the company's goal of creating general artificial intelligence" (Burgess 2017).

Representatives of DeepMind have sometimes scaled back expectations—see, for instance, the chief executive office (CEO) Demis Hassabis interviewed by Wiggers (2018)—but it is clear that the literal interpretation of neurocomputation has a prominent role in shaping researchers' anticipations of how progress will be achieved: "Distilling intelligence into an algorithmic construct and comparing it to the human brain might yield insights into some of the deepest and the most enduring mysteries of the mind, such as the nature of creativity, dreams, and perhaps one day, even consciousness" (Hassabis et al. 2017, 255).

display in dressage competitions. But naturally such predictions would be laughed at because the success of steam technology is not a marker of it having begun to capture the essence of biological locomotion. Dragging heavy loads along roads or tracks is not some core capacity of biological motion, but a very simplified kind of movement, which happens to have enormous economic significance. There is no reason to say that the locomotive is on the path to full-blown biofunctionality since it is no more than a machine built to do one task that the horse does, at performance levels beyond what would be possible for the animal. Indeed, its superbiological performance is *due* to it being so unnaturally specialized.

Just as a shire horse is like a locomotive in some respects—one can make functional comparisons with respect to the one task of dragging loads along flat surfaces, and hence engineers coined the term "horsepower," applicable to machines as well—AlphaGo is like the human Lee Sedol in one respect, in the capacity for playing Go, and ChatGPT is like a person in that it can generate sentences. The analogy between the horse and locomotive breaks down when considering motor capacities other than haulage. I am saying that ANN-brain analogies should be expected to break down as soon as we consider tasks beyond those that the machine has been designed to replicate. Hence, ChatGPT should not be thought of as containing the germ of a humanlike intelligence. The idea that a scaled -up, more powerful version of AlphaGo, or ChatGPT, will become a HAL 9000 then looks as absurd as our Victorian technooptimists expecting the emergence of hurdle-jumping steam engines.

The context of the invention of computers is actually similar to the one that saw the invention of steam engines, and machine manufacture. As historians have attested, computers, no less than locomotives, were invented to perform one particular job for which there used to be full reliance on animal labor—but the animals in question were humans and the labor was cognitive rather than muscular (Daston 1994, 2018; Schaffer 1994). It is clear from Turing's account of his conception of the computing machine that it was to perform all and only the cognitive work done by a human computer—a clerical laborer whose job it was to perform arithmetical calculations: "Electronic computers are intended to carry out any definite rule of thumb process which could have been done by a human operator working in a disciplined but unintelligent manner" (Turing c.1950, quoted in

Copeland 2020).[16] We should not ignore the fact that the behavior imitated by the invention is here classed as "unintelligent."

Moreover, a general feature of machines is that they are precisely designed for reliable, predictable, expert performance in controlled conditions, such as in a factory or office or on roads or railways, and as such require a human-created infrastructure to function. The unmade environment, being uncontrolled and less predictable, makes demands on behavioral flexibility not needed when conditions are artificially held fixed. Machines, operating within human-created cocoons or "micro-worlds," can afford to be inflexible and superspecialized (Collins 1996), which is at least part of the explanation of their superhuman power. It is common to treat the brain as a computing machine whose designer is natural selection (e.g., Dennett 1995). This gets the machine-organ relationship wrong. Machines, including computers, are tools designed to duplicate just one of the countless functions that living organs, like the brain, achieve. You should expect organs to have many capacities that machines do not have because unlike machines, they are not designed, and a fortiori, they are not designed with just one task in mind, to be performed repetitively within steady-state conditions.

Still, there is a difference that we must acknowledge between these cases of nineteenth- and twenty-first-century machine development, which is that horses were not relevant to the design of locomotives in the same way that brains have been the inspiration for specific features crucial to the design of the most advanced AI systems. As we saw in chapter 5, Fukushima's deployment of Hubel and Wiesel's hierarchical model of the visual cortex led to the "neocognitron," and from there to current DCNNs. Similarly, the architecture of *deep reinforcement learning* took its lead from behaviorist psychology and neuroscience (Hassabis et al. 2017, 246–247). Taking the case of DCNNs used to model primate vision, it might reasonably be argued that they are much more biologically inspired than either the locomotive or the Turing machine. Therefore, it could be claimed that they work as well as they do because they replicate, on an abstract level, computations and representations actually occurring in the primate

16. Cf. "The class of problems capable of solution by the machine can be defined fairly specifically. They are [a subset of] those problems which can be solved by human clerical labour, working to fixed rules, and without understanding." Turing (1945/2005), quoted in Copeland (2019).

cortex.[17] This pushes us back toward a literal interpretation of these ANNs, as instantiating neural computations, which is the germ of the expectation that larger-scale versions of such technologies will show more humanlike intelligence.

My first reply is that while the initial seed of an invention has often come from neuroscience, the pattern is that the majority of the subsequent engineering advances occur without reference to the brain (Marblestone et al. 2016, 1–2). Later work on the correlation between bioplausibility and performance in DCNNs weighed against the idea that the machines work as well as they do just because of their brainlike features (Schrimpf et al. 2020). While in the earlier days of development, better performance on a visual object classification was achieved by adding brain-inspired features, this pattern is now broken, and performance gains are no longer achieved by making networks more biorealistic.

A recognized source of disanalogies between the characteristics of neural systems, like the primate ventral stream, and those of machines that are functionally equivalent in some respect is due to neural systems being involved with multiple operations at any one time, not neatly demarcated, whereas ANNs work exclusively on specific tasks. DCNNs trained for core object recognition currently provide the best predictions of responses in ventral-stream neurons to visual stimuli but they account only for a proportion of the variance of IT neuronal responses. The thing to bear in mind here is that IT not only does core object recognition; it is also implicated in other behaviors, such as people's remarkable capacity to recall having

17. This argument is as old as Aristotle. Newman (2004, 18–19) writes:

> At one point (IV 3 381b3–9) [Meteorology,] Aristotle justifies his use of terms taken from cooking, an artificial activity, to describe processes in nature. He claims that "art imitates nature," using this fact to justify the imposition of technical terms such as "boiling" and "roasting" onto natural phenomena. Since artisans have learned their operations by imitating nature, it is unproblematic to use their technical language in describing the natural processes that they have copied. If one takes this to mean that these human artisanal processes are identical to their analogues in the natural world, it opens an avenue by which the imitation of nature—from which the processes are learned—could lead to the very perfecting about which Aristotle speaks at *Physics* II 8 199a15–17. Since this type of imitation would utilize natural processes, one could legitimately argue that it leads to a natural product and that it is in fact perfective.

That said, the erasure of the divide between nature and artifact is not a consistent position of Aristotle's. As Newman (2004, 12–13) also reports, there is the example in the *De Anima* of Daedalus's automaton that *seems* to be alive but actually is not.

viewed thousands of different images (Standing 1973; Brady et al. 2008; Meyer and Rust 2018).

The existence of these kinds of dissimilarity—that brain areas but not ANNs are multitasking—is not contestable. The further point is that the absence of multitasking may help to explain why some ANN responses are strikingly different from human ones. A nice case in point is the tendency seen in DCNNs trained over standard image databases to make object classifications according to the texture of an item rather than its shape (Baker et al. 2018; Geirhos et al. 2019; also see figure 9.1). In contrast, the human tendency is overwhelmingly toward classifications based on shape. Consideration of human psychology and behavior—the fact that we are actors and conceptualizers as well as perceivers—makes sense of our shape bias. If I shave a cat (thus transforming its texture), it is *still* a cat for all intents and purposes. My concept of what a cat *is* needs to be based on more inherent features than its surface texture. Texture, though a reliable statistical marker of object identity, does not have the conceptual import that the form of an object does, and it does not have the same significance for deciding how to interact with most things: a pet cat might be chosen for its lovely, soft fur, but furriness alone does not make for a good pet (or else every cat lover could make do with a furry scarf).[18]

The upshot of this discussion of multitasking in the brain is just that scaling up by itself, building bigger and bigger networks of the expert system type, should not be expected to bring about humanlike AGI, with consciousness, if major disimilarities between animal and machine intelligence arise from multitasking in neural systems. It is not coincidental, therefore, that some have proposed that building multitasking machines, rather than expert systems, is one route to the engineering of AGI. To achieve more humanlike object recognition, one could conceivably build a fancier machine that does

18. Geirhos et al. (2019) show that a network can be made to classify by shape rather than texture, training it on an image set in which texture is not diagnostic for object identity. But this is something of a kludge since in the "training set" of human vision (in images of ordinary objects), texture *is* diagnostic for identity (if it weren't, a DCNN could not learn to employ texture for recognition), and yet we classify by shape anyway. The argument of Hermann, Chen, and Kornblith (2020) regarding "internal workings," that the texture bias of DCNNs is due to a difference in the data they are exposed to, not a dissimilarity from the brain, is not convincing, as the study is not able to reproduce the overwhelming shape bias seen in human perception.

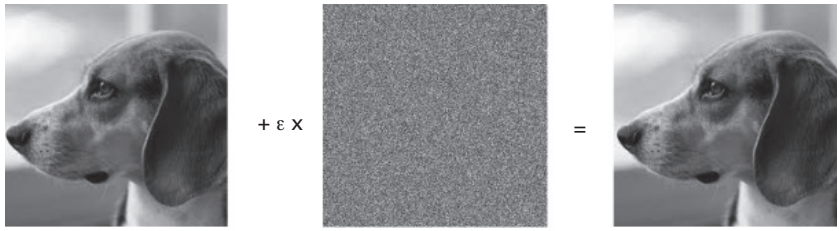


**Figure 9.1**
Illustration of object classification by a DCNN altered by superposition of a very-low-contrast texture pattern (center) to the image of a beagle (left), resulting in the high-confidence classification of the photo as an ostrich, even though it looks identical to a human observer. (Source: Joshua Clymer, CC-BY-SA 4.0: https://commons.wikimedia.org/w/index.php?curid=126027330.)

some of the multitasking of the human ventral stream, which is involved in memorization, concept formation, and functionally integrated with brain areas involved in decision making and action planning. For example, Park et al. (2020) present VisualCOMET, a GPT-2 transformer architecture trained both on textual and pictorial representations of everyday scenes. It was found to be better at common sense reasoning than a purely text based system, even when given text prompts. The aim of this research is to produce an ANN with "visual understanding."[19]

However, if the idea is to engineer multifunctionality to arrive at an AI with consciousness, we hit a stumbling block with the question of what task would need to be added to grant the network visual phenomenal consciousness when it does object recognition. Consciousness is not a capacity, like memorization, that is easy to benchmark and that could be tacked on to a machine with capabilities for object recognition. Recall from section 9.1 that the precondition of the computational framework in AI and theoretical neuroscience is that consciousness is functionally epiphenomenal—that it makes no difference to cognitive performance. It is a presupposition of the practice of computational explanation of cognitive capacities that

19. The Bolt-On approach to engineering AGI does, of course, have its skeptics, such as Loukides and Lorica (quoted in Mitchell 2019, 41): "A pile of narrow intelligences will never add up to a general intelligence. General intelligence isn't about the number of abilities, but about the integration between those abilities."

consciousness is irrelevant to those performances since the explanations may refer only to computations realized in nonsentient machines. The fact that consciousness is not functionally definable, in the sense that we cannot say what extra thing a system with consciousness can *do* in comparison with a nonconscious peer, is a consequence of the computational framework. But this prejudice against the functionality of consciousness presents the advocate of machine sentience with an obstacle. Because the framework of computational explanation stipulates that consciousness is irrelevant to cognition (being a characteristic not common between computers and animals), the framework can say nothing about the added value that consciousness brings to the cognitive economy of an animal. Therefore, someone wishing to engineer consciousness through multitasking is given no clue as to what kind of cognitive and behavioral abilities would need to be added to have consciousness play a role in the cognitive economy of the machine. Technooptimists are left with no more than the idea that if a machine is built with as many nonconscious cognitive capacities as possible—in the limit, as many as the animal has—then consciousness will somehow follow along. This is just another version of the scaling-up idea—by building bigger and fancier AI, it will eventually spawn some consciousness.[20]

20. It might be argued that even if there are obstacles to engineering-conscious machines, it is still possible to build machines that acquire generally intelligent capacities that transcend the limitations of expert systems. In particular, the idea is that through reinforcement learning in the right training environment, the agent will be pushed to acquire the ability to achieve a wide range of goals, the ability that Turner et al. (2021) call "power." (I thank Jacob Pfau for this point.) There is, of course, interesting work in this area, and it is to be expected that ANNs trained through different methods will be more supple than networks, like the DCNNs for visual recognition discussed previously, which undergo supervised learning with very specific task functions. That said, we still need to appreciate that what a digital computer essentially is a model of a very particular cognitive performance originally performed by human computers (see Turing, quoted in section 9.3), and this model disregards the aspects of human cognition that do not involve execution of coded instructions—just as the steam engine is a model of one isolated kind of horse locomotion. I have argued elsewhere (Chirimuuta 2023b) that this basic fact puts limitations on the kind of general abilities that digital computers could acquire. I am skeptical about claims that sophisticated twenty-first-century computers have emergent capabilities that amount to something qualitatively different from their being an immense aggregate of the basic operations of a digital machine. At the same time, I do not think that human cognition can be reduced to those basic digital operations.

## 9.4 Against Conscious Isomorphs

"Technooptimism" is actually the name given by Susan Schneider (2019, 18) for the view that conscious computational systems are on the horizon. This view is set against biological naturalism, which asserts that "even the most sophisticated forms of AI will be devoid of inner experience" (Schneider 2019, 18).[21] In this section, I will defend biological naturalism by showing that the objections to this view, from both Schneider and David Chalmers, are guilty of a fallacy of misplaced concreteness, failing to appreciate that substrate-independent computational models of the brain could only be impoverished abstractions. Both these philosophers' arguments are in principle ones, but I will show how they are founded on the problematic, literal interpretation of neurocomputational models.

The position set up by Schneider (2019, 24) in opposition to biological naturalism is called *computationalism about consciousness* (CAC). It is the idea that "consciousness can be explained computationally, and further, the computational details of a system fix the kind of conscious experiences that it has and whether it has any." To flesh this out, Schneider introduces the notion of a *precise isomorph*, a system that mimics the computational organization—the "precise functional organization" (2019, 26)—of the brain of a sentient animal. CAC says that the isomorph will be conscious.[22] The crucial notion is that of the "precise functional organisation" of a brain. Schneider (2019, 27) says that it is an "abstract pattern of causal interactions between the different components of your brain," which could be represented in a graph. With

21. Schneider herself does not endorse either of these views, opting for a wait-and-see approach, whereby "conscious machines, if they exist at all, may occur in certain architectures and not others, and they may require a deliberate engineering effort, called consciousness engineering'" (Schneider 2019, 34). Since this approach holds it as an open possibility, compatible with neuroscience as we know it, that inorganic computers may become conscious, it is just as much the target of my criticisms as the technooptimist one. Searle (1992) is cited for "biological naturalism." See note 3 in this chapter on how the definition of the term has changed since Searle's introduction of it. My defense of "biological naturalism," via rejection of hypothetical electronic "isomorphs," should not be taken to imply an endorsement of Searle's positive case for the view.

22. "What CAC amounts to is an in-principle endorsement of machine consciousness: *if* we could create a precise isomorph, then it would be conscious" (Schneider 2019, 25; emphasis in original).

all the interactions of each neuron in the brain mapped, Schneider asserts that in principle, each neuron could be replaced by a silicon-based substitute neuron. The resulting being would be a precise isomorph, and with the functional organization of the brain perfectly duplicated, the new creature would have all the mental characteristics of the original, including consciousness. The crucial move in the argument for CAC is that during the process of swapping organic neurons for artificial ones, it is implausible that consciousness will fade out, or abruptly switch off, since artificial neurons are stipulated to preserve the functions of the old ones. Hence the most likely scenario, so the argument goes, is that consciousness holds steady, supporting CAC, the claim that the computational organization of the system determines its consciousness, independently of material constitution.

Schneider's argument draws from one presented in Chalmers (1996) and reiterated in Chalmers (2014). Chalmers (2014, 105) admits that the claim that "functional isomorphs" are possible is a substantive one.[23] And it is this claim that I will challenge. The isomorph argument that animal sentience can be replicated in an inorganic machine depends on the assumption that components of the brain (i.e., neurons) can have their input-output functions duplicated by objects made from an utterly different substrate—electronics, as opposed to living tissue. The plausibility of this assumption stems from people's familiarity with quite simple neurocomputational models, such as receptive field models, that are sometimes presented as encapsulating all the input/output behavior of a neuron. The literal interpretation of such models fosters the idea that there is nothing more to the neuron, functionally speaking, than that—nothing left out that is relevant to the cognition, all that's missing being the inessential matter of implementation, the gooey details

23. Chalmers (2014, 105) does mention that functional isomorphs are not possible if some neurons "function in a noncomputable way, for example, so that a neuron's input/output behavior cannot even be computationally simulated." My argument here does not rest on any technical claims of noncomputability, but rather on the denial that neural structure and function can be separated in the way that is required for these simulations in inorganic machines. I argue elsewhere for the interdependence of structure and function in the brain, based on neuroscientists' newfound appreciation for chemical signaling (Chirimuuta 2022b; see also Maley 2021). My rejection of these hypothetical simulations has a precursor in Haugeland's (1981/1998a) notion of "second order messy" analog systems, like the metabolic network of a rat. Such systems do not afford digital simulation because it is not safe to leave any of their structural details out of the simulation, right down to the finest level of resolution, and hope to preserve function.

of the stuff that the neuron is made of. But, as we have seen, that is the modeler's convenient fiction, pretty harmless within the context of scientific research but requiring more scrutiny when exported to philosophical debate. Without the assumption of a division—marked unambiguously within the brain itself—between cognitive activity and mere metabolic support, there are no grounds to think that a model of a neuron that leaves the details of its actual operations could have a chance of replicating its role within a cognizing system.

We must grant Schneider and Chalmers that their scenarios are hypothetical, and it is open to them to stipulate that the artificial neurons of the isomorph are ideally precise models—they re-create *all* of the input-output behavior of an actual neuron, down to its barely detectible fluctuations in membrane voltage, due to minute changes in permeability to ions. And now the assumption that I would like to challenge is that the perfect replication of that neuron's input-output behavior, down to the level of precision required for maintenance of organic cognition, could be anything other than a neuron. Just as in the tale by Borges (1998), "On Exactitude in Science," where the only perfect enough map of the empire had to be on the scale of the empire itself, the perfect input-output equivalent of the neuron would have to be something more like a duplicate than a model of the cell. Although neuroscientists do not tend to think that all the most minute of alterations in microchemistry are determinative of whole-neuron response profiles, given findings of maintenance of neuronal performance across changes in cell membrane constitution (O'Leary et al. 2013), for functional equivalence, we would still need to go well outside the realm of Schneider's and Chalmers's electronics substitution scenarios.

The critical point here is that given what is known about the workings of neurons, there is no reason to think that a functional equivalent could be achieved with a material substrate so unlike biological tissue as an electronic computer (cf. Ginsburg and Jablonka 2019, 467). Thinking that this is possible is just one of the seductions of the literal interpretation, its intimation that the material details of neurons and the embodied context of the nervous system are merely the background to cognition and separable from it. Cognition itself, on this view, is *substrate-independent*. The technooptimistic fantasy of uploading of consciousness gets its appeal from the idea that your conscious mind could be recreated in a nonliving machine with no material similarity to your brain and body, and given that the machine was never alive, it—unlike your body—never has to die. It is, of course, a fantasy

of immortality; but if the functionality of neurons is as utterly dependent on their being made from living materials as I contend, this is a complete impossibility.

A philosopher's objection that will be raised here is that I have not specified my sense of possibility or said what I meant by "in principle." Schneider (2019, 25) is clear that her argument is about the logical or conceptual possibility of machine consciousness, whereas my critique boils down to facts about how actual neurons work, and as such can do no more than state the *nomological* impossibility of consciousness in inorganic machines[24]—its incompatibility with the workings of nature as we happen to find them. However, an argument for nomological impossibility is good enough for me. The case I want to make is that in *this* world (not all the possible worlds where the laws of nature are postulated differently), inorganic machines will not become conscious. Biological naturalism is, as I would have it, a claim about the world as we know it—that consciousness is, and only ever will be, a characteristic of some living organisms. Schneider's and Chalmers's arguments might still stand as reasons to hold on to the conceptual possibility of machine consciousness, but that is not much use to their readers hankering after silicon-based immortality. Not even the most ambitious of the technophiles would invest in a space program for transport to adjacent possible worlds.

## 9.5 Missing Understanding

> Those alone think who do not passively accept the already given.
> —Theodor Adorno (2005, 264)

*What is thought, such that a machine must lack it?* One answer to this question is to be found in "The Latest Attack on Metaphysics," an essay by Max Horkheimer published in 1937 that has surprising resonance in our age of automated science. That piece was a polemic against logical empiricism,[25] a theory of the nature of scientific knowledge and understanding which, as

24. Note that my argument does not rule out the possibility of consciousness in hypothetical machines made through synthetic biology—machines that are alive and made from organic (metabolizing) substrates.

25. Discussed by Dahms (1994) and O'Neill and Uebel (2004), although their accounts of the dispute are quite partisan toward the logical empiricist side and indicate a lack of comprehension of the issues at stake for Horkheimer.

we saw in section 8.3.3, is exemplified in the programs for automated science arising nowadays with the accumulation of supersized data sets and machine learning methods for digesting them. I noted before that both logical empiricists and promotors of automated science were happy to redefine *scientific understanding* as the ability to make accurate predictions on the basis of acquired data. I postponed criticisms of this move until this chapter, where the background to the discussion is a more general examination of claims for mechanized consciousness and general intelligence. Consideration of the difference between science as performed by humans and machines will bring the notion of understanding into relief and make obvious why it should not be devalued or downgraded.

So far in this book, I have gone along with the term "AI," which encapsulates the thought that certain devices already in existence, while artificial, are also intelligent.[26] But there is actually still an open question about whether such creations are intelligent, properly speaking—whether they are thinking things. It is helpful first to review the analysis of automated science put forward in section 8.3.3 of chapter 8. The difference between a machine, like a deep ANN that can model complex, high-dimensional data sets, and a human scientist is that the machine has no ontological posits in addition to the data it is fed in order for it to find statistical regularities for the purposes of prediction, generation of samples, and classification. I brought Ernst Mach into the discussion there because on his ideal of science, an ontology of actual things with essential properties was vestigial and ill placed. The task of science, according to Mach—at least according to a caricature that focuses on his depiction of science as the project of minimizing mental effort (Patton 2021)—is to order and represent the data economically, as a means to accurate prediction of new data, with the ultimate aim of instrumental control. There is no question of any things beyond or behind the data stream. This

26. To preempt my conclusions, I would be happy with the term in the colloquial but nonetymological sense of the word "artificial," as "fake" or "imitation" (e.g., "artificial cream"). I assume that this meaning is not operative in the world of AI, where "artificial" has its etymological meaning of "made through technique" (i.e., human-created). Herbert Simon (1969, 4) discusses this same point about the difference between the etymological and idiomatic senses of "artificial," shortly after the baptism of AI. "Artificial" has too many negative connotations for Simon's liking because, he says, "our language seems to reflect man's deep distrust of his own products." How times have changed!

would make the ANN, a statistics powerhouse built to order and generate data, the perfect scientific agent.

What's jarring about this data-adhering mindset is revealed when one considers the vulnerability of visual ANNs to adversarial attacks. These occur when, for example, a DCNN trained to classify everyday objects is presented with photographs with small perturbations (or in some cases, photographs of objects with a few specially designed stickers placed on them), which would not lead human perceivers to alter their categorization, but rather result in dramatic changes in classification for the DCNN. Initial hypotheses about adversarial vulnerability assumed that the networks were succumbing to some noisiness in their systems (i.e., that the adversarial misclassifications were not connected with the learned data structures that enable successful classification). However, later work showed, surprisingly, that the features in the data that cause the networks to make adversarial misclassifications, are also ones relied on in successful cases. As Ilyas et al. (2019) summarize their finding, "Adversarial vulnerability is a direct result of our models' sensitivity to well generalizing features in the data."

To appreciate this point intuitively, examine figure 9.1 from earlier in this chapter. It shows how an adversarial image can be generated by taking an ordinary photograph of an object and superimposing a very-low-contrast texture (the "adversarial noise" pattern) that is diagnostic of another kind of object. It was mentioned in section 9.3 that when the typical texture and shape cues of different objects are combined in one figure, human perceivers overwhelmingly make the object identification on the basis of shape, whereas a DCNN's identification will be determined by texture. Texture is a "well-generalizing feature" in data sets comprising images of everyday objects: when the DCNN under supervised learning comes to associate each name label with the texture of those objects when presented in the training data, it can reliably use those learned texture-name associations to classify objects in images not previously presented in the training set. So the texture bias of DCNNs accounts for some cases of adversarial vulnerability, and it illustrates Ilyas et al.'s general point that adversarial vulnerability is due to the DCNN learning features of image data that are actually diagnostic for classification but would not be relied upon by a human perceiver.

In their commentary on Ilyas et al.'s findings, Gilmer and Hendrycks (2019) write that the problem of adversarial vulnerability is due to the tendency that an ANN "latches onto superficial statistics in the data." This is a

very telling remark. From the human perspective, we encounter the data as being relatable to *things*, objects that have core properties and surface properties, essential and inessential features. A furry texture should not be diagnostic for the classification of a cat because a cat can lose its fur and still be a cat; the presence of glasses should not determine the identification of a person because people take glasses on and off, and this never changes who they *are*. Our thinking is constrained in this way regardless of the statistical regularities that we have experienced. In my lifetime of accumulated cat data, I have never looked at a picture of a shaved cat, and yet if I did, I would still visually classify it with the furry images whose statistics are so different. But one cannot make this distinction between "superficial statistics" and deep ones unless one has some informal ontology—a theory of things, processes, and their essential versus accidental properties—in the background. And so, from the ultraempiricist perspective of the ANN (lacking our informal, essentialist ontology), this distinction cannot be made, for data are all that there is, and statistics over them may be more or less stable and projectible (and hence more or less useful for prediction and classification), *but they cannot be more or less superficial*. The fact that we humans cannot help but think that the data features picked out by DCNNs that are nonrobust to adversarial attacks (even if predictively useful in the majority of cases) reveal ignorance about what something (e.g., a cat or a butterfly) actually *is*, is some indication that the ontology beyond the data, however informal—some notion of beings, with essential and inessential properties—is not dispensable for human thought.

It remains to be seen whether adversarial vulnerability presents a serious barrier to the rollout of automated science restricted to the instrumental aims of prediction and control (Buckner 2020). What it demonstrates is that we should not hope for automated science, restricted as it is to making inductions on occurrent data, to deliver anything beyond these aims. Moreover, these observations are a springboard for a new characterization of understanding, and they help us to account for the significance of the machine's lack of understanding. I contend that the ANN's restriction to its data stream, and the human's irrepressible tendency to think beyond the data, make the difference between the absence and presence of understanding. Understanding is the activity of sense making performed by human beings who ceaselessly act in among things.

As suggested by the previous discussion of the shape bias, our thinking about objects with essential and nonessential properties is a manifestation

of the fact that human thought is about our dealings in the world around us, and no less than our bodily movements, it is an activity. As asserted previously, human visual object recognition is not a module detachable from action and conceptualization. We cannot carve off object recognition from an understanding of things—a sense of what the objects are, and what their presence means—because perceiving, thinking, and moving are interconnected for us, and thinking is not the passive ingestion of data, but the active, communal production of a network of meanings.[27]

This opinion on what understanding is, and why twentieth-century empiricism has no account of it, is expressed by Horkheimer (1937/2002, 145):

> In the eyes of the empiricist, science is no more than a system for the arrangement and rearrangement of facts, and it matters not what facts are selected from the infinite number that present themselves. He proceeds as if the selection, description, acceptance, and synthesis of facts in this society have neither emphasis nor direction. Science is thus treated like a set of containers which are continually filled higher and kept in good condition by constant repair. This process, which was previously identified with the activity of the understanding, is unconnected with any activity which could react on it and thereby invest it with direction and meaning.

The point is that the processing of data ("facts") by itself, and without some additional activity of thought, has no path toward meaning, and hence understanding. It is a view of broadly Kantian origin, also expressed

27. Comparable here is Vallor (2021) on the "sense-making labor of understanding". Another term that could be used here is to say that human thought is "world forming." Heidegger (1995, part 2) contrasts this with the condition of animals that are in his view "world poor" because their behavior is relatively more conditioned by environmental triggers. My position is that AIs, lacking understanding and sentience, simply do not have a world—are neither "world forming" nor "world poor," which is the condition of inanimate objects. I develop the Heideggerian point elsewhere (Chirimuuta in preparation). We may appreciate here a connection between understanding and sentience, lost in too many contemporary discussions in which consciousness is carved off from cognition and treated as nonfunctional. To be a sentient creature is to have a world around you which you are aware of and which is inherently meaningful, and this does not happen without some activity of understanding. Recent accounts in philosophy of cognition, sensitive to this point, are in Thompson (2007, 228), who follows Maine de Biran in considering consciousness a, "sentiment de l'existence"; and in Ginsburg and Jablonka (2019, 7), who treat consciousness not as a capacity or property of a system (like having sight), but as a mode of being.

in Cassirer's friendlier, though still critical, discussion of Mach (Cassirer 1910/1923, 261).[28]

A central feature of this broadly Kantian view is that it rejects the empiricist notion that knowledge rests on some foundation of pure data, facts that are immediately given and unadulterated by any activity of the mind.[29] In Horkheimer's version, thought, even when resting on facts, must always be in the business of evaluating them, which means that empirical thought, as much as it is immersed in the factual, should never aspire to a value-free ideal:

> In countering experience, the intellect must itself appeal to experience, for its concepts are not inborn or inspired. The answer is that it is precisely because facts are referred to when other facts are being exposed or abolished, and because facts, as it were, are involved in everything on every hand, that constructive thought which evaluates facts and discriminates between surface and pith is of such supreme importance in every decision. (Horkheimer 1937/2002, 151–152)

The notion explicated in this section, that of understanding being a constructive and normative activity, is consistent with the *verum factum* account of scientific understanding depicted in chapter 5 and explicated in section 8.2 of chapter 8. A genuine cognizer has a role in shaping the data (which are not just given on a plate), and with this process of making and interpreting, can achieve an understanding. This shaping cannot be extricated from the goals and values of the agent. With this in mind, the fundamental disanalogy between humans and machines, responsible for the presence and absence of understanding, respectively, comes down to the artificial creatures having no

28. A nice point of connection here can be found in Rosenblatt's (1958, 404–405) report of the deficit in the "symbolic behaviour" of the perceptron. His reference is to the analysis of Kurt Goldstein (1940) on the inability of some patients with brain lesions to perform certain kinds of abstractions. As argued in Chirimuuta (2020a), Goldstein's account of abstraction and symbolic behavior is the result of a strong mutual influence between the neurologist and his cousin, Ernst Cassirer. Thus, I would conjecture, Rosenblatt is making an early report of precisely the lack of spontaneous meaning-making that I am diagnosing as the barrier to artificial systems being genuinely intelligent.

29. As Horkheimer (1937/2002, 158) recounts, "The given is not only expressed by speech but fashioned by it; it is mediated in many ways. In accordance with its philosophical presuppositions, Neo-Kantianism has understood the activity which produces and organizes the facts to be an intellectual process." Famously, this is the rejection of what Sellars (1956) called the "Myth of the Given."

self-propelled activity.[30] ANNs, including the most advanced large language models (LLMs), are devices that find statistical regularities in enormous, high-dimensional data sets that are inscrutable to humans. Such machines model data, but they do not shape their own data, nor do they posit anything behind the data, because they lack spontaneity of thought,[31] and this is, more generally, because they lack self-propelled activity—an ANN is not a doer, and it is not the kind of thing that could be because it is not alive (cf. Thompson 2007). Only living systems are self-making and self-propelling, and as such, can spontaneously weave a nexus of meaning around themselves. However, recent models of organic cognition, somewhat transfixed by the analogy with ANNs, present a picture of the brain as no more than an engine, like an ANN, for prediction ("interpolation") over massive data sets (Hasson, Nastase, and Goldstein 2020).[32]

From the various opinions of Horkheimer and others, I have assembled an account of what understanding is and why machines lack it. That was the main purpose of this section. Before closing, however, I would like to dwell on an additional point of view to be taken from Horkheimer's essay, on why the prospect of fully automated science should be looked at with concern. A theme of his criticism of logical empiricism is that in spite of the progressive rhetoric associated with the movement, its recommendations for scientific practice would bring about purely conservative results, only strengthening, and never destabilizing an exploitative status quo. This line of argument is summarized in a tale of a repressive and brutal state in which scientists, including social scientists, perfectly conform to the logical empiricist methodology (Horkheimer 1937/2002, 159–160). They are fully content to make predictions based on the surface data that they collect about the people of

30. The ground-level point about "self-propulsion" is that no artifacts come into existence other than through human production. We should not be tempted to equate organic reproduction with manufacture of machines. Living beings build themselves up in the processes of development in a way that artifacts do not.

31. The association between thought and spontaneity comes from Kant (Pippin 1987), but I am emphasizing, in addition, an association between spontaneity and aliveness.

32. There is currently a lively discussion of LLMs, such as ChatGPT, and whether they can be said to understand the text they produce. See Coelho Mollo and Millière (2023) for a helpful overview and positive proposal. Elsewhere, I present the argument that understanding (in the sense proposed in this chapter) is lacking in LLMs (Chirimuuta, in preparation).

this nation, and to ignore the inner truths of discontent and injustice. This is the scenario that "end or theory" enthusiasts, and even the more impartial observers of big-data science, must at least consider: in the offloading of scientific thought to machines, our knowledge (so called) will inherit the passivity of machines and can do no more than accept, at face value, what is given; moreover, the activity and spontaneity of thought, so disanalogous with automated procedures but so essential for the betterment of social, economic, and ecological conditions, risk falling out of relevance.

## 9.6 Conclusions

Like Mach, Whitehead prized the facility of mathematics to disburden the mind. He described the expanse of thoughtlessness as a marker of civilizational progress, and seemed fairly pleased about it, writing: "It is a profoundly erroneous truism . . . that we should cultivate the habit of thinking of what we are doing. The precise opposite is the case. Civilization advances by extending the number of important operations which we can perform without thinking about them" (1911/1948, 41–42, quoted in Vallor 2021).[33]

Whitehead also said that the fallacy of misplaced concreteness is a source of philosophical ruination:

> The great characteristic of the mathematical mind is its capacity for dealing with abstractions; and for eliciting from them clear-cut demonstrative trains of reasoning, entirely satisfactory so long as it is those abstractions which you want to think about. The enormous success of the scientific abstractions, yielding on the one hand matter with its simple location in space and time, on the other hand mind, perceiving, suffering, reasoning, but not interfering, has foisted onto philosophy the task of accepting them as the most concrete rendering of fact. Thereby, modern philosophy has been ruined. (1925/1967, 55)

I am left wondering if Whitehead ever connected these two points; for the fallacy of misplaced concreteness occurs not only because of some osmosis of abstractions from the mathematical sciences into philosophy, but also because, by education or acculturation, people become content in their thought to deal only with the operable versions of complex things, as

33. Compare this with Mach (1883/1919, 488), quoted in section 8.3.3 of chapter 8. Lindsay (2021, 8–9) gives a paraphrased version of this comment, "The ultimate goal of mathematics is to eliminate all need for intelligent thought," near the start of her book on computational models in neuroscience.

depicted in formal models; and because they are not in the habit of demanding that we think more deeply into things than is possible through quantitative, and ultimately automated, techniques. The depth of thought relevant here is orthogonal to deductive and inductive rigor; it comes with a willingness to be bogged down with details, to be concerned about the inner truths of situations, as seen in Horkheimer's tale, even though that process is inefficient and unrewarding in its production of exact and utilizable results; it demands an ethical stake.

In this chapter, I have argued that if we see through the fallacy of misplaced concreteness and do not settle for the quantified, operable depictions of mind and brain that computational neuroscience provides, we can appreciate how and why organic cognition can be so different from its machinic imitation. Computation, I have urged, is not the essence of biological intelligence. We are then left with the question of what to say positively about the mind. The position of biological naturalism regarding consciousness and understanding obviously fits well with the doctrine of embodied cognition, as put, for example, by Evan Thompson (2009, 81), that "mind is life-like, and life is mind-like." In fact, this chapter is a modern-dress reenactment of an earlier twentieth-century argument for embodied and embedded cognition on the grounds of a critique of abstraction (Chirimuuta 2020d).

It would seem, therefore, that my final task, in chapter 10, should be to outline an agenda for embodied, embedded, noncomputationalist neuroscience. But this is not what I will do. While I believe that the theory of embodied cognition is closer to the truth of how the mind and brain work—that there are countless processes in the whole body that are deeply interconnected with the operation of the brain, and hence cognition, and likewise that the body is very much entangled with the environmental circumstances that our activities respond to—all these interconnections are nevertheless more than can be put into a tractable model or workable theory. Embodied, ecological approaches fall down, as programs in science rather than philosophies, because they make the inverse of the fallacy of misplaced concreteness: they fail to appreciate how much the task of scientific representation is to simplify, rather than to incorporate the truths of things, as far as possible, as they stand.

The computational approach modularizes the whole nervous system and posits that subsystems within the brain can be understood in isolation from the body, and from one another, using highly abstract mathematical models.

This provides clarity and precision, which leads to the prestige and success of computationalism over embodied paradigms. But it should not be forgotten that the abstraction is a departure from the truth—cognition must be very different from how it is presented by the simplified models. In the next chapter, I will examine the philosophical problem of the mind-body relationship, with the idea in place that the notional separation between brain/mind and body is itself a kind of idealization.

# 10 Cartesian Idealization

> Our calculations would be easy if there were only two bodies colliding, and these were perfectly hard, and so isolated from all other bodies that no surrounding bodies impeded or augmented their motions. In this case they would obey the rules that follow.
>
> —Descartes (1985, 244)

> These men will be composed, as we are, of a soul and a body. First I must describe the body on its own; then the soul, again on its own; and finally I must show how these two natures would have to be joined and united in order to constitute men who resemble us.
>
> —Descartes, quoted in Simmons (2011, 54)

## 10.1 Immortal, Invincible

Dualism is the repressed that always returns. As much as philosophers disavow it, the thought recurs that mind and matter are just fundamentally different—two opposite, mutually repelling poles out of all that exists.[1] The irrepressibility of the thought is suggestive of its having a wider role in a network of ideas, not yet accounted for. Even without the Christian doctrine of the immortal soul, and without a scientifically respectable notion of nonmaterial substance, dualism has persisted in some way or other—as we encountered it, for example, in the incorruptible, uploadable form of computational structure in chapter 9. No less than Descartes did himself, the present-day adherent to computational cognition believes that mind can in principle detach itself from a living body and have an existence free of it.

1. Pecere (2020) is a useful comparison of contemporary and historical opinions.

John Haugeland's essay "Mind Embodied and Embedded" finds the source of the trouble in the conception of the mind as an isolatable subsystem, interacting with the body and environment only through limited, prespecified channels; and his suggested remedy is recognition of the "*intimacy* of the mind's embodiment and embeddedness in the world," where the supposed opposite poles *comingle* and are *integral* to one another (Haugeland 1998b, 208; emphasis in original). With this I concur. However, in this chapter, I will show that more effort is needed to understand why the assumption of mutual near-isolation is so tenacious. My argument is that this form of idealization is indispensable as a way of simplifying the whole environment-body-mind conglomerate into the self-contained modules that are manageable targets of scientific research. So even if mind and body are integral to one another, a scientific perspective will struggle to see it that way. Consistent with the analysis of chapter 9, philosophy need not and should not burden itself with the simplifying assumptions of science; but in order for those abstractions to be dispensed with, they need to be recognized.

The bulk of what follows will be an examination of the workings of this form of simplification, taking Herbert Simon as our representative twentieth-century Cartesian, showing how the treatment of the human being as a hierarchical, decomposable system permits both dualism (section 10.4) and radical skepticism (section 10.5). Along the way, I offer suggestions as to the conception of mind that ought to follow from the rejection of this tradition. The conclusion of the chapter, and of this book, will be to vouch for a philosophy of mind that pursues its course independently of neuroscience, mixed with pessimism about the viability of a scientific program that is truly sensitive to the integralness of the mind's place in nature.

## 10.2 Fleshing Out Biological Naturalism

In section 9.4 of chapter 9, I defended the position known as "biological naturalism," which takes the materiality of the brain to be indispensable for consciousness, against the arguments of Chalmers (2014) and Schneider (2019) for the in principle possibility of rehousing human consciousness in an inorganic, silicon-based machine. My case rested on the idea introduced back in chapter 4, that computation is a useful model for animal cognition precisely because it abstracts away from the biological complexities that impede scientific understanding but that most likely make possible consciousness and

the other characteristics of general intelligence. As I use the term, biological naturalism is a specific claim about the reliance of consciousness on the actual material and activities of the nervous system.[2] It is not, by itself, a theory of the mind-body relationship. However, as we will now see, there are implications to be drawn out as to how to conceive of that relationship, given the commitment to biological naturalism and the account of simplification in neuroscience developed in the course of this book.

*Functionalism* is the metaphysics of mind in the background of Chalmers's and Schneider's arguments for "conscious isomorphs." While it is normally taken to be a physicalist or materialist theory—asserting that ordinary physical matter is all that there is—it retains the dualistic intuition that whatever mind may be, it can be conceptualized and examined independently of a study of the intrinsic nature of brain and body; that mind is, in principle, separable from the material body that possesses it. The positive doctrine is, roughly, that an entity has a mind, or specific mental states, by virtue of the functional relationships, a pattern of organization, among some of its internal components, its sensory receptors and motor effectors.[3] Just as a computer is what it is not because of the material that comprises it (which could be plastic lego, silicon composites, or iron cogs and gears), but by virtue of its parts, whatever they are made of, standing in the right kinds of relationships to one another, such that the transitions of states of the physical system can instantiate the steps of a computation, a cognizant being is said to have its

2. This is how Schneider (2019) employs the term, but it is originally comes from Searle (1992, 2), who defines it as the wider view that "mental phenomena are caused by neurophysiological processes in the brain and are themselves features of the brain." I emphasize that I am not endorsing Searle's wider commitments, and my sense of biological naturalism remains neutral on the question of whether the relationship between mental phenomena and neural processes is causal or something else, and whether mental phenomena are features of the brain or some extended system, possibly reaching beyond the central nervous system. Also, I am not in favor of this use of "biological" to mean "pertaining to a living system" since it erases the difference between a scientific discipline (biology) and its subject matter (living organisms); but since this meaning is so widespread, it is practically unavoidable.

3. I am just giving a general characterization here because there are many subspecies of functionalism. The *machine functionalism* of Putnam (1975) was most reliant on the strict comparison to computers, but it was later abandoned by its founder. In my discussion, however, the focus will be on machine functionalism since this is the kind of functionalism that underwrites the ideas, criticized in this book, that mental states are computational states and that the brain is a biological computer.

mind because of the structure of the arrangement of its material, not because of any particularities of the material itself. This structure is an abstraction divorceable from any actual concrete system, and as such, it can be multiply realized—embodied in various kinds of material systems. The concrete material is inessential to mental activity, which gets compared to the running of software.

A thing to appreciate here is that the digital computer is a machine precisely engineered so that its organizational properties float free of many of the specifications of its hardware.[4] This fact is spelled out very clearly by Simon in his *Sciences of the Artificial*:

> No artifact devised by man is so convenient for this kind of functional description as a digital computer. It is truly protean, for almost the only ones of its properties that are detectable in its behavior (when it is operating properly!) are the organizational properties. The speed with which it performs it basic operations may allow us to infer a little about its physical components and their natural laws; speed data, for example, would allow us to rule out certain kinds of "slow" components. For the rest, almost no interesting statement that one can make about an operating computer bears any particular relation to the specific nature of the hardware. (1969, 18)

The point is that digital computers are unique among devices in that their operations (i.e., the tasks they were built to perform) can best be understood at a level abstracted away from concrete hardware; and this is *because* they were designed with this aim in mind. The material details of implementation have no relevance to most of the questions that pertain to the operation of the machine as an executioner of algorithms. In the light of the purposefulness behind the clean separation between hardware and software in such machines, it seems a little crazy to think that an organ of the body—which is evolved and not designed—would have converged on exactly this degree of material indifference. And yet this is the assumption made by the functionalist who believes that in the human being, as much as in the computer, "it is the organization of components, and not their physical properties, that largely determines behaviour" (Simon 1969, 22), and as such, that the computer is not only a model, but an instantiator of thought.

The presumption that both evolution and artifice have converged on this separation of hardware/software levels appears less crazy when related to a

4. This is not the case for analog computers (Maley 2021).

prima facie defensible claim about where to find simplicity in nature. As will be discussed in more detail in the next section, the idea is that even a system as complex as the human being would be arranged hierarchically in approximately autonomous levels, such that the comprehension of a higher level, like that of cognition, is possible with minimal knowledge of the lower-level entities and activities that constitute it.[5] This is one way that Simon makes the point about autonomy among levels in order to show how an artificial system could take on the behavior of a natural one, though sharing none of its material basis:

> Resemblance in behavior of systems without identity of the inner systems is particularly feasible if the aspects in which we are interested arise out of the organization of the parts, independently of all but a few properties of the individual components. Thus for many purposes we may be interested in only such characteristics of a material as its tensile and compressive strength. We may be profoundly unconcerned about its chemical properties, or even whether it is wood or iron. (1969, 17)

A crucial thing to notice here is that Simon talks, in the same breath, of an order in nature, due to the independence of outer behavior from most of the properties of "inner systems," and the irrelevance *to the purposes of the modeler* of most material details. Indeed, Simon's book is victim to a long-running conflation of these, at many points picturing abstractions (which are by definition the product of a modeling procedure) as if they are part of the furniture of the world.[6]

There are obvious reasons why we need to attend to the difference between irrelevance of some details to a modeler, and irrelevance tout court. No two people's brains have the same number of neurons, and no one neuron in my brain would be fully functionally equivalent to a neuron in yours. And yet these particularities in neuronal details would be irrelevant to almost all modeling projects since a model that did attend to such details would be of little use, precisely because it would not generalize. It would take an enormous effort to build, with no payoff in inductive generality, and with

5. This hypothesis is revisited by Ballard (2015). See Chirimuuta (2022a) for further discussion.

6. The following sentence nicely exemplifies this ambiguity: "This skyhook-skyscraper construction of science from the roof down to the yet unconstructed foundations was possible because the behavior of the system at each level depended on only a very approximate, simplified, abstracted characterization of the system at the level next beneath" (Simon 1969, 17).

predictive utility restricted to the one person that it targets, at a particular period in time (given the brain's changeability). However, those differences are certainly not irrelevant to us as people: it matters to me that my ideas, decisions, and actions change from one year to the next; I am an individual, not identical to you, because of material particularities that are of no import to the modeler. The idea that every person's brain is, in its details, different from everyone else's has significance, even if those details will be left unregistered by science. Without them, there could be no uniqueness of personality and memory. The same observations apply to targets of modeling other than the human brain. The question always needs to be raised: in whose interests, from what perspective, are we claiming that the details don't matter?

We therefore recognize that functionalism is committed to a certain idea about how to simplify the brain via the independence of functional organization from material realization. Wary of committing the fallacy of misplaced concreteness, I decline to project onto the fabric of the world a convenient simplification that could at best hold only approximately. The rejection of functionalism, as well as the assertion that for the actualization of human and animal cognition in all its richness (as opposed to a crude and partial imitation), the material details probably do matter, lead me into the territory of biological naturalism. Biological naturalism denies the functionalist tenet of multiple realizability, the notion that there could be minds instantiated in radically different, inorganic substrates. A positive view remains to be stated.

The regular dialectical opponent of functionalism is the mind-brain identity theory—the view that the mind just is the brain since descriptions of mental states and processes are reducible to descriptions of neural ones.[7] Although the identity theory is a species of biological naturalism, I reject it along with functionalism, for it also involves the dubious ossification of a scientific simplifying strategy. Reductionism was the mode of simplification under scrutiny back in chapter 3. Its bet is on there being stable, elementary components of the system that do not interact in complicated ways, such that knowledge obtained about the parts is foundational to explaining the

7. An early statement is from Smart (1959), and a more recent defense is by Polger (2006).

behavior and properties of the whole.[8] We saw in chapter 3 that the reductionist methodology, with its obvious departure from the observable facts concerning the context dependency of activity in the nervous system, under normal ecological conditions, is justifiable only on instrumentalist grounds. A perspective so narrowly justified is not a viable basis for a philosophical theory of mind and body. And yet this narrowness infects philosophical accounts that subscribe to reductionism, with their spurious claims that pain could be C-fibers twitching, or depression a chemical imbalance. Reductionists are in denial about what these so apparently are; namely, complex, multifaceted, neural, psychological, and social conditions.

One way to characterize the trajectory of theoretical neuroscience in the twentieth century, from reflexology to computationalism, is as a growth in the sophistication of its simplifying strategies, from an implausible reductionism to a computationalism granting that there is a certain kind of "organised complexity" (Simon 1969, 86n4) in the brain and nervous system. But as has been argued since chapter 4, the kind of complexity compatible with the computational framework will still leave too much of the mind-brain system unaccounted for. Thus, the version of biological naturalism I endorse is one that does not rest on the simplifications of reductionism or functionalism. I will refer to it as *embodied mind*,[9] and I will return to it toward the end of the chapter, following an argument that the functionalist's favored simplifications lend themselves to dualist and skeptical results.

One last point before moving on is to say that a lesson of my approach is that it is a mistake to equate multiple realization (substrate independence of function) with degeneracy and structural variation in organic systems.

---

8. This reductionism makes the central nervous system a simple system, in Simon's terms. Churchland (1994) outlines a sophisticated reductionist methodology for neurobiology that allows for some top-down investigation.

9. For brevity, I'm labeling the view as *embodied mind* rather than *embodied and embedded mind*. As will become clear in section 10.5, it is just as important to reject the assumption of near-independence of mind or brain from environment (asserting embeddedness) as it is to assert embodiment, rejecting the assumption of near-independence of mind or brain from the body. See Ward and Stapleton (2012) for the argument that the embodied, embedded, and enactive views of cognition are mutually supportive. While I do not discuss action and cognition in this chapter, my view is consistent with the enactivist idea that these two are fundamentally linked. This is discussed in the account of understanding presented in the previous chapter, in section 9.5.

Across and within living species, we continually encounter variations on a theme, where virtually identical functions are carried out by notably different structures and processes. Within the organism, including the nervous system, degeneracy—the maintenance of vital functions in spite of turnover of components and recalibration of processes—has been commonly observed and thought to be ubiquitous in evolved systems (Edelman and Gally 2001). That something persists, across the flux, is a general characteristic of living organisms (Dupré and Nicholson 2018). The combination of functional stability and material plasticity is omnipresent in the living world. But this gives no grounds to say that function can float free of material realization, and that it is transferable into inorganic material, as the thesis of multiple realizability would have it. It does mean, however, that the theorist of embodied mind need not be worried by the observations that were problematic for the identity theory (e.g., that it seems right to say that an octopus could feel pain even without having any of the C-fibers of mammals). These phenomena of variation and degeneracy justify attention—where explanatory context permits—to some coarser-grain stabilities encompassing a range of more or less subtly different tokens. Moreover, the phenomenon of convergent evolution may justify attribution of same or similar functions to neural structures in phylogenetically distant animals like humans and octopuses (Godfrey-Smith 2016b). This means that fine-grained details, varying from one individual to another and throughout the lifetime of an individual, need not always be privileged according to the theoretical agenda of embodied mind; it is just that they should not be excluded from consideration at the outset, as the functionalist would prefer. It may well be that these coarse-grained stabilities provide a foothold for some kinds of simplifications, but it will not be the radical form of abstraction that Simon argues for, and to which we now turn.

## 10.3 The Assumption of Near-Decomposability

I am following Haugeland in his characterization of the theory of embodied and embedded cognition as the rejection of the idea that a certain kind of simplicity is to be found in the brain and nervous system—that the mind itself is, and is part of, a near-decomposable system made of semi-independent components whose pattern of organization is responsible for the sophisticated behavior of the system. To better understand the contrast between embodied

mind and the tradition it breaks from, I will now discuss the assumption of near-decomposability in more detail.

The first thing to appreciate is that the assumption of near-decomposability is a path to scientific understanding. As Haugeland writes: "*Finding*, in something complicated and hard to understand, a set of simple reliable interfaces, dividing it into relatively independent components, is a way of rendering it *intelligible*" (1998b, 216). For example, if you take an organism and see no more than an undifferentiated whole, you have little hope of accounting for the ways that it is sensitive to its environment, how its behavior is generated, and how self-maintenance occurs. In contrast, appreciating that its structure is differentiated into organs, and positing that these are somewhat independent, licenses more focused investigation of those components and the possibility of grasping the principles of their joint operation (Bechtel and Richardson 2010). To borrow Haugeland's own example, a television set would be unintelligible if cut up into 1-centimeter cubes. But breaking it down systematically into likely component parts creates a chance of understanding how it works.

In a near-decomposable system, components are defined by "intensity of interaction" (Simon 1969, 90). A component is a part of a system, such that the number of interactions within the part is an order of magnitude higher than the number of interactions that a part has with others in the system (Simon 1969, 99).[10] This is why we are to think of components as semi-independent from one another. Embodied mind is the denial of the thesis that the mind, brain, and rest of the body are components in this sense. It denies the quasi-independence of the mind from both lower-level subcomponents and the systemwide context in which it is embedded.

One way to think about Simon's notion of a component is to say that from the perspective of the wider system, each component is a *black box*. As such, its place in the system is clearly defined in terms of its function, the inputs it can receive and the outputs it will generate, but its inner workings—the procedure by which this input-output relationship is maintained—do not

10. This allows Simon then to sum up his approach in two propositions: "(a) In a nearly decomposable system the short-run behavior of each of the component subsystems is approximately independent of the short-run behavior of the other components; (b) in the long run the behavior of any one of the components depends in only an aggregate way on the behavior of the other components" (1969, 100).

matter. This affords the investigator of the system a handy simplification. In order to understand the operating principles of the system, she need only characterize the function of each component, deferring the specification of their inner mechanisms. Moreover, there is a layered picture of "boxes-within-boxes," such that components have subcomponents, but likewise the details of those submechanisms can be black-boxed, allowing the same abstraction to occur at the various levels of the system. This is how Simon describes the situation:

> The basic idea is that the several components in any complex system will perform particular sub functions that contribute to the overall function. Just as the "inner environment" of the whole system may be defined by describing its functions, without detailed specification of its mechanisms, so the "inner environment" of each of the subsystems may be defined by describing the functions of that subsystem, without detailed specification of its submechanisms. (1969, 73)

It is important here to note the connection to the computer, which as we saw previously conforms more closely than anything else to the ideal of a system in which only functions, not mechanistic or implementational details, are relevant to its behavior (Simon 1969, 18). But to reiterate the point from section 10.2, it would seem outlandish to presume that a brain could be like a computer in this respect.

In the course of the well-known essay on the "Architecture of Complexity," Simon offers an evolutionary argument as to why we might nonetheless expect this. It is an argument based on an analogy with the process of watchmaking: it would be near impossible to make a watch by hand if, each time the phone rang and the work were interrupted, all the pieces assembled up to that point fell apart in a heap—far better to have a process of manufacture in which stable subassemblies are made separately, then pieced together to form more complex systems. Since, as Simon argues, the same requirement of there being stable subassemblies would apply to the evolution of organisms, one should expect to find, in nature, hierarchical complex systems made up of stable, semi-independent components and subcomponents (Simon 1969, 90–94). A problem with this argument is that even if granted the point that evolution must involve the "assembly" of more complex organisms out of simpler, but still stable and viable ones (as in the major evolutionary transition from unicellular to multicellular life forms), this does not mean that stable subassemblies must have the characteristics outlined in Simon's definition of a component. In particular,

there need not be a low intensity of interaction beyond the confines of the evolved subassembly, and it may well not be the case that the subassembly can be characterized functionally, without reference to material details. These putative properties of subassemblies are, of course, valuable to a human being, such as a watch-maker, who makes things from material components. It is undesirable for any component subassembly to have many modes of interaction with its context or to be too sensitive to fine grained material conditions, as this will increase the chance that it will depart, in unforeseeable ways, from the behavior required by the design.

In chapter 8, we encountered Descartes among the early scientists whose explanatory program presupposed the denial of any fundamental difference between natural and artifactual objects. This is a simplifying assumption because machines and other technological entities are less complex than organisms, but they afford a model, a simplifying lens through which to view the works of nature. We can call this a "Cartesian idealization" and recognize that Simon is employing it both in his evolutionary argument, which takes natural selection to fall under the same constraints as the human process of design, and his promotion of the computer as the model for the mind, with its characterization of functional systems floating free of material concerns. In the epigraph to this chapter, we see that the convenience of positing isolated systems is already noted by Descartes. Simon's positing of almost-independent components is a species of this sort of Cartesian idealization.[11] Cartesian dualism is the notorious view that mental substance (*res cogitans*) is radically different from physical substance (*res extensa*).[12] If we bracket the ontological commitments of Descartes's

11. I do not mean to suggest that Descartes holds the copyright on these idealizations. Rather, they are characteristic of a tradition of physical science that has in turn shaped the course of cognitive science and neuroscience more recently. Indeed, mind-body dualism, of some sort, is much older than Descartes. My deployment of the term "Cartesian" is self consciously polemical, and as with all such polemics, it risks making a caricature of the historical figure (Roux 2013). I emphasize that in talking of "Cartesian idealization," I am not claiming that Descartes was the inventor or propagator of these idealizations. Importantly, the assumption of the self-containment of mind, which I call "Cartesian" for the purposes of this discussion, is probably not consistent with the notion of "intermingling" between mind and body in human beings, which Descartes invokes but does not properly theorize (Simmons 2011).

12. Descartes's own views on the mind-body relationship are more complicated than the standard reading, focused on the *Meditations*, allows. From other texts such as

own dualism, we see that the more tenacious dualist idea stems from the idealization of isolation, as Haugeland (1998b, 207) rightly puts it, of cognitions as "self-standing and determinate on their own, without essential regard to other entities." To treat the mind in this way, as conceptually separate from the brain, and the brain as separate from the body is to treat a person as a near-decomposable system

Before moving onto a closer examination of dualism, we should here acknowledge that Simon does at least float the idea that complex natural systems appear to be near-decomposable because that is the way that humans can understand them. I quote this important passage at length:

> The fact then that many complex systems have a nearly decomposable, hierarchic structure is a major facilitating factor enabling us to understand, describe, and even "see" such systems and their parts. Or perhaps the proposition should be put the other way round. If there are important systems in the world that are complex without being hierarchic, they may to a considerable extent escape our observation and understanding. Analysis of their behavior would involve such detailed knowledge and calculation of the interactions of their elementary parts that it would be beyond our capacities of memory or computation.
>
> I shall not try to settle which is chicken and which is egg: whether we are able to understand the world because it is hierarchic or whether it appears hierarchic because those aspects of it which are not elude our understanding and observation. I have already given some reasons for supposing that the former is at least half the truth that evolving complexity would tend to be hierarchic but it may not be the whole truth. (Simon 1969, 108)

One thing to point out is that a possibility not considered by Simon here is the one frequently argued for in this book: that simplicity (e.g., in this case, near-decomposability) is not merely *found*, and intractable complexity *ignored*, but that simplicity, instead of being discovered, is *projected, massaged, or in some way elicited from the world*. Many systems can be assumed to be, to some degree of approximation, near-decomposable. In addition, experimental methods can suppress the channels of interactions among parts of systems, making them better conform to Simon's definition of a component. What must be remembered is that the success of scientific work that

---

*Traité de l'homme*, it is possible to recover an "embodied Descartes," in which the body by itself is endowed with flexibility and intelligence. But this does not disrupt the core point at issue in this chapter, which is that "Descartes does take the mind and the body to be radically distinct—and to be fully separable, at least in principle" (Hutchins, Eriksen, and Wolfe 2016, 301). See also Simmons (2011).

represents systems as near-decomposable should not be taken as evidence that this is simply how things *are*.

## 10.4 Dualism

> Historically the modern theory of transformational linguistics and the information-processing theory of cognition were born in the same matrix of ideas produced by the development of the modern digital computer, and in the realization that, though the computer was embodied in hardware, its soul was a program.
>
> —Herbert Simon (1969, 47)

Various contemporary philosophers, and even neuroscientists such as Antonio Damasio, have cast Descartes as the originator of a pernicious idea about the radical difference between mind and body—an idea that has weedlike tenacity, which many have attempted to dig out once and for all, but which always seems to grow back from the fragments left in the soil. Gregory McCulloch (1995) is one philosopher who, in his own effort to root out dualism, pinpoints the modern manifestation as resting in the notion that minds are "self contained."[13] It is easy to buy into a picture of Descartes bequeathing a conception of the mind to future generations, a poisoned inheritance that has persisted long past substance dualism, morphing into different versions, and because it is somewhat changed, it goes undetected and is almost irresistible. The problem with this diagnosis of dualistic thinking as the result of an individual philosopher's influence is that it fails to consider that there may be broader and still active causes of its appeal. What is left unconsidered is the possibility that dualism is symptomatic of the wider tendencies of the scientific culture that Descartes, among others, represents, and it persists not because of the long shadow of one philosopher, but because the essentials of this intellectual culture remain.

That the supposed "self containment" of the mind is the idea most significant to philosophy and science today is, of course, consistent with the analysis put forward by Haugeland (1998b), and endorsed in this chapter. The treatment of mind and body as different *components* of a near-decomposable

13. "The idea that minds are more or less self contained with respect to their material surroundings continues to exert a powerful influence in contemporary philosophy, psychology, cognitive science and artificial intelligence design" (McCullough 1995, 16–17).

system is one instance of this idea. The key assumption is that there is *relatively* little interaction between these two components—the interface between them is "narrow-bandwidth" (Haugeland 1998b, 220)—such that the mind can be characterized in terms of its rich internal organization (within component interactions), plus the small number of interactions it has with the components outside it via its input and output channels. The coupling between soul and body at the pineal gland suggests a narrow bandwidth interface; but even the contemporary image of the brain in a vat runs on the assumption that the brain (now taking the place of the mind) is hooked up to the rest of the world via a relatively small number of nerve fibers, so that the input and output communications typical of an embodied brain could be re-created artificially. Thus we see that the treatment of a human being as a near-decomposable system grounds the conception of the mind (or brain) as a separable component, which is one way to express the dualist commitment to the self-containment of the mind.

The additional contribution of my analysis is the point that the demand of science to make complex systems intelligible by imposing simplifying assumptions will create a pressure toward treating human beings, and other creatures, as if near-decomposable systems, with self contained minds, even when they are not. Whereas Haugeland (1998b, 228–229) treats it as a straightforward matter of empirical discovery that a picture of the integralness of brain and body will be favored once there is a weight of neuroanatomical evidence in support of the view that the interface between them is extremely "high-bandwidth," my thought is that even given these well-known facts, the pressure remains for scientists to idealize away from them, retaining the picture of cleanly separable systems and subsystems. The richness and breadth of the interconnections between the nervous system and all the other bodily systems—immune, endocrine, digestive, muscular, skeletal—are more than can be encompassed from any one modeling perspective that aims at a minimum of clarity and precision. Thus the tendency toward a form of dualism will remain.

The choice of Herbert Simon as the representative scientist is convenient because in his writings, the connection between dualism, functionalism, and techniques of abstraction is present in a uniquely salient way. The following passage exemplifies the way that functionalism turns out to be a version of dualism:

> I have discussed the organization of the mind without saying anything about the structure of the brain.
>
> The main reason for this disembodiment of mind is of course the thesis that I have just been discussing. The difference between the hardware of a computer and the "hardware" of the brain has not prevented computers from simulating a wide spectrum of kinds of human thinking just because both computer and brain, when engaged in thought, are adaptive systems, seeking to mold themselves to the shape of the task environment. (Simon 1969, 54)

In this passage, the simplification at work is the assumption that computers and brains are adaptive systems, and as such, their behavior is explainable by reference to the externally imposed constraints to which they must conform.[14] This assumption permits factors such as material constitution and processes in the body to be excluded from consideration, which is a convenient way to reduce the number of variables to be studied, and of course also justifies the use of the relatively simple computer as a model for the brain.

However, it might be objected that the legacy theory closest to functionalism is not dualism but hylomorphism.[15] The identification of Aristotle's hylomorphism as the precursor to functionalism is made at the start of various introductory texts (e.g., Block 1980) and was promoted by Hilary Putnam, one of the early proponents of functionalism; it has generated its own controversy among scholars of Aristotle (Rorty and Nussbaum 1992). While I grant that the characterization of mind as form, or organization, is shared between functionalism and hylomorphism, that does not detract from the point crucial to my discussion: to wit, in functionalism, the intrinsic features of the body are inessential to what the mind is, as opposed to the functional roles of constituent parts of the body. For the functionalist, any system made from components duplicating those functional roles could have a mind, such that in principle, mind is possible without embodiment and mental life has nothing inherently to do with being alive.

14. *Pace* Haugeland (1998b, 210), I do not take Simon to be saying that humans are so different from ants in that, as apparent in the quotation here, our complexity is also said to be driven by external factors. This is the "environmental complexity thesis" described by Godfrey-Smith (1996), and comparable to the hypothesis of adaptationism in evolutionary biology.

15. Briefly, hylomorphism uses form as the basic ontological and explanatory principle in the philosophy of mind. Jaworski (2016) is a recent proponent.

As Burnyeat writes, "The whole point of functionalism is to free our mental life from dependence on any *particular* material set-up" (1992, 17; emphasis in original). But, he argues, it is a mistake to take Aristotle's form-matter distinction as lending itself to this kind of independence, where mind is the form and brain is the matter; rather, hylomorphism asserts an *interdependence* of form and matter, or as Burnyeat puts it, the assertion is that the forms essential to organisms are not contingently related to matter:

> Life and perceptual awareness are not something contingently added to animal bodies in the way in which shape is contingently added to the bronze to make a statue. Aristotle states explicitly in [*De Anima*] 2.i that the only bodies which are potentially alive are those that are actually alive. A dead animal is an animal in name alone. And this homonymy principle is no mere linguistic ruling. It is a physical thesis to the effect that the flesh, bones, organs, etc. of which we are composed are *essentially* alive, *essentially* capable of awareness. (1992, 26; emphasis in original)[16]

It is also interesting that Burnyeat identifies the artifact analogies used by Aristotle to illustrate hylomorphism (such as the bronze statue) as having the misleading connotation that form and matter may only be contingently related in living beings, arguing that this analogy should not be read into *De Anima*.[17] It is the denial of the interdependence of mind and body that is the crucial commonality between functionalism and dualism, and the construal of the artifact as a model for the living organism was noted by me already as one of the relevant Cartesian idealizations. Nussbaum and Putnam (1992, 33), in their response to Burnyeat, uphold the contingency of matter-form relationships and argue that "plasticity"—the many-to-one relationship between structure and function—in living organisms speaks to the appropriateness of this assertion when theorizing mind and brain. But as argued previously, it is a mistake to equate these phenomena of plasticity, degeneracy, and convergent evolution with the much more radical independence of multiple realizability maintained by functionalists.

It is worth pausing to note the underlying interconnection between our two Cartesian idealizations. The assertion that an artifact could be a perfect model for an organ or organism (i.e., one not suffering from glaring

16. See Whiting (1992, 85–88) for a related discussion.

17. "There are in any case strong independent grounds for rejecting, where proper substances [e.g., organisms] are concerned, the artefact model and the idea of a merely contingent relation between matter and form" (Burnyeat 1992, 26).

and misleading disanalogies) is a commitment to the idea that an organ like the brain could operate, in its essentials, in the same way as a machine like a computer does. Descartes is well known for arguing in his *Treatise on Man* that a living body could operate, in its essentials, in the same way as a machine because bodies are mechanisms. Independently of the question of whether Descartes was historically the most important propagator of this idea, it is true that this is now a prevalent, if not dominant, conception of the body and its systems. But what does it mean to say that the body is a suite of mechanisms? Among the many characterizations of mechanism in the literature, one feature is particularly significant to our study. It is that mechanisms are assemblages of parts that are in principle separable from one another. The components of mechanisms are *partes extra partes*—things all sitting externally to one another with no inherent connection among them.[18] They interact in limited, clearly specifiable ways, being components in Simon's sense. Indeed, the parts of artificial mechanisms must be that way, or else they could not be assembled. So the notion of this isolated form of existence—of entities that are not inherently dependent on or constituted by what is beyond their outer boundaries, even if, as a matter of empirical fact, they always occur within particular contexts—can be found at the root of the idea of mechanism, of decomposable systems, and hence of the Cartesian mind, with the mind not inherently dependent on the body. The assumption of isolated existences is a prerequisite for the kind of conceptual clarity demanded in scientific thought. If the assumption is not employed, the boundaries around objects of study remain vague and indeterminate, and we are beginning to deal with a worldview in which resonance and mutual influence replace workable relationships of demarcated cause and effect, and where relationality has precedence over entities related. The mindset of dualism, which is the mindset of the so-called mechanistic worldview, is the rejection of this outlook.[19]

18. See Guttinger (2018, 306), reporting on the ideas of biologists Birch and Cobb:

> In a mechanical system. . . . the nature of an entity is not affected by the relations it has with other things or processes. The cogwheel or the steel rod are not affected in their nature by their (external) relations or by the change (turning, expanding, contracting) they undergo. The way they react to changes in their context is set by their material constitution. In an ecological system, *what a thing is depends on the relations it has* (emphasis added).

19. The contrast between these two worldviews is the theme that runs through Hesse's *Forces and Fields*. Action at a distance was associated with obscure modes of influence,

There are some philosophical precursors to the diagnosis of dualism as being fostered by the scientific preference for dealing with neatly separated systems. In the late nineteenth century, mind-body parallelism was a dualist theory popular among scientists, including the neurologist Hughlings Jackson (Chirimuuta 2017b). The idea was that mental states ran along a parallel track in synchrony or "concomitance" with the series of states of the nervous system, but without mutual influence. William James was scathing about concomitance, which he called an "utterly irrational notion" (1890/1950, 136). Name-calling aside, James makes the helpful suggestion that this view, which requires commitment to the unlikely scenario of "absolute separateness" together with perfect correspondence of mental and neural states, gains its appeal because it offers to scientists conceptual neatness and the chance to exclude hazy mental factors from their investigations:

> The desire on the part of men educated in laboratories not to have their physical reasonings mixed up with such incommensurable factors as feelings is certainly very strong. I have heard a most intelligent biologist say: "It is high time for scientific men to protest against the recognition of any such thing as consciousness in a scientific investigation." In a word, feeling constitutes the "unscientific" half of existence, and anyone who enjoys calling himself a "scientist" will be too happy to purchase an untrammelled homogeneity of terms in the studies of his predilection, at the slight cost of admitting a dualism which, in the same breath that it allows to mind an independent status of being, banishes it to a limbo of causal inertness, from whence no intrusion or interruption on its part need ever be feared. (James 1890/1950, 134–135)

I should also mention that in his presentation of the fallacy of misplaced concreteness, Whitehead (1925/1967) argues that the demand for mathematical precision in physical science makes fertile ground for dualism because the quick way to deal with all that is qualitative, and as such, refractory to precise definition, is to exclude it from matter and impose it on mind.

---

whereas the restriction to action-by-contact came with a conception of material bodies as impassive, bounded entities only able to be affected by immediate impulse. As she notes:

> The preference for action-by-contact theories in physics was historically connected with the objectification and depersonalisation of nature and the desire to eliminate from explanations of it the "psychological" analogies of organism, command, and attraction in favour of the analogy of mechanism, and it was a fact that most familiar mechanical devices acted by contact. (Hesse 1962, 291)

In Cassirer's *Philosophy of Symbolic Forms*, there is a variant of this account, which I will revisit at the end of this chapter. It is not science per se, but the tradition of metaphysical inquiry from which science emerged, that fosters dualism. According to Cassirer, it is the conceptual clarity and determinateness demanded by metaphysical theorizing that generate the mind-body problem out of what, phenomenologically, presents itself in untutored experience as the basic unity of mind and body:

> The body-soul relationship proves ever and again elusive, regardless of whether thought seeks to catch it in the meshes of an empirical causality or of a purely intelligible determination. For every kind of determination makes body and soul appear as two independent, self-subsistent entities, one of which is conditioned and determined by the other: and the peculiar mode of "in-volvement," of mutual interwovenness disclosed by the body-soul relationship, never ceases to resist this form of determination. (Cassirer 1929/1957, 99)

Cassirer observes that Aristotle is not as far down this path of separation as later metaphysicians since for him, "the soul is still the entelechy of the body and thus its most proper 'reality'" (1957, 103). We might add to this remark that hylomorphism is exemplary of the way that Aristotle's philosophy is a product of the twin demands of systematic theorizing and faithfulness to what is available in sensory experience.

With my favored analysis in place, we can end this section with a deeper understanding of the ways in which contemporary, materialist philosophies of mind are dualistic. Searle rightly finds fault with materialist theories, such as functionalism, for their acceptance of the Cartesian categories of the mental versus the physical:

> The weird feature about this entire discussion is that materialism inherits the worst assumption of dualism. In denying the dualist's claim that there are two kinds of substances in the world or in denying the property dualist's claim that there are two kinds of properties in the world, materialism inadvertently accepts the categories and the vocabulary of dualism. It accepts the terms in which Descartes set the debate. It accepts, in short, the idea that the vocabulary of the mental and the physical, of material and immaterial, of mind and body, is perfectly adequate as it stands. (1992, 54)

However, this does not account for why the terms of the debate remain so intuitive and appealing, other than with the hint that Descartes arranged things that way, and no one has bothered to give them an overhaul in the meantime. The more plausible explanation is that dualism has something

else to offer, and for this reason it has been reinvented many times. Yet, if the division of substances and properties into two schedules under a mental and physical heading is itself a convenient way to clarify terms and simplify the subject matter of science, perhaps it is not so dispensable after all. We will return to this issue in section 10.6, after a short examination of skepticism and disjunctivism.

## 10.5 Skepticism

The habit of thought underlying dualism takes the world to be made up of items not inherently related to one other, which are more or less, in principle, isolatable. These separate entities are linked to one another by cause and effect, but not by the deeper bond of constitution. The mind as separate from the body, and the body as separate from the world, are just two instances of a more generalized picture. However, the separateness of mind from everything else is marked out from other cases in its generating a unique set of philosophical concerns. We have already examined dualism, which creates the puzzle of how mind and body could be so tightly synchronized with one another if so radically different. Skepticism shows even more clearly how a problem arises with the assumption of isolatability and how it can be resolved by removing that assumption. The Cartesian skeptical predicament is of a mind absolutely isolated but deceived into thinking that it perceives an external world by the manipulations of an evil demon, or—in the updated version of the thought experiment—a mad neuroscientist tweaking the nerve impulses sent into a brain in a vat. Once, conceptually, the mind is cut off from the rest of existence, it can in principle only have certain knowledge of its own contents—that it is experiencing a sensory perception, but not that there is anything in the world beyond the confines of the mind that the experience is a perception of.

That the Cartesian predicament is more general than an epistemological puzzle has been appreciated elsewhere. At the start of *Mind and World*, John McDowell writes of "an inchoately felt threat that a way of thinking we find ourselves falling into leaves minds simply out of touch with the rest of reality, not just questionably capable of getting to know about it. A problem about crediting ourselves with knowledge is one shape, and not the most fundamental, in which that anxiety can make itself felt" (1996, xiii–xiv).

We will now see how *disjunctivism*, one strand of McDowell's response to the worry about the failure of the mind to make contact with the world, is in essence a denial of the assumption of isolatability, for disjunctivism asserts that in the good cases, where perception affords knowledge of things in the world around me, those things do not merely cause but also constitute my mental state. We will see that McDowell's case for disjunctivism converges on the central points of this chapter. However, disjunctivism has faced condemnation from Tyler Burge for alleged incompatibility with perceptual science. I will argue that the incompatibility is not the one that Burge takes it to be, and that there are indeed good reasons for philosophy to pursue inquiries detached from the conceptual frameworks of the sciences.

### 10.5.1 Disjunctivism

Disjunctivism is a theory in the philosophy of perception that states that where you have a veridical perceptual state (e.g., of seeing a purple balloon drift past your window) that has an illusory or hallucinatory counterpart indistinguishable to the subject, even though the veridical and nonveridical states are subjectively indiscriminable, it is *not* the case that they have the same epistemic significance (McDowell 2013, 259–260, 263). Disjunctivism is most often presented as a solution to the problem posed by illusions and hallucinations to the naive realist theory whereby veridical perceptual states involve a relation of acquaintance with the external object of perception (Soteriou 2020; Crane and French 2021). For this reason, it is not always made obvious that disjunctivism, at least on McDowell's account, is in essence a response to Cartesian skepticism.[20]

However, the connection between self-containment and skepticism and the rejection of these afforded by disjunctivism is quite clear in McDowell's presentation. The idea at fault, according to McDowell (1998, 242) is of a "self-contained subjective realm, in which things are as they are independently of external reality." In such a view, the mind just seems to make no contact with the external world. McDowell uses various locutions to describe the account that he opposes: it is of "the inner realm autonomous" in which

20. I am only considering McDowell's version of disjunctivism. See Byrne and Logue (2009), Haddock and MacPherson (2011), and Soteriou (2016) for surveys of the topic.

"we deny interpenetration between inner and outer" (1998, 245); it is "a conception of a realm whose layout is independent of external reality" (1998, 257). With this isolation of the mind, the idea that perception could give it access to things around it becomes doubtful, and hence radical skeptical scenarios rise up as coherent possibilities. Most of the recent post-Cartesian epistemologies accept the fallibility of perceptual knowledge and neglect the root issue.[21]

McDowell concurs with the view defended in this chapter—namely, that what is most fundamentally problematic about the Cartesian frame in philosophy of mind is its positing of the self containment of mind, not its ontology of mental substance—and McDowell also sees this as a fault within functionalism even though it is a materialist theory (1998, 246). Moreover, he pinpoints the demands of scientific causal explanation as giving the initial impetus for the Cartesian separation of the mental as a self standing explanandum:

> It seems scarcely more than common sense that a science of the way organisms relate to their environment should look for states of the organisms whose intrinsic nature can be described independently of the environment; this would allow explanations of the presence of such states in terms of the environment's impact, and explanations of interventions in the environment in terms of the causal influence of such states, to fit into a kind of explanation whose enormous power to make the world intelligible was becoming clear with the rise of modern science, and is even clearer to us than it would have been to Descartes. (McDowell 1998, 243–244)

Functionalism reinhabits the same explanatory framework, identifying the "autonomous explanatory states" with organizational states of the nervous system. Again, the point is not that functionalism is directly influenced by

21. The anti-Cartesian agenda behind disjunctivism is recapitulated in McDowell's first response to Burge's attack on disjunctivism:

> We can express the idea with a disjunction: an appearance is either a case of things being thus and so in a way that is manifest to the subject or a case of its merely seeming to the subject that that is how things are. If we go on regarding appearances as elements in a subject's inner world, this disjunctive conception embodies a recognizably non-Cartesian conception of that world. When a state of affairs that conforms to the first of those two disjuncts is an element in a subject's inner world, how things are in that world cannot be fully specified without a commitment as to how things are in the subject's environment. On this conception, a subject's inner world does not have the characteristic Cartesian independence from the outer world. (McDowell 2010, 244)

the historical Descartes, but that the same "fundamental motivation" is shared between Cartesianism and functionalism.[22]

McDowell's point of arrival also bears similarities to the view defended in this chapter, of mind embodied and embedded whose capacities are inherently due to its belonging to a living body. There is acceptance of Searle's biological approach, but rejection of the identification of mind with brain: "It is an insight on Searle's part that intentionality is a biological phenomenon. But intentionality needs to be understood in the context of an organism's life in the world. We cannot understand it, or even keep it in view, if we try to think of it in the context of the brain's "life" inside the head" (McDowell 1998, 258, n57).

Like Haugeland, McDowell rejects the usual ways of placing boundaries on the cognitive, and he sees this as a direct implication of disjunctivism: "Allowing intrinsic object dependence, we have to set whatever literally spatial boundaries are in question outside the subject's skin or skull. Cognitive space incorporates the relevant portions of the 'external' world" (McDowell 1998, 258). The striking difference between Haugeland's and McDowell's presentation of the view is that the former, but not the latter, makes ample reference to ideas and results from neuroscience and cognitive science. Indeed, McDowell (2013) asserts that his project is tangential to those activities.

This claim for the autonomy of philosophical studies of perception is what most seems to have exercised Burge (2005), and his attack on disjunctivism boils down to the charge that it has been refuted by empirical findings since the science requires, but disjunctivism denies, a "common factor" between subjectively indistinguishable veridical and nonveridical perceptual states. The controversy between Burge and McDowell is a collision of large-scale

22. "Now this intellectual impulse is gratified also in a modern way of purportedly bringing the mind within the scope of theory, in which the interiority of the inner realm is literally spatial; the autonomous explanatory states are in ultimate fact states of the nervous system, although, in order to protect the claim that the explanations they figure in are psychological, they are envisaged as conceptualized by theories of mind in something like functionalist terms. This conception of mind shares what I have suggested we should regard as the fundamental motivation of the classically Cartesian conception; and I think this is much more significant than the difference between them." (McDowell 1998, 244)

In section 10.4, I focused on the functionalist separation of mind from the material brain. As McDowell points out, functionalism is equally committed to a division between the interior subject and the outside environment.

philosophical worldviews, which Fish (2021) has compared to a clash of Kuhnian paradigms. But the treatment of the controversy as a matter of competing research programs, analogous to scientific ones, neglects the crucial particularity of the case, which is that disjunctivism declines to define its explanatory objects in the way most conducive to scientific research. For this reason, there is more of a tension with science than McDowell admits—it goes beyond the situation of two parties holding orthogonal interests. Yet, as I will now argue, this does not invalidate disjunctivism.[23]

### 10.5.2 Inner States

Burge's work displays much familiarity with the details of experimental and theoretical research on perception, especially vision. I do not have the space here to review the many facets of the theory that he has developed out of consideration of these results. My focus here is on the way that he follows the science, tacitly, in its idealizations. Insofar as Burge subscribes to a *proximality principle*, he inherits the idealization discussed at length in this chapter, that of mind and environment being separate systems interacting with each other in relatively minimal ways so that *particular* perceptual states can get an adequate characterization in terms of factors within the organism (internal to the system).[24] This leads Burge to enforce the separation between mind and world, disallowing the interpenetration of inner and outer, which is exactly the move that disjunctivism is set up against.[25]

To see how this Cartesian idealization follows directly from Burge's incorporation of the mainstream computational theory in perceptual science, it is worth quoting him at length:

---

23. I discuss the controversy at greater length elsewhere (Chirimuuta 2022c).

24. Burge's (2010, 61ff.) externalism or anti-individualism requires that certain general relations hold between the causal structure of the environment and certain *types* of perceptual processes and states. However, these general conditions do not negate the proximality principle—the point that any particular perceptual state is to be characterized by its internal (proximal) stimulus and relationships to other mental states.

25. Again, this observation is consistent with Burge's externalism or anti-individualism: "Anti-individualism *per se* does *not* claim that mental states *are* relations to the environment, or that mental states are not in the head, or that entities in the environment are part of the mental state or of the state's representational content. I reject these claims" (2005, 64; emphasis in original).

> The reason why the science's basic principles cite a common factor is that the kinds of perceptual states that are formed—including conscious state kinds that are the perceivings and misperceivings by individuals—depend purely on (a) the registration of proximal stimulation, (b) the antecedent psychological and physical states of the individuals, and (c) the quasi-deterministic laws of transition between registration of proximal stimulation and the perceptual states that are formed. This is a statement of what I call the science's "Proximality Principle". . . . Differences among the [subjectively indistinguishable veridical and nonveridical] cases are individuated by reference to the occasion-specific "distal inputs"—the causal chains that lead from the environment to the same registration of proximal stimulation. *The shared factor is separable from the unshared factors. It is separated by the science. Explanation of the formation of the perceptual states centers on that shared factor.* (Burge 2011, 44, emphasis added)

The "proximal stimulation" is the first reception of a stimulus at a sensory organ. It is the product of the transduction of the stimulus into a pattern of activity in the nervous system. The proximal stimulus (e.g., the pattern of light falling on the retina) contrasts with the distal stimulus, the object in the world that one would ordinarily think of as the target of perception, such as the surface that the light was originally reflected from. When comparing the veridical and nonveridical cases, Burge assumes a clean division between factors that occur before and after this moment of transduction. The proximal stimulation serves as the absolute divide between factors that are essential (inside the perceiver, from the point of sensory transduction onward) and inessential (the distal ones) to the characterization of a perceptual state, the explanation of how it is formed. Since the veridical and nonveridical cases differ only in their distal conditions, they are not different for the purposes of scientific explanation. The mistake of disjunctivism, as Burge sees it, is its neglect of the importance of this identity with respect to proximal factors.[26]

---

26. In chapter 6 I argued that the positing of neural representations invokes a nonproximal research heuristic, in which proximal, causal explanations are bypassed in order to attend to the relationship between the neural activation and the distal cause. This would seem to be inconsistent with the point here, which is that perceptual scientists (including sensory neuroscientists) are committed to a *proximality principle*. The apparent tension can be resolved once we appreciate that Burge's proximality principle is just one instance of a very general constraint on causal explanations, which is that proximal causes screen off distal ones. (Schematically, if the more distant cause *C* brings about an effect *E* via a more proximal intermediary cause *I*, then

Haugeland's essay contains an extended criticism of the assumption that sensory transduction marks the point of interface, the boundary between mind and world and that transduction in the opposite direction, from symbolic motor command in the brain to muscle movement, is the point of interface between mind and body. The view that he urges us to take up is one in which the signals, codes, or symbols, which are the hallmark of the cognitive domain and are supposed to reside solely on the inside of the transduction boundary, make sense and are decodable only in the context of bodily and worldly states. In other words, he rejects the supposition of "inner symbols" housed in the mind, which have their intentionality and meanings independently of happenings beyond the mind. This is how Haugeland makes the case for the radical dependency of the putative symbolic realm on bodily context, denying in principle the clean division, at the point of transduction, between distal and proximal factors:

> That some particular pulse pattern, on some occasion, should result in my typing an "A" depends on many contingencies, over and above just which pattern of pulses it happens to be. In the first place, it depends on the lengths of my fingers, the strengths and quicknesses of my muscles, the shapes of my joints, and the like. Of course, whatever else I might do with my hands, from typing the rest of the alphabet to tying my shoes, would likewise depend simultaneously on particular pulse patterns and these other concrete contingencies. But there need be no way to "factor out" the respective contributions of these different dependencies, such that contents could consistently be assigned to pulse patterns independent of which fingers they're destined for. That is to say, there need be no way—even in principle, and with God's own microsurgery—to reconnect my neurons to anyone else's fingers, such that I could reliably type or tie my shoes with them. (1998b, 225)

In a striking metaphor that conveys the way that the embodied mind theory rejects the division between what is cognitive (i.e., symbolic) and what is material (i.e., corporeal), Haugeland (1998b, 226) speaks of the body as a large and ever-changing encryption key for neural motor commands.

We have seen in chapters 5 and 7 that the default tendency of theoretical neuroscience has been to treat sensory and motor cortex responses as being fixed representations that always *mean* the same particular stimulus feature

---

the effect that *C* has on *E* can be no different from that of any other distal cause that works via *I*.) The nonproximal, representation positing explanations under discussion in chapter 6 do not actually depart from this general constraint; it is just that they decline to investigate the intermediate, more proximal causes.

or muscle movement. We also saw that this is a simplifying assumption that led to elegant theories of visual and motor cortex, but when set against data collected in long-term, naturalistic situations, it becomes clear that it is a strong idealization—indeed, the phenomenon of representational drift is gathering increasing attention as something that might shake up the foundations of theoretical neuroscience (Schoonover et al. 2021). The assumption holds that whatever goes on more widely beyond the brain is irrelevant to the significance of neural activity: so long as a neuron is made to fire, however the firing is caused, it will always *mean* the same thing. The result is that all the malevolent neuroscientist needs to do is cause the same set of neural activations that would occur in ordinary life, and the disembodied mind will be perfectly deluded by its sensory array. In short, the idealization of fixed representations implies that neural activations have meanings autonomously of anything beyond the brain, and this Cartesian idealization lends itself to Cartesian skepticism. This is also the view encapsulated in Burge's "proximality principle," and which Haugeland rejects, precisely because he doubts that transduction provides a hard border between the brain and its surroundings.

The upshot is that the clash between Burge and McDowell is generated by Burge's incorporation of a scientific framework that is itself in the business of making the Cartesian idealization of the separability of inner and outer factors. Burge is correct to recognize a tension—an incompatibility even—between the scientific framework and a disjunctivism that rejects its core assumption. But he is wrong to uphold the authority of that framework over independent philosophical inquiries. A philosophical position is not refuted because it is inconsistent with a thesis that is not a discovery, but a working assumption, of the empirical science. Burge (2011, 44) writes, "Science is our best guide to determining the basic natures of kinds that it describes and explains." Accordingly, for him, any philosophical methodology not closely attending to scientific results and deferential to its conceptual schemes is invalid. What Burge fails to appreciate is the way that the abstractions and idealizations of science—and, of course, he is aware that models depend on them—invalidate the authority of science within such inquiries into "natures" and "kinds." These simplifications are introduced at least in part for pragmatic reasons, and as many examples in this book have shown, they involve departures, in the scientific representation, from how things actually are. A philosophical inquiry concerned, for example,

with "human nature" should not satisfy itself with the caricature given in a scientific model that must necessarily abstract away from the variety, complications, and subjectivity that make human existence what it is. The same simplifying assumption may offer an innocuous convenience in a scientific context, but cause an endless, exhausting headache once embedded in philosophical inquiry—as is the case with the idealization of the self contained mind.

## 10.6 Philosophy without Science, Science without Simplification?

> Our mind has an irresistible tendency to consider that idea clearest which is most often useful to it.
>
> —Henri Bergson (1903/1912, 53)

Haugeland writes as if the Cartesian idealization of the separable mind is dispensable in the science of brain and nervous system. He highlights the assumption that the interface between mind and the rest of creation is a narrowband one, and he takes it to be a straightforward hypothesis that stands open to empirical refutation. He argues that observations in neuroanatomy of the density of connections between brain and body confirm, rather, that the interface must be broadband; he asserts that the intensity of interaction between brain and body, body and environment, tell against the idea that these are component systems, in Simon's sense. Thus science, by itself, can lead us beyond Cartesianism.

The worry left neglected is that even if the intermingling of systems and processes is an observable, empirical fact, it would not be feasible for science to accept all these high-bandwidth interfaces at face value and attempt to represent them in its theories and models. Science needs to limit its objects of investigation to bite-sized proportions. Of course, there is a long-standing and still active research tradition in embodied and embedded cognitive science, which is to say that some scientists, sometimes willed on by philosophers, have tried to dispense with the Cartesian idealization. But it is telling that this is a minority approach and it has sometimes struggled to achieve recognition.[27] The persistence with which it has been pur-

27. The ecological psychology of J. J. Gibson is probably the most successful example of work in the 4E (embodied, embedded, extended, and enactive) tradition, in

sued attests to its grounding in empirical fact, the various phenomena that tell against the separability of brain, mind, body, and environment. At the same time, this unwillingness to overlay those observations with idealizing distortions results in a loss of tractability and precise theoretical articulation, in comparison with the mainstream computationalist approach.[28]

The picture we are left with is one in which science, bound to make certain simplifications, and philosophy, more free to dispense with them, take divergent paths of inquiry. But a concern should arise here about the claim that philosophy is any less reliant than science on simplifications in the conceptual domain. We saw in the quotation from Cassirer in section 10.4 that it is possible to accuse metaphysical theorizing, which has its own standards of precision and determinacy, of the division of mind and body into separate, opposed tendencies. To the extent that philosophy, as much as science, needs to delimit topics and clarify its terms, then it too will be saddled with conceptual schemas that do not do justice to the complexity and interwovenness of life, experience, and embodiment. Indeed, so much of current analytic

terms of wider impact in cognitive science, but Burge (2005, 70–71, n21) is dismissive about it. For more recent statements and defenses, see Chemero (2011), Di Paolo, Buhrmann, and Barandiaran (2017), and Varela, Thompson, and Rosch (1991/2016). I should also mention that there is scientific work under the 4E umbrella that is just as reliant on mathematical abstractions as standard computationalism—in particular, work using dynamical systems theory. This raises (broadly) the same worries about abstraction away from biological complexity that come up with computationalist models.

28. I should admit here that the conclusion to this chapter is a criticism of my former self. In my previous book (Chirimuuta 2015), I endorsed an ecological theory of color perception, but like Haugeland, I assumed that it would be a straightforward matter to defend it on both philosophical and scientific grounds. I now think this was a mistake, and that arguing for the thesis that colors are properties not locatable on either side of a firm perceiver-environment boundary must involve a deeper scrutinization of the assumptions and idealizations of mainstream perceptual science. In advocating for the independence of some philosophical forms of inquiry into perception, and implicitly criticizing naturalistic research in philosophy of mind for failing to take seriously the way that scientific abstractions and idealizations distort the issues at stake, I have arrived at a methodological position closer to McDowell's than Haugeland's. But this has come about through a much closer examination of the science than to be found in the nonnaturalistic literature in philosophy of mind. The autonomy of philosophy cannot be properly defended by treating it as a walled garden and keeping the context of scientific activity—which has a dominating effect on our intellectual culture—out of view.

philosophy, with its reliance on cooked-up scenarios and toy models, reads like a transference into a nonempirical domain of the habit of thought that leads scientists to prefer controlled artificial conditions and unbelievable idealizations over the indeterminacies of uncontrived situations, where quantification is unavailable and the cataloguing of variables is incomplete. As Whitehead (1925/1967, 59) said, "You cannot think without abstractions," and this holds for everyone, regardless of academic job description.

The point to appreciate here is that philosophy need not commit itself to the same abstractions as science. Since philosophy is not bound by the requirement to design conceptual tools that serve material purposes, like prediction and manipulation of physiological effects, it has more latitude in how it goes about its abstractions, and it is also in a position to evaluate scientific abstractions by standards different from the instrumental ones of technoscience. We see this in McDowell's rejection of the Cartesian self-containment of mind for being a source of philosophical anxieties concerning the disconnection between mind and world, troubles to which scientists using that abstraction would generally be oblivious. Attunement to the wider implications of abstractions is one of the more valuable activities of philosophy, its role being that of the "critic of abstractions" (Whitehead 1925/1967, 59).

Thus, we should not accept that philosophy must be saddled with the same abstractions as science, not even given the shared history of these disciplines and common intellectual culture. The thing to remember, from chapter 8, is that science (unlike philosophy) has been conjoined with engineering, and this deeply shapes the particular ways that it simplifies its subject matter. The quotation from Bergson at the beginning of this section notes a cozy association between the clarity and utility of a concept. There are very many ways that, with our concepts, we simplify and make complex things seem more clear. And there are divergent ends with which we put those concepts to use. Cartesian idealization cannot be disentangled from the Cartesian agenda, which is for science to make mortals "masters and possessors" of nature. More generally, the form of simplification that science imposes is the one most conducive to the making of its object into a "manipulandum" (Merleau-Ponty 1961/2001, 289). These are ends that, as philosophers, we should at least reflect upon, and at times condemn.

The clash between Burge and McDowell involved a disagreement over whether philosophy could claim to have a subject matter of its own. McDowell carves out for himself the notion of the state of a perceiver as opposed to

states of perceptual systems, whereas Burge finds all these terms incorporated into the field of investigation of perceptual science. And it is true that explanations in that science occur as much at the person-level as the subpersonal one. However, if we appreciate that the divergent agendas of the two disciplines make available or unavailable different kinds of abstractions, we see that their notions and subject matter can indeed be different, even if they seem to be referring to the same thing. The philosopher, but not the scientist, has available a concept of personhood that takes persons to be embedded in an expansive network of circumstances and does not abstract away from subjectivity, intersubjectivity, and the normativity that is an omnipresent aspect of this form of existence.

When theorizing minds and perception, with that notion of the personal in place, it is requisite that the self-contained Cartesian ideal be rejected. For their person-level accounts, however, perceptual scientists are better off abstracting away from all of those—to their ends—extraneous factors. All that matters for the achievement of their tasks is what can be clearly defined as either external and environmental or inner and proximal, so the kind of person-level explanation arising from perceptual science is compatible with Cartesianism but antithetical to a philosophy concerned with personhood as a complicated site of normativity and intersubjectivity. This is why the theory of mind embodied and embedded must be retained, even if it cannot come up with all the scientific credentials of other accounts.

# References


Abbott, Laurence F., Stefano Fusi, and Kenneth D. Miller. 2013. "Appendix F, Theoretical Approaches to Neuroscience: Examples from Single Neurons to Networks." In *Principles of Neural Science*, 5th ed., edited by Sarah Mack, Eric R. Kandel, Thomas M. Jessell, James H. Schwartz, Steven A. Siegelbaum and A. J. Hudspeth. New York: McGraw Hill Professional.

Abraham, Tara. 2004. "Nicolas Rashevsky's Mathematical Biophysics." *Journal of the History of Biology* 37: 333–385.

Abraham, Tara. 2016. *Rebel Genius: Warren S. McCulloch's Transdisciplinary Life in Science*. Cambridge, MA: MIT Press.

Adelson, E. H., and J. R. Bergen. 1985. "Spatiotemporal Energy Models for the Perception of Motion." *Journal of the Optical Society of America, Series A 2*: 284–299.

Adorno, T. W. 2005. *Critical Models: Interventions and Catchwords*. Translated by Henry W. Pickford. New York: Columbia University Press.

Adrian, E. D. 1954. "Address of the President Dr E. D. Adrian, O.M., at the Anniversary Meeting, 30 November 1953." *Proceedings of the Royal Society of London, B* 142 (906): 1–9.

Aizawa, Kenneth. 2018. "Multiple Realization and Multiple 'Ways' of Realization: A Progress Report." *Studies in History and Philosophy of Science* 68: 3–9.

Albright, T. D., and G. R. Stoner. 2002. "Contextual Influences on Visual Processing." *Annual Review of Neuroscience* 25: 339–379.

Alivisatos, A. Paul, Miyoung Chun, George M. Church, Ralph J. Greenspan, Michael L. Roukes, and Rafael Yuste. 2012. "The Brain Activity Map Project and the Challenge of Functional Connectomics." *Neuron* 74: 970–974.

Allais, Lucy. 2015. *Manifest Reality: Kant's Idealism and His Realism*. Oxford: Oxford University Press.

Anderson, Chris. 2008. "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete." *Wired*. https://www.wired.com/2008/06/pb-theory/. Accessed May 10, 2021.

Anderson, P. W. 1972. "More Is Different: Broken Symmetry and the Nature of the Hierarchical Structure of Science." *Science* 177 (4047): 393–396.

Anderson, R. Lanier. 1998. "Truth and Objectivity in Perspectivism." *Synthese* 115: 1–32.

Arbib, Michael A. 2016. "Afterword: Warren McCulloch's Search for the Logic of the Nervous System." In *Embodiments of Mind*, edited by Warren S. McCulloch. Cambridge, MA: MIT Press.

Arieli, Amos, Alexander Sterkin, Amiram Grinvald, and Ad Aertsen. 1996. "Dynamics of Ongoing Activity: Explanation of the Large Variability in Evoked Cortical Responses." *Science* 273 (5283): 1868–1871.

Arkhipov, Anton, Nathan W. Gouwens, Yazan N. Billeh, et al. 2018. "Visual Physiology of the Layer 4 Cortical Circuit in Silico." *PLoS Computational Biology* 14 (11).

Ashby, W. Ross. 1954. *Design for a Brain*. New York: John Wiley & Sons.

Attneave, Fred. 1954. "Some Informational Aspects of Visual Perception." *Psychological Review* 61 (3): 183–193.

Baker, N., H. Lu, G. Erlikhman, and P. J. Kellman. 2018. "Deep Convolutional Networks Do Not Classify Based on Global Object Shape." *PLoS Computational Biology* 14 (12): e1006613.

Ballard, Dana. 2015. *Brain Computation as Hierarchical Abstraction*. Cambridge, MA: MIT Press.

Banks, E. C. 2004. "The Philosophical Roots of Ernst Mach's Economy of Thought." *Synthese* 139 (1): 23–53.

Barack, David L. 2019. "Mental Machines." *Biology and Philosophy* 34 (63). https://doi.org/10.1007/s10539-019-9719-6.

Barack, David L. 2020. "Mental Kinematics: Dynamics and Mechanics of Neurocognitive Systems." *Synthese*. https://doi.org/10.1007/s11229-020-02766-1.

Barack, David L., and John W. Krakauer. 2021. "Two Views on the Cognitive Brain." *Nature Reviews Neuroscience*. https://doi-org.ezproxy.is.ed.ac.uk/10.1038/s41583-021-00448-6.

Barlow, H. B. 1953. "Summation and Inhibition in the Frog's Retina." *Journal of Physiology* 119: 69–88.

Barlow, H. B. 1961. "Possible Principles Underlying the Transformation of Sensory Messages." In *Sensory Communication*, edited by W. A. Rosenblith, 217–234. Cambridge, MA: MIT Press.

Barlow, H. B. 1990. "The Mechanical Mind." *Annual Review of Neuroscience* 13: 15–24.

Barlow, Horace. 1972. "Single Units and Sensation: A Neuron Doctrine for Perceptual Psychology?" *Perception* 1: 371–394.

Bartha, Paul. 2016. "Analogy and Analogical Reasoning." *Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/entries/reasoning-analogy/.

Bashivan, Pouya, Kohitij Kar, and James J. DiCarlo. 2019. "Neural Population Control via Deep Image Synthesis." *Science* 364 (eaav9436): 1–11.

Bateson, P., and P. Gluckman. 2011. *Plasticity, Robustness, Development and Evolution*. Cambridge: Cambridge University Press.

Batterman, Robert W. 2010. "On the Explanatory Role of Mathematics in Empirical Science." *British Journal for the Philosophy of Science* 61: 1–25.

Batterman, Robert W. 2018. "Autonomy of Theories: An Explanatory Problem." *Noûs* 52 (4): 858–873.

Beatty, John. 1994. "The Proximate/Ultimate Distinction in the Multiple Careers of Ernst Mayr." *Biology and Philosophy* 9: 333–356.

Bechtel, William. 2015. "Can Mechanistic Explanation Be Reconciled with Scale-Free Constitution and Dynamics?" *Studies in History and Philosophy of Science* C 53: 84–89.

Bechtel, William. 2016. "Investigating Neural Representations: The Tale of Place Cells." *Synthese* 193: 1287–1321.

Bechtel, William, and Robert C. Richardson. 2010. *Discovering Complexity*. 2nd ed. Cambridge, MA: MIT Press.

Beer, R. D., and Paul L. Williams. 2015. "Information Processing and Dynamics in Minimally Cognitive Agents." *Cognitive Science* 39: 1–38.

Behrens, Timothy E. J., Timothy H. Muller, James C. R. Whittington, et al. 2018. "What Is a Cognitive Map? Organizing Knowledge for Flexible Behavior." *Neuron* 100: 490–509.

Beiser, Frederick. 2010. *Diotima's Children*. Oxford: Oxford University Press.

Beiser, Frederick. 2014. *After Hegel*. Princeton, NJ: Princeton University Press.

Bergson, Henri. 2001. *Time and Free Will*. Mineola, NY: Dover Publications.

Originally published in 1889.

Bergson, Henri. 1912. *An Introduction to Metaphysics*. Translated by T. E. Hulme. New York: G. P. Putnam's Sons.

Originally published in 1903.

Bergson, Henri. 1944. *Creative Evolution*. New York: Random House.

Originally published in 1907.

Berk, Michael, and Andrew Nierenberg. 2015. "Three Paths to Drug Discovery in Psychiatry." *American Journal of Psychiatry* 172 (5): 412–414.

Berkowitz, Carin. 2015. *Charles Bell and the Anatomy of Reform*. Chicago: University of Chicago Press.

Bermudez-Contreras, Edgar, Benjamin J. Clark, and Aaron Wilber. 2020. "The Neuroscience of Spatial Navigation and the Relationship to Artificial Intelligence." *Frontiers in Computational Neuroscience* 14: 63. https://doi.org/10.3389/fncom.2020.00063.

Bickle, John. 2008. "Real Reduction in Real Neuroscience: Metascience, not Philosophy of Science (and Certainly not Metaphysics!)." In *Being Reduced: New Essays on Reduction, Explanation, and Causation*, edited by Jakob Hohwy and Jesper Kallestrup, 34–51. Oxford: Oxford University Press.

Bickle, John. 2022. "Tinkering in the Lab." In *The Tools of Neuroscience Experiment*, edited by John Bickle, Carl F. Craver, and Ann-Sophie Barwich, 13–36. New York: Routledge.

Blakemore, C, and E. A. Tobin. 1972. "Lateral Inhibition between Orientation Detectors in the Cat's Visual Cortex." *Experimental Brain Research* 15: 439–440.

Block, Ned. 1980. "What Is Functionalism?" In *Readings in Philosophy of Psychology*, vol. 1, edited by Ned Block, 171–184. Cambridge, MA: Harvard University Press.

Bogen, J., and J. F. Woodward. 1988. "Saving the Phenomena." *Philosophical Review* 97 (3): 303–352.

Bokulich, Alisa. 2012. "Distinguishing Explanatory from Nonexplanatory Fictions." *Philosophy of Science* 79 (5): 725–737.

Bonds, A. B. 1989. "Role of Inhibition in the Specification of Orientation Selectivity of Cells in the Cat Striate Cortex." *Visual Neuroscience* 2: 41–55.

Bongard, Joshua, and Michael Levin. 2021. "Living Things Are Not (20th Century) Machines: Updating Mechanism Metaphors in Light of the Modern Science of Machine Behavior." *Frontiers in Ecology and Evolution* 9: 650726.

Boon, Mieke. 2020. "How Scientists Are Brought Back into Science—The Error of Empiricism." In *A Critical Reflection on Automated Science: Will Science Remain Human?* edited by M. Bertolaso and F. Sterpetti, 43–65. Cham, Switzerland: Springer Nature Switzerland.

Borges, Jorge Luis. 1998. *Collected Fictions*. Translated by Andrew Hurley. London: Allen Lane (Penguin).

Boring, Edwin G. 1950. *A History of Experimental Psychology*. 2nd ed. Englewood Cliffs, NJ: Prentice-Hall.

Boyle, Matthew. unpublished. *Kant's Hylomorphism*.

Boyle, Robert. 1686/1996. *A Free Enquiry into the Vulgarly Received Notion of Nature*. Edited by Edward B. Davis, Michael Hunter. Cambridge: Cambridge University Press.

Brady, Timothy F., Talia Konkle, George A. Alvarez, and Aude Oliva. 2008. "Visual Long-Term Memory Has a Massive Storage Capacity for Object Details." *PNAS* 105 (38): 14325–14329.

Breiman, Leo. 2001. "Statistical Modeling: The Two Cultures." *Statistical Science* 16: 199–215.

Breland, Keller, and Marian Breland. 1961. "The Misbehavior of Organisms." *American Psychologist* 16: 681–684.

Brette, Romain. 2019. "Is Coding a Relevant Metaphor for the Brain?" *Behavioral and Brain Sciences* 42: 1–58.

Bridewell, Will, and Alistair Isaac. 2021. "Apophatic Science: How Computational Modeling Can Explain Consciousness." *Neuroscience of Consciousness* 7 (1): niab010.

Bridgman, P. W. 1927. *The Logic of Modern Physics*. New York: MacMillan.

Bubeck, Sébastien, Varun Chandrasekaran, Ronen Eldan, et al. 2023. "Sparks of Artificial General Intelligence: Early Experiments with GPT-4." https://arxiv.org/abs/2303.12712.

Buckley, Kerry W. 1989. *Mechanical Man: John B. Watson and the Beginnings of Behaviorism*. New York: Guilford Press.

Buckner, Cameron. 2018. "Empiricism without Magic: Transformational Abstraction in Deep Convolutional Neural Networks." *Synthese* 195: 5339–5372.

Buckner, Cameron. 2019. "Deep Learning: Philosophical Issues." *Philosophy Compass* 14 (10): e12625. https://doi.org/10.1111/phc3.12625.

Buckner, Cameron. 2020. "Understanding Adversarial Examples Requires a Theory of Artefacts for Deep Learning." *Nature Machine Intelligence* 2: 731–736.

Buckner, Cameron. 2023. *From Deep Learning to Rational Machines*. Oxford: Oxford University Press.

Bullock, Theodore H., Michael V. L. Bennett, Daniel Johnston, Robert Josephson, Eve Marder, and R. Douglas Field. 2005. "The Neuron Doctrine, Redux." *Science* 310: 791–793.

Burge, Tyler. 2005. "Disjunctivism and Perceptual Psychology." *Philosophical Topics* 33 (1): 1–78.

Burge, Tyler. 2010. *Origins of Objectivity*. Oxford: Oxford University Press.

Burge, Tyler. 2011. "Disjunctivism Again." *Philosophical Explorations* 14 (1): 43–80.

Burgess, Matt. 2017. "DeepMind's Latest AI Breakthrough Is Its Most Significant Yet." *Wired*, October 10. https://www.wired.co.uk/article/deepmind-alphago-zero-nature-reinforcement-learning.

Burnyeat, M. F. 1992. "Is an Aristotelian Philosophy of Mind Still Credible (A Draft)?" In *Essays on Aristotle's De Anima*, edited by Martha C. Nussbaum and Amélie Oksenberg Rorty, 15–26. Oxford: Oxford University Press.

Butts, Daniel A. 2019. "Data-Driven Approaches to Understanding Visual Neuron Activity." *Annual Review of Neuroscience* 5: 451–477.

Buytendijk, F. J. J., and H. Plessner. 1936. "Die Physiologische Erklärung des Verhaltens: Eine Kritik an der Theorie Pawlows." *Acta Biotheoretica* 1: 151–172.

Buzzoni, Marco. 2019. "Multilevel Reality, Mechanistic Explanations, and Intertheoretic Reductions." In *Mechanistic Explanations in Physics and Beyond*, edited by Brigette Falkenburg and Gregor Schiemann, 111–141. Berlin: Springer.

Byrne, Alex, and Heather Logue, eds. 2009. *Disjunctivism: Contemporary Readings*. Cambridge, MA: MIT Press.

Cadena, Santiago A., George H. Denfield, Edgar Y. Walker, et al. 2019. "Deep Convolutional Models Improve Predictions of Macaque V1 Responses to Natural Images." *PLoS Computational Biology* 15 (4):e1006897. https://doi.org/10.1371/journal.pcbi.1006897.

Cadieu, Charles F., Ha Hong, Daniel L. K. Yamins, et al. 2014. "Deep Neural Networks Rival the Representation of Primate IT Cortex for Core Visual Object Recognition." *PLoS Computational Biology* 10 (12):e1003963.

Cajal, Ramón y. 1937. *Recollections of My Life*. Translated by E. Horne Craigie. Philadelphia: American Philosophical Society.

Callebaut, W. 2012. "Scientific Perspectivism: A Philosopher of Science's Response to the Challenge of Big Data Biology." *Studies in History and Philosophy of Biological and Biomedical Sciences* 43 (1): 69–80. doi: 10.1016/j.shpsc.2011.10.007.

Canguilhem, Georges. 1963. "The Role of Analogies and Models in Biological Discovery." In *Scientific Change*, edited by A. C. Crombie, 507–520. New York: Basic Books.

Canguilhem, Georges. 1982. "Descartes e a Técnica/Descartes et la Technique." *Trans/Form/Ação* 5: 111–122.

Originally published in 1937.

Canguilhem, Georges. 1994. "The Concept of Reflex." In *A Vital Rationalist: Selected Writings from Georges Canguilhem*, edited by François Delaporte. New York: Zone Books.

Canguilhem, Georges. 2008a. "Aspects of Vitalism." In *Knowledge of Life,* edited by Paola Marrati and Todd Meyers, 59–74. New York: Fordham University Press.

Originally published in 1965.

Canguilhem, Georges. 2008b. "The Living and Its Milieu." In *Knowledge of Life,* edited by Paola Marrati and Todd Meyers, 98–120. New York: Fordham University Press.

Originally published in 1965.

Canguilhem, Georges. 2008c. "Machine and Organism." In *Knowledge of Life,* edited by Paola Marrati and Todd Meyers, 75–97. New York: Fordham University Press.

Originally published in 1965.

Canguilhem, Georges. 2012. *Writings on Medicine*. Translated by Stefanos Geroulanos and Todd Meyers. New York: Fordham University Press.

Originally published in 1989.

Canguilhem, Georges. 2015. *La Formation du Concept de Réflexe aux XVIIe et XVIIIe Siècles*. Paris: J. Vrin.

Originally published in 1955.

Cantwell Smith, Brian. 2019. *The Promise of Artificial Intelligence: Reckoning and Judgment*. Cambridge, MA: MIT Press.

Cao, Rosa. 2014. "Signaling in the Brain: In Search of Functional Units." *Philosophy of Science* 81 (5): 891–901.

Cao, Rosa. 2019. "Computational Explanations and Neural Coding." In *Routledge Handbook of the Computational Mind*, edited by Mark Sprevak and Matteo Colombo, 283–296. Abingdon, UK: Routledge.

Cao, Rosa, and Daniel L. K. Yamins. 2021a. "Explanatory Models in Neuroscience: Part 1—Taking Mechanistic Abstraction Seriously." arXiv:2104.01490. https://arxiv.org/abs/2104.01490.

Cao, Rosa, and Daniel L. K. Yamins. 2021b. "Explanatory Models in Neuroscience: Part 2—Constraint-Based Intelligibility." arXiv:2104.01489. https://arxiv.org/abs/2104.01489.

Carandini, M. 2012. "From Circuits to Behavior: A Bridge Too Far?" *Nature Neuroscience* 15 (4): 507–509. doi: 10.1038/nn.3043.

Carandini, M., J. B. Demb, V. Mante, et al. 2005. "Do We Know What the Early Visual System Does?" *Journal of Neuroscience* 25 (46): 10577–10597. doi: 10.1523/JNEUROSCI.3726–05.2005.

Carandini, M., and D. J. Heeger. 2011. "Normalization as a Canonical Neural Computation." *Nature Reviews Neuroscience* 13 (1): 51–62. doi: 10.1038/nrn3136.

Carlyle, Thomas. 1839. *Critical and Miscellaneous Essays (Thomas Carlyle's Collected Works, Vol. VII)*. London: Chapman and Hall.

Carr, Danielle Judith Zola. 2020. "'Ghastly Marionettes' and the Political Metaphysics of Cognitive Liberalism: Anti-Behaviourism, Language, and the Origins of Totalitarianism." *History of the Human Sciences* 33 (1): 147–174.

Carrier, Martin. 2004. "Knowledge and Control: On the Bearing of Epistemic Values in Applied Science." In *Science, Values and Objectivity*, edited by Peter Machamer and Gereon Wolters, 275–293. Pittsburgh: University of Pittsburgh Press.

Cartwright, Nancy. 1983. *How the Laws of Physics Lie*. Oxford: Oxford University Press.

Cartwright, Nancy. 1999. *The Dappled World*. Cambridge: Cambridge University Press.

Cartwright, Nancy. 2009. "Causal Laws, Policy Predictions, and the Need for Genuine Powers." In *Dispositions and Causes*, edited by Toby Handfield, 127–157. Oxford: Oxford University Press.

Carus, A. W. 2007. *Carnap and Twentieth-Century Thought: Explication as Enlightenment*. Cambridge: Cambridge University Press.

Casper, Stephen T. 2014. "History and Neuroscience: An Integrative Legacy." *Isis* 105: 123–132.

Cassirer, Ernst. 1923. *Substance and Function, and Einstein's Theory of Relativity*. Translated by William Curtis Swabey and Marie Collins Swabey. Chicago: Open Court.

Originally published in 1910.

Cassirer, Ernst. 1957. *The Philosophy of Symbolic Forms, Volume 3: The Phenomenology of Knowledge*. New Haven, CT: Yale University Press.

Originally published in 1929.

Cassirer, Ernst. 1950. *The Problem of Knowledge: Philosophy, Science, and History since Hegel*. Translated by William H. Woglom. New Haven, CT: Yale University Press.

Cassirer, Ernst. 1981. *Kant's Life and Thought*. Translated by James Haden. New Haven, CT: Yale University Press.

Originally published in 1919.

Castelle, Michael. Under review. "Are Neural Networks Neoclassical? The Role of Economic Rationality in Artificial Intelligence." *British Journal for the History of Science*.

Cavanaugh, J., W. Bair, and J. A. Movshon. 2002. "Nature and Interaction of Signals from the Receptive Field Center and Surround in Macaque V1 Neurons." *Journal of Neurophysiology* 88: 2530–2546.

Cavendish, Margaret. 2001. *Observations upon Experimental Philosophy*. Cambridge: Cambridge University Press.

Originally published in 1668.

Cembrowski, Mark S., and Nelson Spruston. 2019. "Heterogeneity within Classical Cell Types Is the Rule: Lessons from Hippocampal Pyramidal Neurons." *Nature Reviews Neuroscience* 20: 193–204.

Chakravartty, Anjan. 2007. *A Metaphysics for Scientific Realism: Knowing the Unobservable*. Cambridge: Cambridge University Press.

Chakravartty, Anjan. 2010. "Perspectivism, Inconsistent Models, and Contrastive Explanation." *Studies in History and Philosophy of Science* 41: 405–412.

Chakravartty, Anjan. 2017. *Scientific Ontology*. Oxford: Oxford University Press.

Chalmers, David J. 1996. *The Conscious Mind*. Oxford: Oxford University Press.

Chalmers, David J. 2012. "A Computational Foundation for the Study of Cognition." *Journal of Cognitive Science* 12: 323–357.

Chalmers, David J. 2014. "Uploading: A Philosophical Analysis." In *Intelligence Unbound: The Future of Uploaded and Machine Minds*, edited by Russell Blackford and Damien Broderick, 102–118. Chichester, UK: John Wiley & Sons.

Chalmers, David J. 2020. "GPT-3 and General Intelligence." *Daily Nous*, July 30. https://dailynous.com/2020/07/30/philosophers-gpt-3/#chalmers.

Chambers, Anna, and Simon Rumpel. 2017. "A Stable Brain from Unstable Components: Emerging Concepts and Implications for Neural Computation." *Neuroscience* 357: 172–184.

Chang, Hasok. 2004. *Inventing Temperature: Measurement and Scientific Progress*. Oxford: Oxford University Press.

Chang, Hasok. 2012. *Is Water H2O? Evidence, Realism and Pluralism*. Dordrecht, Netherlands: Springer.

Chang, Hasok. 2020. "Pragmatism, Perspectivism, and the Historicity of Science." In *Understanding Perspectivism: Scientific Challenges and Methodological Prospects*, edited by Michela Massimi and Casey McCoy, 10–27. New York: Routledge.

Chang, Hasok. 2022. *Realism for Realistic People*. Cambridge: Cambridge University Press.

Chemero, Anthony. 2011. *Radical Embodied Cognitive Science*. Cambridge, MA: MIT Press.

Cheung, Tobias. 2006. "From the Organism of a Body to the Body of an Organism: Occurrence and Meaning of the Word 'Organism' from the Seventeenth to the Nineteenth Centuries." *British Journal for the History of Science* 39 (3): 319–339.

Chirimuuta, M. In preparation. "The Davos Debate over AI."

Chirimuuta, M. 2013. "Extending, Changing, and Explaining the Brain." *Biology and Philosophy* 28: 613–638.

Chirimuuta, M. 2014. "Minimal Models and Canonical Neural Computations: The Distinctness of Computational Explanation in Neuroscience." *Synthese* 191: 127–153.

Chirimuuta, M. 2015. *Outside Color: Perceptual Science and the Puzzle of Color in Philosophy*. Cambridge, MA: MIT Press.

Chirimuuta, M. 2016. "Vision, Perspectivism, and Haptic Realism." *Philosophy of Science* 83: 746–756.

Chirimuuta, M. 2017a. "Crash Testing an Engineering Framework in Neuroscience: Does the Idea of Robustness Break Down?" *Philosophy of Science* 84: 1140–1151.

Chirimuuta, M. 2017b. "Hughlings Jackson and the 'Doctrine of Concomitance': Mind-Brain Theorising between Metaphysics and the Clinic." *History and Philosophy of the Life Sciences* 39: 26.

Chirimuuta, M. 2018. "Explanation in Computational Neuroscience: Causal and Non-Causal." *British Journal for the Philosophy of Science* 69 (3): 849–880.

Chirimuuta, M. 2019. "Synthesis of Contraries: Hughlings Jackson on Sensory-Motor Representation in the Brain." *Studies in History and Philosophy of Biological and Biomedical Sciences* 75: 34–44.

Chirimuuta, M. 2020a. "Cassirer and Goldstein on Abstraction and the Autonomy of Biology." *HOPOS: Journal of the International Society for the History of Philosophy of Science* 10:471–503. https://doi.org/10.1086/710181.

Chirimuuta, M. 2020b. "Charting the Heraclitean Brain: Perspectivism and Simplification in Models of the Motor Cortex." In *Understanding Perspectivism: Scientific Challenges and Methodological Prospects*, edited by Michela Massimi and Casey McCoy, 141–159. New York: Routledge.

Chirimuuta, M. 2020c. "Prediction versus Understanding in Computationally Enhanced Neuroscience." *Synthese* 199: 767–790. doi: 10.1007/s11229-020-02713-0.

Chirimuuta, M. 2020d. "The Reflex Machine and the Cybernetic Brain: The Critique of Abstraction and Its Application to Computationalism." *Perspectives on Science* 28 (3): 421–457.

Chirimuuta, M. 2021. "Reflex Theory, Cautionary Tale: Misleading Simplicity in Early Neuroscience." *Synthese* 199: 12731–12751. https://doi.org/10.1007/s11229-021-03351-w.

Chirimuuta, M. 2022a. "Artefacts and Levels of Abstraction." *Frontiers in Ecology and Evolution*. doi: 10.3389/fevo.2022.952992.

Chirimuuta, M. 2022b. "Comment on Neurocognitive Mechanisms by Gualtiero Piccinini." *Journal of Consciousness Studies* 29 (7–8): 185–194.

Chirimuuta, M. 2022c. "Disjunctivism and Cartesian Idealisation." *Proceedings of the Aristotelian Society* XX: 1–21. https://doi.org/10.1093/arisoc/aoac010.

Chirimuuta, M. 2023a. "Ideal Patterns and Non-Factive Understanding." In *Scientific Understanding and Representation: Modeling in the Physical Sciences*, edited by Kareem Khalifa, Insa Lawler and Elay Shech, 78–95. London: Routledge.

Chirimuuta, M. 2023b. "Rules, Judgment, and Mechanisation." *Journal of Cross-Disciplinary Research in Computational Law*. 1 (3). https://journalcrcl.org/crcl/article/view/22.

Chirimuuta, M. 2023c. "Haptic Realism for Neuroscience." *Synthese*, 202:63 https://doi.org/10.1007/s11229-023-04295-z.

Chirimuuta, M., and I. Gold. 2009. "The Embedded Neuron, the Enactive Field?" In *Handbook of Philosophy and Neuroscience*, edited by John Bickle, 200–225. Oxford: Oxford University Press.

Churchland, Patricia Smith. 1994. "Can Neurobiology Teach Us Anything about Consciousness?" *Proceedings and Addresses of the American Philosophical Association*, 67 (4): 23–40.

Churchland, Patricia Smith, Christof Koch, and Terrence J. Sejnowski. 1994. "What Is Computational Neuroscience?" In *Computational Neuroscience*, edited by Eric L. Schwartz, 46–55. Cambridge: Cambridge University Press.

Churchland, Patricia Smith, and Terrence J. Sejnowski. 2016. "Blending Computational and Experimental Neuroscience." *Nature Reviews Neuroscience* 17: 667–668.

Clark, Andy. 2016. *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*. Oxford: Oxford University Press.

Clark, Andy, and Josefa Toribio. 1994. "Doing without Representing?" *Synthese* 101 (3): 401–431.

Clarke, Edwin, and L. S. Jacyna. 1987. *Nineteenth-Century Origins of Neuroscientific Concepts*. Berkeley: University of California Press.

Coelho Mollo, Dimitri. 2021. "Deflationary Realism: Representation and Idealisation in Cognitive Science." *Mind & Language*. doi: 10.1111/mila.12364.

Coelho Mollo, Dimitri and Raphaël Millière. 2023. "The Vector Grounding Problem." https://arxiv.org/abs/2304.01481.

Colaço, David. 2020. "Recharacterizing Scientific Phenomena." *European Journal for Philosophy of Science* 10 (2): 1–19.

Collins, Harry. 1996. "Embedded or Embodied? A Review of Hubert Dreyfus' What Computers Still Can't Do." *Artificial Intelligence* 80: 99–117.

Colombo, Matteo. 2014. "Explaining Social Norm Compliance: A Plea for Neural Representations." *Phenomenology and the Cognitive Sciences* 13: 217–238.

Copeland, B. Jack. 2020. "The Church-Turing Thesis." In *Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/archives/spr2019/entries/church-turing/.

Crane, Tim, and Craig French. 2021. "The Problem of Perception." In *Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/archives/spr2021/entries/perception-problem/.

Cranz, F. Edward. 1953. "Saint Augustine and Nicholas of Cusa in the Tradition of Western Christian Thought." *Speculum* 28(2): 297-316.

Craver, Carl F. 2007. *Explaining the Brain.* Oxford: Oxford University Press.

Craver, Carl F. 2013. Functions and Mechanisms: A Perspectivalist View. In *Functions: Selection and Mechanisms*, edited by Philippe Huneman, 133–158. Dordrecht, Netherlands: Springer.

Craver, Carl F. 2014. "The Ontic Account of Scientific Explanation." In *Explanation in the Special Sciences: The Case of Biology and History*, edited by M. I. Kaiser, O. R. Scholz, D. Plenge, and A. Hüttemann, 27–52. Dordrecht, Netherlands: Springer.

Craver, Carl F., and Lindley Darden. 2013. *In Search of Mechanisms.* Chicago: University of Chicago Press.

Craver, Carl F., and David Michael Kaplan. 2018. "Are More Details Better? On the Norms of Completeness for Mechanistic Explanations." *British Journal for the Philosophy of Science*. doi: 10.1093/bjps/axy015.

Craver, Carl F., and James Tabery. 2017. "Mechanisms in Science." In *Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/entries/science-mechanisms/

Cunningham, J. P., and B. M. Yu. 2014. "Dimensionality Reduction for Large-Scale Neural Recordings." *Nature Neuroscience* 17 (11): 1500–1509. doi: 10.1038/nn.3776.

Cusanus, Nicolas. 1954. *Of Learned Ignorance.* Translated by Germain Heron. London: Routledge & Kegan Paul.

Dahms, H. J. 1994. *Positivismusstreit: Die Auseinandersetzungen der Frankfurter Schuke mit dem logischen Positivismus, dem amerikanischen Pragmatismus und dem kritischen Rationalismus*. Frankfurt am Main: Suhrkamp.

Danks, David. 2020. "Safe-and-Substantive Perspectivism." In *Understanding Perspectivism: Scientific Challenges and Methodological Prospects*, edited by Michela Massimi and Casey McCoy, 127–140. New York: Routledge.

Daston, Lorraine. 1994. "Enlightenment Calculations." *Critical Inquiry* 21 (1): 182–202.

Daston, Lorraine. 2016. "Cloud Physiognomy." *Representations* 135 (1): 45–71.

Daston, Lorraine. 2018. "Calculation and the Division of Labor, 1750–1950." *Bulletin of the German Historical Institute* 62: 9–30.

Daugman, John G. 2001. "Brain Metaphor and Brain Theory." In *Philosophy and the Neurosciences: A Reader*, edited by William Bechtel, Pete Mandik, Jennifer Mundale and Robert S. Stufflebeam, 23–36. Oxford, UK: Blackwell.

David, S. V., W. E. Vinje, and J. L. Gallant. 2004. "Natural Stimulus Statistics Alter the Receptive Field Structure of V1 Neurons." *Journal of Neuroscience* 24: 6991–7006.

Davis, Ernest, and Gary Marcus. 2015. "Commonsense Reasoning and Commonsense Knowledge in Artificial Intelligence." *Communications of the ACM* 58 (9): 92–103.

de Regt, Henk. 2009. "Understanding and Scientific Explanation." In *Scientific Understanding: Philosophical Perspectives*, edited by Henk de Regt, Sabine Leonelli and Kai Eigner. Pittsburgh: University of Pittsburgh Press.

de Regt, Henk. 2017. *Understanding Scientific Understanding*. Oxford: Oxford University Press.

De Valois, Russell L., and Karen K. De Valois. 1988. *Spatial Vision*. Oxford: Oxford University Press.

Dear, Peter. 2005. "What Is the History of Science the History Of?" *Isis* 96: 390–406.

Dear, Peter. 2006. *The Intelligibility of Nature*. Chicago: University of Chicago Press.

deCharms, R. Christopher, and Anthhony Zador. 2000. "Neural Representation and the Cortical Code." *Annual Review of Neuroscience* 23: 613–647.

Dehghani, Nima. 2018. "Theoretical Principles of Multiscale Spatiotemporal Control of Neuronal Networks: A Complex Systems Perspective." *Frontiers in Computational Neuroscience* 12: Article 81.

Deitch, Daniel, Alon Rubin, and Yaniv Ziv. 2021. "Representational Drift in the Mouse Visual Cortex." *Current Biology* 31: 4327–4339.

Dennett, Daniel, C. 1981. *Brainstorms: Philosophical Essays on Mind and Psychology*. Cambridge, MA: MIT Press.

Dennett, Daniel, C. 1987. *The Intentional Stance*. Cambridge, MA: MIT Press.

Dennett, Daniel, C. 1988. "Précis of the Intentional Stance." *Behavioral and Brain Sciences* 11: 495–546.

Dennett, Daniel, C. 1991. "Real Patterns." *Journal of Philosophy* 88 (1): 27–51.

Dennett, Daniel, C. 1995. "Cognitive Science as Reverse Engineering: Several Meanings of 'Top-Down' and 'Bottom-Up.'" *Logic, Methodology and Philosophy of Science IX (Proceedings of the Ninth International Congress of Logic, Methodology and Philosophy of Science)*, Uppsala, Sweden.

Dennett, Daniel, C. 1997. "True Believers: The Intentional Strategy and Why It Works." In *Mind Design II*, edited by John Haugeland. Cambridge, MA: MIT Press.

Originally published in 1981.

Descartes, René. 1985. *Philosophical Writings, Volume I*. Translated by John Cottingham, Robert Stoothoff, and Dugald Murdoch. Cambridge: Cambridge University Press.

de Vries, Saskia E. J., Jerome A. Lecoq, Michael A. Buice, et al. 2020. "A Large-Scale Standardized Physiological Survey Reveals Functional Organization of the Mouse Visual Cortex." *Nature Neuroscience* 23: 138–151.

Dewey, John. 1896. "The Reflex Arc Concept in Psychology." *Psychological Review* 3 (4): 357–370.

Dewey, John. 1929. *The Quest for Certainty: A Study of the Relation of Knowledge and Action*. New York: Minton, Balch & Company.

Dieks, Dennis. 2019. "Mechanisms, Explanation and Understanding in Physics." In *Mechanistic Explanations in Physics and Beyond*, edited by Brigette Falkenburg and Gregor Schiemann, 47–64. Berlin: Springer.

Dilthey, Wilhelm. 2010. "Ideas for a Descriptive and Analytic Psychology." In *Understanding the Human World—Wilhelm Dilthey, Selected Works, Vol. 2*, edited by Rudolf A. Makkreel and Frithjof Rodi. Princeton, NJ: Princeton University Press.

Originally published in 1894.

Di Paolo, Ezequiel, Thomas Buhrmann, and Xabier Barandiaran. 2017. *Sensorimotor Life: An Enactive Proposal*. Oxford: Oxford University Press.

Drake, Stillman. 1957. *Discoveries and Opinions of Galileo*. New York: Anchor Books.

Dreher, B., and K. J. Sanderson. 1973. "Receptive Field Analysis: Responses to Moving Visual Contours by Single Lateral Geniculate Neurones in the Cat." *Journal of Physiology* 234: 95–118.

Driscoll, L. N., N. L. Pettit, M. Minderer, S. N. Chettih, and C. D. Harvey. 2017. "Dynamic Reorganization of Neuronal Activity Patterns in Parietal Cortex." *Cell* 170: 986–999.e16.

du Bois-Reymond, Emil. 1874. "The Limits of Our Knowledge of Nature." *Popular Science Monthly* 5: 17–32.

Originally published in 1872.

Duhem, Pierre. 1954. *The Aim and Structure of Physical Theory*. Translated by P. Wiener. Princeton, NJ: Princeton University Press.

Originally published in 1906.

Dupré, John. 2012. *Processes of Life*. Oxford: Oxford University Press.

Dupré, John, and Daniel J. Nicholson. 2018. "A Manifesto for a Processual Philosophy of Biology." In *Everything Flows*, edited by Daniel J. Nicholson and John Dupré, 3–45. Oxford: Oxford University Press.

Dupuy, Jean-Pierre. 2009. *On the Origins of Cognitive Science*. Translated by M. B. DeBevoise. Cambridge, MA: MIT Press.

Edelman, Gerald M., and Joseph A. Gally. 2001. "Degeneracy and Complexity in Biological Systems." *PNAS* 98 (24): 13763–13768.

Edwards, José. 2016. "Behaviorism and Control in the History of Economics and Psychology." *History of Political Economy* 48: 170–197.

Edwards, Paul N. 1996. *The Closed World: Computers and the Politics of Discourse in Cold War America*. Cambridge, MA: MIT Press.

Egan, Frances. 2017. "Function-Theoretic Explanation and the Search for Neural Mechanisms." In *Explanation and Integration in Mind and Brain Science*, edited by David Michael Kaplan, 145–163. Oxford: Oxford University Press.

Egan, Frances. 2019. "The Nature and Function of Content in Computational Models." In Routledge *Handbook of the Computational Mind*, edited by Mark Sprevak and Matteo Colombo, 247–258. Abingdon, UK: Routledge.

Egan, Frances. 2020. "A Deflationary Account of Mental Representation." In *Mental Representations*, edited by Joulia Smortchkova, Krzysztof Dolega and Tobias Schlicht, 26–53. Oxford: Oxford University Press.

Einevoll, Gaute T., Alain Destexhe, Markus Diesmann, et al. 2019. "The Scientific Case for Brain Simulations." *Neuron* 102: 735–744.

Einstein, Albert. 1934. "On the Method of Theoretical Physics." *Philosophy of Science* 1 (2): 163–169.

Elgin, Catherine Z. 2017. *True Enough*. Cambridge MA: MIT Press.

Elgin, Catherine Z. 2019. "Nominalism, Realism and Objectivity." *Synthese* 196: 519–534.

Eliasmith, Chris. 2003. "Moving beyond Metaphors: Understanding the Mind for What It Is." *Journal of Philosophy* 100 (10): 493–520.

Evarts, E. V. 1968. "Relation of Pyramidal Tract Activity to Force Exerted during Voluntary Movement." *Journal of Neurophysiology* 31 (1): 14–27.

Fairhall, Adrienne. 2014. "The Receptive Field Is Dead. Long Live the Receptive Field?" *Current Opinion in Neurobiology* 25: ix–xii.

Falkenburg, Brigette. 2007. *Particle Metaphysics: A Critical Account of Subatomic Reality*. Berlin: Springer.

Falkenburg, Brigette. 2012. *Mythos Determinismus: Wieviel erklärt uns die Hirnforschung?* Heidelberg, Germany: Springer.

Falkenburg, Brigette. 2019. "Mechanistic Explanations Generalized: How Far Can We Go?" In *Mechanistic Explanations in Physics and Beyond*, edited by Brigette Falkenburg and Gregor Schiemann, 65–90. Berlin: Springer.

Favela, Luis H. 2021. "The Dynamical Renaissance in Neuroscience." *Synthese* 199: 2103–2127.

Fearing, Franklin. 1930. *Reflex Action: A Study in the History of Physiological Psychology*. New York: Hafner Publishing Company.

Feest, Uljana. 2011. "What Exactly Is Stabilized When Phenomena Are Stabilized?" *Synthese* 182: 57–71.

Ferrier, David. 1873. "Experimental Researches in Cerebral Physiology and Pathology." *Journal of Anatomy and Physiology* 8: 152–155.

Figdor, Carrie. 2018. *Pieces of Mind: The Proper Domain of Psychological Predicates*. Oxford: Oxford University Press.

Fisch, Max H. 1969. "Vico and Pragmatism." In *Giambattista Vico: An International Symposium*, edited by Giorgio Tagliacozzo and Hayden V. White, 401–424. Baltimore: Johns Hopkins Press.

Fiser, A., David Mahringer, Hassana K. Oyibo, Anders V. Petersen, Marcus Leinweber, and Georg B. Keller. 2016. "Experience-Dependent Spatial Expectations in Mouse Visual Cortex." *Nature Neuroscience* 19: 1658–1664.

Fish, William. 2021. "Perceptual Paradigms." In *Purpose and Procedure in Philosophy of Perception*, edited by Heather Logue and Louise Richardson, 23–42. Oxford: Oxford University Press.

Floridi, Luciano, and Massimo Chiriatti. 2020. "GPT-3: Its Nature, Scope, Limits, and Consequences." *Minds and Machines* 30: 681–694.

Forster, A. C., and G. M. Church. 2007. "Synthetic Biology Projects in Vitro." *Genome Research* 17: 1–6.

Freeman, Jeremy, Corey M. Ziemba, David J. Heeger, Eero P. Simoncelli, and J. Anthony Movshon. 2013. "A Functional and Perceptual Signature of the Second Visual Area in Primates." *Nature Neuroscience* 16 (7): 974–981.

Frégnac, Yves. 2017. "Big Data and the Industrialization of Neuroscience: A Safe Roadmap for Understanding the Brain?" *Science* 358: 470–477.

French, Steven. 2014. *The Structure of the World*. Oxford: Oxford University Press.

Freudenthal, Gideon, and Peter McLaughlin. 2009. "Classical Marxist Historiography of Science: The Hessen-Grossmann-Thesis." In *The Social and Economic Roots of the Scientific Revolution: Texts by Boris Hessen and Henryk Grossmann*, edited by Gideon Freudenthal and Peter McLaughlin, 1–40. Berlin: Springer.

Friedman, M. 2000. *A Parting of the Ways*. Chicago: Open Court.

Fritsch, G., and E. Hitzig. 1870. "Über die elektrische Erregbarkeit des Grosshirns." *Arch Anat Physiol Wissen* 37: 300–332.

Fukushima, K. 1980. "Neocognitron: A Self-Organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position." *Biological Cybernetics* 36 (4): 193–202.

Furness, J. B. 2006. *The Enteric Nervous System*. Oxford, UK: Blackwell Publishing.

Fusi, S., E. K. Miller, and M. Rigotti. 2016. "Why Neurons Mix: High Dimensionality for Higher Cognition." *Current Opinion in Neurobiology* 37: 66–74.

Gadamer, Hans-Georg. 2004. *Truth and Method*. Translated by Joel Weinsheimer and Donald G. Marshall. 2nd ed. London: Continuum.

Originally published in 1975.

Galison, P. 1994. "The Ontology of the Enemy: Norbert Wiener and the Cybernetic Vision." *Critical Inquiry* 21 (1): 228–266.

Gallego, Juan A., Matthew G. Perich, Lee E. Miller, and Sara A. Solla. 2017. "Neural Manifolds for the Control of Movement." *Neuron* 94: 978–984.

Gallego, Juan A., Matthew G. Perich, Stephanie N. Naufel, Christian Ethier, Sara A. Solla, and Lee E. Miller. 2018. "Cortical Population Activity within a Preserved Neural Manifold Underlies Multiple Motor Behaviors." *Nature Communications* 9 (4233). doi: 10.1038/s41467-018-06560-z.

Gao, Peiran, and Surya Ganguli. 2015. "On Simplicity and Complexity in the Brave New World of Large-Scale Neuroscience." *Current Opinion in Neurobiology* 32: 148–155.

Gao, Peiran, Eric Trautmann, Byron Yu, et al. 2017. "A Theory of Multineuronal Dimensionality, Dynamics and Measurement." bioRxiv. https://doi.org/10.1101/214262.

Geirhos, Robert, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. 2019. "ImageNet-Trained CNNs Are Biased towards

Texture; Increasing Shape Bias Improves Accuracy and Robustness." https://arxiv.org/abs/1811.12231.

Georgopoulos, A., J. F. Kalaska, R. Caminiti, and J. T. Massey. 1982. "On the Relations between the Direction of Two-Dimensional Arm Movements and Cell Discharge in Primate Motor Cortex." *Journal of Neuroscience* 2:1527–1537.

Georgopoulos, A., A. Schwartz, and R. Kettner. 1986. "Neuronal Population Coding of Movement Direction." *Science* 233 (4771): 1416–1419. doi: 10.1126/science.3749885.

Gerovitch, Slava. 2002. *From Newspeak to Cyberspeak: A History of Soviet Cybernetics*. Cambridge, MA: MIT Press.

Gershman, Samuel J., Petra E. M. Balbi, C. Randy Gallistel, and Jeremy Gunawardena. 2021. "Reconsidering the Evidence for Learning in Single Cells." *eLife* 10. doi: https://doi.org/10.7554/eLife.61907.

Gidon, Albert, Timothy Adam Zolnik, Pawel Fidzinski, et al. 2020. "Dendritic Action Potentials and Computation in Human Layer 2/3 Cortical Neurons." *Science* 367: 83–87.

Giere, Ron. 1999. *Science without Laws*. Chicago: University of Chicago Press.

Giere, Ron. 2006a. "Perspectival Pluralism." In *Minnesota Studies in the Philosophy of Science, Vol. 19: Scientific Pluralism*, edited by S. H. Kellert, H. E. Longino, and C. K. Waters. Minneapolis: University of Minnesota Press.

Giere, Ron. 2006b. *Scientific Perspectivism*. Chicago: University of Chicago Press.

Gilbert, C. D., M. Sigman, and R. E. Crist. 2001. "The Neural Basis of Perceptual Learning." *Neuron* 31: 681–697.

Gill, Michael B. 2021. "Shaftesbury on the Beauty of Nature." *Journal of Modern Philosophy* 3 (1): 1–28.

Gillett, Carl. 2016. *Reduction and Emergence in Science and Philosophy*. Cambridge: Cambridge University Press.

Gilmer, Justin, and Dan Hendrycks. 2019. Adversarial Example Researchers Need to Expand What Is Meant by 'Robustness.'" *Distill* https://distill.pub/2019/advex-bugs-discussion/response-1/. doi: 10.23915/distill.00019.1.

Ginsburg, Simona, and Eva Jablonka. 2019. *The Evolution of the Sensitive Soul*. Cambridge, MA: MIT Press.

Glennan, Stuart. 2002. "Rethinking Mechanistic Explanation." *Philosophy of Science* 69: S342–S353.

Godfrey-Smith, Peter. 1996. *Complexity and the Function of Mind in Nature*. Cambridge: Cambridge University Press.

Godfrey-Smith, Peter. 2004. "Mental Representation, Naturalism, and Teleosemantics." In *Teleosemantics: New Philosophical Essays*, edited by David Papineau and Graham MacDonald, 42–68. Oxford: Oxford University Press.

Godfrey-Smith, Peter. 2009. "Triviality Arguments against Functionalism." *Philosophical Studies* 145: 273–295.

Godfrey-Smith, Peter. 2016a. "Mind, Matter, and Metabolism." *Journal of Philosophy* 113 (10): 481–506.

Godfrey-Smith, Peter. 2016b. *Other Minds: The Octopus, the Sea, and the Deep Origins of Consciousness*. New York: Farrar, Straus and Giroux.

Goldenfeld, N., and L. P. Kadanoff. 1999. "Simple Lessons from Complexity." *Science* 284: 87–89.

Goldstein, Kurt. 1939. *The Organism: A Holistic Approach to Biology Derived from Pathological Data in Man*. New York: American Book Company.

Originally published in 1934.

Goldstein, Kurt. 1940. *Human Nature in the Light of Psychopathology*. Cambridge, MA: Harvard University Press.

Goldstein, Kurt. 1959. "Health as Value." In *New Knowledge in Human Values*, edited by A. Maslow, 178–188. New York: Harper.

Gould, Stephen Jay. 1981. *The Mismeasure of Man.* London: Penguin.

Graham Brown, Thomas. 1914. "On the Nature of the Fundamental Activity of the Nervous Centres; Together With an Analysis of the Conditioning of Rhythmic Activity in Progression, and a Theory of the Evolution of Function in the Nervous System." *Journal of Physiology* 48: 18–46.

Grant. 2018. "Synapse Molecular Complexity and the Plasticity Behaviour Problem." *Brain and Neuroscience Advances* 2: 1–7.

Green, Sara, Michael R. Dietrich, Sabine Leonelli, and Rachel Ankeny. 2018. "'Extreme' Organisms and the Problem of Generalization: Interpreting the Krogh Principle." *History and Philosophy of the Life Sciences* 40: 1–22.

Green, Sara, Maria Serban, Raphael Scholl, Nicholaos Jones, Ingo Brigandt, and William Becht. 2018. "Network Analyses in Systems Biology: New Strategies for Dealing with Biological Complexity." *Synthese* 195: 1751–1777.

Grene, Marjorie. 1963. *A Portrait of Aristotle.* London: Faber and Faber.

Grossmann, Henryk. 2009. "The Social Foundations of the Mechanistic Philosophy and Manufacture." In *The Social and Economic Roots of the Scientific Revolution: Texts by Boris Hessen and Henryk Grossmann*, edited by Gideon Freudenthal and Peter McLaughlin, 103–156. Berlin: Springer.

Guttinger, Stephan. 2018. "A Process Ontology for Macromolecular Biology." In *Everything Flows*, edited by Daniel J. Nicholson and John Dupré, 303–320. Oxford: Oxford University Press.

Guzzardi, Luca. 2021. "Holding the Hand of History: Mach on the History of Science, the Analysis of Sensations, and the Economy of Thought." In *Interpreting Mach: Critical Essays*, edited by John Preston, 14–183. Cambridge: Cambridge University Press.

Hacking, Ian. 1983. *Representing and Intervening: Introductory Topics in the Philosophy of Natural Sciences*. Cambridge: Cambridge University Press.

Haddock, Adrian, and Fiona MacPherson, eds. 2011. *Disjunctivism: Perception, Action, Knowledge*. Oxford: Oxford University Press.

Hadot, Pierre. 1995. *Philosophy as a Way of Life*. Oxford, UK: Blackwell.

Hadot, Pierre. 2006. *The Veil of Isis: An Essay on the History of the Idea of Nature*. Translated by Michael Chase. Cambridge, MA: Belknap Press of Harvard University Press.

Haesemeyer, Martin, Alexander F. Schier, and Florian Engert. 2019. "Convergent Temperature Representations in Artificial and Biological Neural Networks." *Neuron* 103: 1123–1134.

Halina, Marta. 2021. "Insightful Artificial Intelligence." *Mind and Language* 36: 315–329.

Hardalupas, Mahi. 2021. *How Neural Is a Neural Net? Bio-inspired Computational Models and Their Impact on the Multiple Realization Debate*. Unpublished doctoral dissertation, University of Pittsburgh. http://d-scholarship.pitt.edu/40634/.

Harrington, A. 1996. *Reenchanted Science: Holism in German Culture from Wilhelm II to Hitler*. Princeton, NJ: Princeton University Press.

Hartline, H. K. 1938. "The Response of Single Optic Nerve Fibres of the Vertebrate Eye to Illumination of the Retina." *American Journal of Physiology* 121: 400–415.

Hassabis, Demis, Dharshan Kumaran, Christopher Summerfield, and Matthew Botvinick. 2017. "Neuroscience-Inspired Artificial Intelligence." *Neuron* 95: 245–258.

Hasson, Uri, Samuel A. Nastase, and Ariel Goldstein. 2020. "Direct Fit to Nature: An Evolutionary Perspective on Biological and Artificial Neural Networks." *Neuron* 105: 416–434.

Haueis, Phillip. 2018. "Beyond Cognitive Myopia: A Patchwork Approach to the Concept of Neural Function." *Synthese* 195: 5373–5402.

Haugeland, John. 1978. "The Nature and Plausibility of Cognitivism." *Behavioral and Brain Sciences* 2: 215–226.

Haugeland, John. 1998a. "Analog and Analog." In *Having Thought: Essays in the Metaphysics of Mind*, 75–88. Cambridge, MA: Harvard University Press.

Originally published in 1981.

Haugeland, John. 1998b. "Mind Embodied and Embedded." In *Having Thought: Essays in the Metaphysics of Mind*, 207–237. Cambridge, MA: Harvard University Press.

Hawking, Stephen. 1995. *A Brief History of Time*. London: Bantam Books.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. "Deep Residual Learning for Image Recognition." arXiv 1606.03490. https://arxiv.org/abs/1512.03385.

Hebb, D. O. 1949. *The Organization of Behavior: A Neuropsychological Theory*. New York: John Wiley and Sons.

Hebb, D. O. 1960. "The American Revolution." *American Psychologist* 15 (12): 735–745.

Heeger, D. J. 1992. "Normalization of Cell Responses in Cat Striate Cortex." *Visual Neuroscience* 9: 181–197.

Heidegger, Martin. 1995. *The Fundamental Concepts of Metaphysics*. Translated by William McNeill and Nicholas Walker. Bloomington: Indiana University Press.

Hempel, Carl G. 1965. *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*. New York: Free Press.

Hermann, Katherine L., Ting Chen, and Simon Kornblith. 2020. "The Origins and Prevalence of Texture Bias in Convolutional Neural Networks." *NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing Systems*. Article No.: 1595: 19000–19015. https://dl.acm.org/doi/abs/10.5555/3495724.3497319.

Hesse, Mary B. 1955. "Action at a Distance in Classical Physics." *Isis* 46 (4): 337–353.

Hesse, Mary B. 1962. *Force and Fields: A Study of Action at a Distance in the History of Physics*. New York: Philosophical Library.

Hesse, Mary B. 1966. *Models and Analogies in Science*. Notre Dame, IN: Notre Dame University Press.

Hesse, Mary B. 1994. "How to Be Postmodern without Being a Feminist." *The Monist* 77 (4): 445–461.

Hesse, Mary B. 1995. "Models, Metaphors and Truth." In *From a Metaphorical Point of View: A Multidisciplinary Approach to the Cognitive Content of Metaphor*, edited by Zdravko Radman, 351–372. Berlin: De Gruyter.

Hilgetag, Claus C., and Alexandros Goulas. 2020. "'Hierarchy' in the Organization of Brain Networks." *Philosophical Transactions of the Royal Society B* 375: 20190319. doi: http://dx.doi.org/10.1098/rstb.2019.0319.

Hoff, Johannes. 2013. *The Analogical Turn: Rethinking Modernity With Nicholas of Cusa*. Grand Rapids, MI: William B. Eerdmans.

Hohwy, Jakob. 2014. *The Predictive Mind*. Oxford: Oxford University Press.

Hooker, Giles, and Cliff Hooker. 2018. "Machine Learning and the Future of Realism." *Spontaneous Generations* 9 (1): 174–182.

Horkheimer, M. 2002. "The Latest Attack on Metaphysics." In *Critical Theory: Selected Essays*, 132–187. New York: Continuum.

Originally published in 1937.

Horkheimer, M. 2013. *Eclipse of Reason*. Mansfield Centre, CT: Martino Publishing.

Originally published in 1947.

Horkheimer, M., and T. W. Adorno. 2002. *Dialectic of Enlightenment: Philosophical Fragments*. Translated by Edmund Jephcott. Stanford, CA: Stanford University Press.

Originally published in 1947.

Hough, Theodore. 1915. "The Classification of Nervous Reactions." *Science* 41 (1055): 407–418.

Hubel, David H. 1957. "Tungsten Microelectrode for Recording from Single Units." *Science* 125: 549–550.

Hubel, David H. 1959. "Single Unit Activity in Striate Cortex of Unrestrained Cats." *Journal of Physiology* 147: 226–238.

Hubel, David H., and Torsten N. Wiesel. 1962. "Receptive Fields, Binocular Interaction and Functional Architecture in the Cat's Visual Cortex." *Journal of Physiology* 160: 106–154.

Hubel, David H., and Torsten N. Wiesel. 1968. "Receptive Fields and Functional Architecture of Monkey Striate Cortex." *Journal of Physiology* 195: 215–244.

Hubel, David H., and Torsten N. Wiesel. 1977. "Ferrier Lecture: Functional Architecture of Macaque Monkey Visual Cortex." *Proceedings of the Royal Society of London. Series B, Biological Sciences* 198 (1130): 1–59.

Hubel, David H., and Torsten N. Wiesel. 1998. "Early Exploration of the Visual Cortex." *Neuron* 20: 401–412.

Hull, Clark, and H. D. Baernstein. 1929. "A Mechanical Parallel to the Conditioned Reflex." *Science* 70 (1801): 14–15.

Husbands, Philip, and Owen Holland. 2008. "The Ratio Club: A Hub of British Cybernetics." In *The Mechanical Mind in History*, edited by Philip Husbands, Owen Holland and Michael Wheeler, 91–148. Cambridge, MA: MIT Press.

Husserl, E. 1970. *The Crisis of European Sciences and Transcendental Phenomenology*. Translated by D. Carr. Evanston, IL: Northwestern University Press.

Hutchins, Barnaby R., Christoffer Basse Eriksen, and Charles T. Wolfe. 2016. "The Embodied Descartes: Contemporary Readings of L'Homme." In *Descartes' Treatise on*

*Man and Its Reception*, edited by Delphine Antoine-Mahut and Stephen Gaukroger, 287–304. Cham, Switzerland: Springer Nature.

Hutto, Daniel D., and Erik Myin. 2014. "Neural Representations Not Needed—No More Pleas, Please." *Phenomenology and the Cognitive Sciences* 13: 241–256.

Illari, Phyllis. 2013. "Mechanistic Explanation: Integrating the Ontic and Epistemic." *Erkenntnis* 78: 237–255.

Ilyas, Andrew, Logan Engstrom, Shibani Santurkar, Brandon Tran, Dimitris Tsipras, and Aleksander Madry. 2019. "Adversarial Examples Are Not Bugs, They Are Features." https://arxiv.org/abs/1905.02175v4.

Ivanova, Milena. 2020. "Beauty, Truth and Understanding." In *The Aesthetics of Science: Beauty, Imagination and Understanding*, edited by Milena Ivanova and Steven French, 86–103. London: Routledge.

Jackson, Frank. 1998. *From Metaphysics to Ethics: A Defence of Conceptual Analysis*. Oxford: Oxford University Press.

James, William. 1890. *Principles of Psychology. Vol. 1*. New York: Henry Holt & Co.

James, William. 1902. *Varieties of Religious Experience: A Study in Human Nature*. New York: Modern Library.

Jarosiewicz, Beata, Steven M. Chase, George W. Fraser, Meel Velliste, Robert E. Kass, and Andrew B. Schwartz. 2008. "Functional Network Reorganization during Learning in a Brain-Computer Interface Paradigm." *Proceedings of the National Academy of Sciences USA* 105 (49): 19486–19491.

Jaworski, William. 2016. *Structure and the Metaphysics of Mind*. Oxford: Oxford University Press.

Jennings, Herbert Spencer. 1906. *Behavior of the Lower Organisms*. New York: Columbia University Press.

Jolley, Nicholas. 1997. "Introduction." In *Malebranche: Dialogues on Metaphysics and Religion*, edited by Nicholas Jolley and David Scott, viii–xxxiv. Cambridge: Cambridge University Press.

Jonas, E., and K. Kording. 2017. "Could a Neuroscientist Understand a Microprocessor?" *PLoS Computational Biology* 13: e1005268.

Jonas, Hans. 1954. "The Nobility of Sight." *Philosophy and Phenomenological Research* 14 (4): 507–519.

Jones, Kelly E., Patrick K. Campbell, and Richard A. Normann. 1992. "A Glass/Silicon Composite Intracortical Electrode Array." *Annals of Biomedical Engineering* 20: 423–437.

Jorgenson, Lyric A., William T. Newsome, David J. Anderson, et al. 2015. "The BRAIN Initiative: Developing Technology to Catalyse Neuroscience Discovery." *Philosophical Transactions of the Royal Society B* 370: 20140164.

Juavinett, Ashley L., Jeffrey C. Erlich, and Anne K. Churchland. 2018. "Decision-Making Behaviors: Weighing Ethology, Complexity, and Sensorimotor Compatibility." *Current Opinion in Neurobiology* 49: 42–50.

Kant, Immanuel. 1998. *Critique of Pure Reason*. Translated by Paul Guyer and Allen W. Wood. Cambridge Edition of the Works of Immanuel Kant. Cambridge: Cambridge University Press.

Previously published in 1781 and 1787.

Kant, Immanuel. 1929. *The Critique of Pure Reason*. Translated by Norman Kemp Smith. Basingstoke, UK: Palgrave.

Originally published in 1781 and 1787.

Kanwisher, Nancy, and Galit Yovel. 2006. "The Fusiform Face Area: A Cortical Region Specialized for the Perception of Faces." *Philosophical Transactions of the Royal Society, B* 361: 2109–2128.

Kapadia, M. K., G. Westheimer, and C. D. Gilbert. 1999. "Dynamics of Spatial Summation in Primary Visual Cortex of Alert Monkeys." *Proceedings of National Academy of Science USA* 96 (21): 12073–12078.

Kaplan, David Michael. 2011. "Explanation and Description in Computational Neuroscience." *Synthese* 183 (3): 339–73.

Kar, K., Jonas Kubilius, Kailyn Schmidt, Elias B. Issa, and J. J. DiCarlo. 2019. "Evidence That Recurrent Circuits Are Critical to the Ventral Stream's Execution of Core Object Recognition Behavior." *Nature Neuroscience* 22: 974–983.

Kastenhofer, Karen. 2013. "Two Sides of the Same Coin? The (Techno)epistemic Cultures of Systems and Synthetic Biology." *Studies in History and Philosophy of Biological and Biomedical Sciences* 44: 130–140.

Kastenhofer, Karen, and Jan C. Schmidt. 2011. "Technoscientia Est Potentia? Contemplative, Interventionist, Constructionist and Creationist Idea(l)s in (Techno) science." *Poiesis Prax* 8: 125–149.

Katz, Yarden. 2012. "Noam Chomsky on Where Artificial Intelligence Went Wrong." *The Atlantic*, November 1.

Kay, Lily E. 2001. "From Logical Neurons to Poetic Embodiments of Mind: Warren S. McCulloch's Project in Neuroscience." *Science in Context* 14 (4): 591–614.

Keijzer, Fred. 2015. "Moving and Sensing without Input and Output: Early Nervous Systems and the Origins of the Animal Sensorimotor Organization." *Biology and Philosophy* 30: 311–331.

Kell, Alexander J. E., and Josh H. McDermott. 2019. "Deep Neural Network Models of Sensory Systems: Windows onto the Role of Task Constraints." *Current Opinion in Neurobiology* 55: 121–132.

Khaligh-Razavi, S.-M., and N. Kriegeskorte. 2014. "Deep Supervised, But Not Unsupervised, Models May Explain IT Cortical Representation." *PLoS Computational Biology* 10 (11): e1003915.

Kheradpisheh, Saeed Reza, Masoud Ghodrati, Mohammad Ganjtabesh, and Timothée Masquelier. 2016. "Deep Networks Can Resemble Human Feed-forward Vision in Invariant Object Recognition." *Scientific Reports* 6: 32672.

Kline, Ronald R. 2015. *The Cybernetics Moment: Or Why We Call Our Age the Information Age*. Baltimore: John Hopkins University Press.

Knuuttila, Tarja, and Andrea Loettgers. 2014. "Varieties of Noise: Analogical Reasoning in Synthetic Biology." *Studies in History and Philosophy of Science* 48: 76–88.

Koch, Christof, and Michael Buice. 2015. "A Biological Imitation Game." *Cell* 163: 277–280.

Koyama, S., S. M. Chase, A. S. Whitford, M. Velliste, A. B. Schwartz, and R. E. Kass. 2010. "Comparison of Brain-Computer Interface Decoding Algorithms in Open-Loop and Closed-Loop Control." *Journal of Computational Neuroscience* 29 (1–2): 73–87. doi: 10.1007/s10827-009-0196-9.

Krakauer, John W. 2022. "Representation in Cognitive Science by Nicholas Shea: But Is It Thinking? The Philosophy of Representations Meets Systems Neuroscience." *Studies in History & Philosophy of Science* 92: 267–269. https://doi.org/10.1016/j.shpsa.2021.05.014.

Krakauer, John W., Asif A. Ghazanfar, Alex Gomez-Marin, Malcolm A. MacIver, and David Poeppel. 2017. "Neuroscience Needs Behavior: Correcting a Reductionist Bias." Neuron 93: 480–490.

Kriegeskorte, Nikolaus, and Jörn Diedrichsen. 2019. "Peeling the Onion of Brain Representations." *Annual Review of Neuroscience* 42: 407–432.

Krizhevsky, A., I. Sutskever, and G. E. Hinton. 2012. "ImageNet Classification with Deep Convolutional Neural Networks." *Proceedings of NeurIPS*: 1106–1114. https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

Kuffler, S. W. 1953. "Discharge Patterns and Functional Organization of Mammalian Retina." *Journal of Neurophysiology* 16: 37–68.

Kunkel, Susanne, Tobias C. Potjans, Jochen M. Eppler, Hans Ekkehard Plesser, Abigail Morrison, and Markus Diesmann. 2012. "Meeting the Memory Challenges of Brain-Scale Network Simulation." *Frontiers in Neuroinformatics*. https://doi.org/10.3389/fninf.2011.00035.

Ladyman, James, James Lambert, and Karoline Wiesner. 2013. "What Is a Complex System?" *European Journal for Philosophy of Science* 3: 33–67.

Ladyman, James, and Karoline Wiesner. 2020. *What Is a Complex System?* New Haven, CT: Yale University Press.

Laiwalla, Farah, and Arto Nurmikko. 2019. "Future of Neural Interfaces." In *Neural Interface: Frontiers and Applications*, edited by Xiaoxiang Zheng, 225–241. Singapore: Springer Singapore.

Lake, Brenden M., Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman. 2017. "Building Machines That Learn and Think Like People." *Behavioral and Brain Sciences* 40: e253. doi:10.1017/S0140525X16001837.

Lande, Kevin. 2019. "Do You Compute?" *Aeon*. https://aeon.co/essays/your-brain-probably-is-a-computer-whatever-that-means. Accessed May 10, 2021.

Landeweerd, Laurens. 2021. T*ime, Life & Memory: Bergson and Contemporary Science, Library of Ethics and Applied Philosophy*. Cham, Switzerland: Springer Nature.

Lappin, Shalom. 2021. *Deep Learning and Linguistic Representation.* Boca Raton, FL: CRC Press (Taylor & Francis).

Largent, Mark A. 2009. "Darwin's Analogy between Artificial and Natural Selection in the Origin of Species." In *Cambridge Companion to the "Origin of Species,"* edited by Michael Ruse and Robert J. Richards, 15–29. Cambridge: Cambridge University Press.

Larkum, Matthew Evan. 2022. "Are Dendrites Conceptually Useful?" *Neuroscience* 489: 4–14.

Lashley, Karl S. 1931. "Cerebral Control Versus Reflexology: A Reply to Professor Hunter." *Journal of General Psychology* 5: 3–19.

Latour, Bruno. 1992. "Pasteur on Lactic Acid Yeast: A Partial Semiotic Analysis." *Configurations* 1 (1): 129–146.

LeCun, Y., B. Boser, J. S. Denker, et al. 1989. "Backpropagation Applied to Handwritten Zip Code Recognition." *Neural Computation* 1: 541–555.

Lee, Jonny, and Joe Dewhurst. 2021. "The Mechanistic Stance." *European Journal for Philosophy of Science* 11 (20). doi: 10.1007/s13194-020-00341-6.

Lehky, Sidney R., and Terrence J. Sejnowski. 1988. "Network Model of Shape-from-Shading: Neural Function Arises from Both Receptive and Projective Fields." *Nature* 333 (6172): 452–454.

Lenk, Hans. 2017. "A Scheme-Interpretationist and Actionistic Scientific Realism." In *Varieties of Scientific Realism: Objectivity and Truth in Science*, edited by Evandro Agazzi, 257–276. Cham, Switzerland: Springer.

Lenk, Hans. 2019. "A Methodological Interpretation of Mechanistic Explanations." In *Mechanistic Explanations in Physics and Beyond,* edited by Brigette Falkenburg and Gregor Schiemann, 143–162. Berlin: Springer.

Lent, Roberto, Frederico A. C. Azevedo, Carlos H. Andrade-Moraes, and Ana V. O. Pinto. 2012. "How Many Neurons Do You Have? Some Dogmas of Quantitative Neuroscience under Revision." *European Journal of Neuroscience* 35: 1–9.

Leuba, G., and R. Kraftsik. 1994. "Changes in Volume, Surface Estimate, Three-Dimensional Shape and Total Number of Neurons of the Human Primary Visual Cortex from Midgestation until Old Age." *Anatomy and Embryology* 190: 351–366.

Levy, Arnon. 2014. "What Was Hodgkin and Huxley's Achievement?" *British Journal for the Philosophy of Science* 65: 469–492.

Levy, Arnon. 2018. "Idealization and Abstraction: Refining the Distinction." *Synthese* 198(24): 5855–5872. doi: https://doi.org/10.1007/s11229-018-1721-z.

Liberti, William A., Tobias A. Schmid, Angelo Forli, Madeleine Snyder, and Michael M. Yartsev. 2022. "A Stable Hippocampal Code in Freely Flying Bats." *Nature* 604: 98–103. https://doi.org/10.1038/s41586-022-04560-0.

Lillicrap, Timothy P., and K. Kording. 2019. "What Does It Mean to Understand a Neural Network?" https://arxiv.org/abs/1907.06374.

Lindsay, Grace. W. 2020. "Convolutional Neural Networks as a Model of the Visual System: Past, Present, and Future." *Journal of Cognitive Neuroscience*. https://doi.org/10.1162/jocn_a_01544.

Lindsay, Grace. W. 2021. *Models of the Mind: How Physics, Engineering and Mathematics Have Shaped Our Understanding of the Brain*. London: Bloomsbury.

Loeb, Jacques. 1900. *Comparative Physiology of the Brain and Comparative Psychology*. New York: J. P. Putnam's Sons.

Loeb, Jacques. 1912. *The Mechanistic Conception of Life*. Chicago: University of Chicago Press.

Longino, Helen E. 1995. "Gender, Politics, and the Theoretical Virtues." *Synthese* 104 (3): 383–397.

Longino, Helen. 2006. "Theoretical Pluralism and the Scientific Study of Behavior." In *Minnesota Studies in the Philosophy of Science, Vol. 19, Scientific Pluralism*, edited by S. H. Kellert, H. E. Longino, and C. K. Waters, 102–131. Minneapolis: University of Minnesota Press.

Longino, Helen. 2013. *Studying Human Behavior*. Chicago: University of Chicago Press.

Longuenesse, Béatrice. 1998. *Kant and the Capacity to Judge*. Translated by Charles T. Wolfe. Princeton, NJ: Princeton University Press.

Originally published in 1993.

Mach, Ernst. 1895. "The Economical Nature of Physical Inquiry." In *Popular Scientific Lectures*, edited by Thomas McCormack. Chicago: Open Court.

Originally published in 1882.

Mach, Ernst. 1910. *Populär-wissenschaftliche Vorlesungen.* Leipzig, Germany: Johann Barth.

Mach, Ernst. 1911. *History and Root of the Principle of the Conservation of Energy*. Translated by P. E. B. Jourdain. Open Court.

Originally published in 1872.

Mach, Ernst. 1914. *The Analysis of Sensations, and the Relation of the Physical to the Psychical*. Translated by C. M. Williams. Chicago: Open Court.

Originally published in 1886.

Mach, Ernst. 1919. *The Science of Mechanics*. Translated by Thomas J. McCormack. Chicago: Open Court.

Originally published in 1883.

Machamer, Peter. 2009. "Learning, Neuroscience, and the Return of Behaviorism." In *Handbook of Philosophy and Neuroscience*, edited by John Bickle, 166–176. Oxford: Oxford University Press.

Machamer, Peter, Lindley Darden, and Carl F. Craver. 2000. "Thinking about Mechanisms." *Philosophy of Science* 67 (1): 1–25.

Maffei, L., and A. Fiorentini. 1976. "The Unresponsive Regions of Visual Cortical Receptive Fields." *Vision Research* 16: 1131–1139.

Maley, Corey J. 2021. "The Physicality of Representation." *Synthese* 199: 14725–14750. doi: https://doi.org/10.1007/s11229-021-03441-9.

Mao, Bu-Qing, Farid Hamzei-Sichani, Dmitriy Aronov, Robert C Froemke, and Rafael Yuste. 2001. "Dynamics of Spontaneous Activity in Neocortical Slices." *Neuron* 32 (5): 883–898.

Marcus, Gary. 2015. "The Computational Brain." In *The Future of the Brain*, edited by Gary Marcus and Jeremy Freeman, 205–218. Princeton, NJ: Princeton University Press.

Marder, Eve, and J. M. Goaillard. 2006. "Variability, Compensation and Homeostasis in Neuron and Network Function." *Nature Reviews Neuroscience* 7 (7): 563–74. doi: 10.1038/nrn1949.

Marder, Eve, Timothy O'Leary, and Sonal Shruti. 2014. "Neuromodulation of Circuits with Variable Parameters: Single Neurons and Small Circuits Reveal Principles of State-Dependent and Robust Neuromodulation." *Annual Review of Neuroscience* 37: 329–346.

Marr, David. 1982. *Vision*. San Francisco: W. H. Freeman.

Marr, David, and Shimon Ullman. 1981. "Directional Selectivity and Its Use in Early Visual Processing." *Proceedings of the Royal Society of London B* 211: 151–180.

Martin, Gottfried. 1955. *Kant's Metaphysics and Theory of Science*. Translated by P. G. Lucas. Manchester, UK: University of Manchester Press.

Originally published in 1951.

Masland, Richard H., and Paul R. Martin. 2007. "The Unsolved Mystery of Vision." *Current Biology* 17 (15): R577–R582.

Massimi, Michela. 2011. "From Data to Phenomena: A Kantian Stance." *Synthese* 182: 101–116.

Massimi, Michela. 2018a. "Four Kinds of Perspectival Truth." Philosophy and Phenomenological Research 96 (2): 342–359.

Massimi, Michela. 2018b. "Perspectival Modeling." *Philosophy of Science* 85 (3): 335–359.

Massimi, Michela. 2018c. "Perspectivism." In *Routledge Handbook of Scientific Realism*, edited by J. Saatsi, 164–175. Oxford, UK: Routledge.

Massimi, Michela. 2022. *Perspectival Realism*. Oxford: Oxford University Press.

Massimi, Michela, and Casey McCoy, eds. 2020. *Understanding Perspectivism: Scientific Challenges and Methodological Prospects*. New York: Routledge.

Mayr, Ernst. 1961. "Cause and Effect in Biology." *Science* 134: 1501–1506.

Mayr, Ernst. 1988. "The Multiple Meanings of Teleological." In *Toward a New Philosophy of Biology*, edited by Ernst Mayr, 38–66. Cambridge, MA: Belknap Press of Harvard University Press.

McAllister, J. 1997. "Phenomena and Patterns in Data Sets." *Erkenntnis* 47 (1): 217–228.

McAllister, James, W. 2007. "Model Selection and the Multiplicity of Patterns in Empirical Data." *Philosophy of Science* 74 (5): 884–894.

McCulloch, Gregory. 1995. *The Mind and Its World*. London: Routledge.

McCulloch, Warren S. 2016. *Embodiments of Mind*. Cambridge, MA: MIT Press.

Originally published in 1965.

McCulloch, Warren S., and Walter Pitts. 1943. "A Logical Calculus of the Ideas Immanent in Nervous Activity." *Bulletin of Mathematical Biophysics* 5: 115–133.

McDowell, John. 1996. *Mind and World*. Cambridge, MA: Harvard University Press.

McDowell, John. 1998. "Singular Thought and the Extent of Inner Space." In *Meaning, Knowledge, and Reality*, 228–259. Cambridge, MA: Harvard University Press.

McDowell, John. 2010. "Tyler Burge on Disjunctivism." *Philosophical Explorations* 13 (3): 243–255.

McDowell, John. 2013. "Tyler Burge on Disjunctivism (II)." *Philosophical Explorations* 16 (3): 259–279.

McIntosh, L. T., N. Maheswaranathan, A. Nayebi, S. Ganguli, and S. A. Baccus. 2016. "Deep Learning Models of the Retinal Response to Natural Scenes." *Advances in Neural Information Processing Systems* 29: 1369–1377.

McNulty, Jacob. 2021. "From Analytic Pragmatism to Historical Materialism: Frankfurt School Critical Theory and the Quine-Duhem Thesis." *European Journal for Philosophy*. doi: 10.1111/ejop.12737.

Mechler, F., and D. L. Ringach. 2002. "On the Classification of Simple and Complex Cells." *Vision Research* 42: 1017–1033.

Medina, Eden. 2014. *Cybernetic Revolutionaries: Technology and Politics in Allende's Chile*. Cambridge, MA: MIT Press.

Menzies, Peter. 2007. "Causation in Context." In *Causation, Physics, and the Constitution of Reality: Russell's Republic Revisited*, edited by Huw Price and Richard Corry, 191–223. Oxford: Oxford University Press.

Merleau-Ponty, Maurice. 1967. *The Structure of Behaviour*. Translated by Alden L. Fisher. Boston: Beacon Press.

Originally published in 1942.

Merleau-Ponty, Maurice. 2001. "Eye and Mind." In *Continental Aesthetics: An Anthology*, edited by Richard Kearney and David Rasmussen, 288–306. Oxford, UK: Blackwell.

Originally published in 1961.

Merton, R. K. 1970. *Science, Technology & Society in Seventeeth-Century England*. New York: Howard Fertig.

Originally published in 1938.

Meyer, Travis, and N. C. Rust. 2018. "Single-Exposure Visual Memory Judgments Are Reflected in Inferotemporal Cortex." *eLife* 7: e32259.

Miłkowski, Marcin. 2018. "From Computer Metaphor to Computational Modeling: The Evolution of Computationalism." *Minds and Machines* 28(3): 515–541. doi: 10.1007/s11023-018-9468-3.

Millikan, Ruth Garrett. 2020. "Neuroscience and Teleosemantics." *Synthese* 199: 2457–2465. doi: 10.1007/s11229-020-02893-9.

Mills, John A. 1998. *Control: A History of Behavioral Psychology*. New York: New York University Press.

Milner, David, and Mel Goodale. 1995. *The Visual Brain in Action*. Oxford: Oxford University Press.

Mitchell, Melanie. 2009a. *Complexity: A Guided Tour*. Oxford: Oxford University Press.

Mitchell, Melanie. 2019. *Artificial Intelligence: A Guide for Thinking Humans*. London: Pelican.

Mitchell, Sandra D. 2003. *Biological Complexity and Integrative Pluralism*. Cambridge: Cambridge University Press.

Mitchell, Sandra D. 2009b. *Unsimple Truths: Science, Complexity, and Policy*. Chicago: University of Chicago Press.

Mitchell, Sandra D. 2020. "Perspectives, Representation, and Integration." In *Understanding Perspectivism: Scientific Challenges and Methodological Prospects*, edited by Michela Massimi and Casey McCoy, 178–193. New York: Routledge.

Mitchell, Sandra D., and Angela M. Gronenborn. 2017. "After Fifty Years, Why Are Protein X-ray Crystallographers Still in Business?" *British Journal for the Philosophy of Science* 68 (3): 703–723. doi: 10.1093/bjps/axv051.

Moinat, Frédéric. 2012. *Le vivant et sa naturalisation: Le problème du naturalisme en biologie chez Husserl et le jeune Merleau-Ponty*. Dordrecht, Netherlands: Springer.

Montijn, J. S., G. T. Meijer, C. S. Lansink, and C. M. A. Pennartz. 2016. "Population-Level Neural Codes Are Robust to Single-Neuron Variability from a Multidimensional Coding Perspective." *Cell Reports* 16: 2486–2498.

Moore, J. 2005. "Some Historical and Conceptual Background to the Development of B. F. Skinner's 'Radical Behaviorism'—Part 2." *Journal of Mind and Behavior* 26 (1–2): 95–124.

Morar, Florin-Stefan. 2015. "Reinventing Machines: The Transmission History of the Leibniz Calculator." *British Society for the History of Science* 48 (1): 123–146.

Morrison, Margaret. 2011. "One Phenomenon, Many Models: Inconsistency and Complementarity." *Studies in History and Philosophy of Science* 42: 342–351.

Movshon, J. A. 2021. "Obituary: Horace Basil Barlow (1921–2020)." *Perception* 50 (2): 183–192.

Movshon, J. A., I. D. Thompson, and D. J. Tolhurst. 1978. "Spatial Summation in the Receptive Fields of Simple Cells in the Cat's Striate Cortex." *Journal of Physiology* 283: 53–77.

Musall, Simon, Matthew T. Kaufman, Ashley L. Juavinett, Steven Gluf, and Anne K. Churchland. 2019. "Single-Trial Neural Dynamics Are Dominated by Richly Varied Movements." *Nature Neuroscience* 22: 1677–1686.

Musall, Simon, Anne E Urai, David Sussillo, and Anne K Churchland. 2019. "Harnessing Behavioral Diversity to Understand Neural Computations for Cognition." *Current Opinion in Neurobiology* 58: 229–238.

Napoletani, D., M. Panza, and D. C. Struppa. 2011. "Agnostic Science: Towards a Philosophy of Data Analysis." *Foundations of Science* 16: 1–20.

Nastase, Samuel A., Ariel Goldstein, and Uri Hasson. 2020. "Keep It Real: Rethinking the Primacy of Experimental Control in Cognitive Neuroscience." *NeuroImage* 222: 117254.

Neander, Karen. 2017. *A Mark of the Mental*. Cambridge, MA: MIT Press.

Newman, William R. 2004. *Promethean Ambitions: Alchemy and the Quest to Perfect Nature*. Chicago: University of Chicago Press.

Niell, C. M., and M. P. Stryker. 2010. "Modulation of Visual Responses by Behavioral State in Mouse Visual Cortex." *Neuron* 65: 472–479.

Nietzsche, Friedrich. 1994. *On the Genealogy of Morality*. Cambridge: Cambridge University Press.

Originally published in 1887.

Nordmann, A. 2006. "Collapse of Distance: Epistemic Strategies of Science and Technoscience." *Danish Yearbook of Philosophy* 41: 7–34.

Northcott, Robert. 2022. "Pandemic Modeling, Good and Bad." *Philosophy of Medicine* 3 (1): 1–20.

Norton, John D. 2000. "'Nature Is the Realisation of the Simplest Conceivable Mathematical Ideas': Einstein and the Canon of Mathematical Simplicity." *Studies in History & Philosophy of Modern Physics* 31 (2): 135–170.

Norton, John D. 2007. "Causation as Folk Science." In *Causation, Physics, and the Constitution of Reality: Russell's Republic Revisited*, edited by Huw Price and Richard Corry, 11–44. Oxford: Oxford University Press.

Norvig, Peter. 2012. "On Chomsky and the Two Cultures of Statistical Learning." accessed September 7, 2022. http://norvig.com/chomsky.html.

Nowogrodzki, Anna. 2018. "How Cerebral Organoids Are Guiding Brain-Cancer Research and Therapies." *Nature* 561: S48–S49.

Nussbaum, Martha C., and H. Putnam. 1992. "Changing Aristotle's Mind." In *Essays on Aristotle's De Anima*, edited by Martha C. Nussbaum and Amélie Oksenberg Rorty, 27–56. Oxford: Oxford University Press.

Nutt, David, and Anna Need. 2014. "Where Now for Schizophrenia Research?" *European Neuropsychopharmacology* 24: 1181–1187.

O'Keefe, J. A., and D. H. Conway. 1978. "Hippocampal Place Units in the Freely Moving Rat: Why They Fire Where They Fire." *Experimental Brain Research* 31: 573–590.

O'Leary, Timothy, Alex H. Williams, Jonathan S. Caplan, and Eve Marder. 2013. "Correlations in Ion Channel Expression Emerge from Homeostatic Tuning Rules." *PNAS* 110 (28): E2645–E2654.

Olshausen, B. A., and David J. Field. 2005. "How Close Are We to Understanding V1?" *Neural Computation* 17 (8): 1665–1699.

Olshausen, B. A., and David J. Field. 2006. "What Is the Other 85 Percent of V1 Doing?" In *23 Problems in Systems Neuroscience*, edited by J. Leo van Hemmen and Terrence J. Sejnowski, 182–211. Oxford: Oxford University Press.

O'Malley, M. 2009. "Making Knowledge in Synthetic Biology: Design Meets Kludge." *Biological Theory* 4 (4): 378–389.

Omrani, Mohsen, Matthew T. Kaufman, Nicholas G. Hatsopoulos, and Paul D. Cheney. 2017. "Perspectives on Classical Controversies about the Motor Cortex." *Journal of Neurophysiology* 118: 1828–1848.

O'Neill, John, and Thomas Uebel. 2004. "Horkheimer and Neurath: Restarting a Disrupted Debate." *European Journal of Philosophy* 12 (1): 75–105.

Otis, L. 2007. *Müller's Lab*. Oxford: Oxford University Press.

Papert, Seymour. 2016. "Introduction." In *Embodiments of Mind*, edited by Warren S. McCulloch, xxix–xxxv. Cambridge, MA: MIT Press.

Originally published in 1965.

Parisi, G. I., R. Kemker, J. L. Part, C. Kanan, and S. Wermter. 2019. "Continual Lifelong Learning with Neural Networks: A Review." *Neural Networks* 113: 54–71.

Park, Jae Sung, Chandra Bhagavatula, Roozbeh Mottaghi, Ali Farhadi, and Yejin Choi. 2020. "VisualCOMET: Reasoning about the Dynamic Context of a Still Image." In *Computer Vision—ECCV 2020. Lecture Notes in Computer Science*, vol. 12350, edited by A. Vedaldi, H. Bischof, T. Brox, and J. M. Fram. Springer, Cham. https://doi.org/10.1007/978-3-030-58558-7_30

Parker, Philip R. L., Morgan A. Brown, Matthew C. Smear, and Cristopher M. Niell. 2020. "Movement-Related Signals in Sensory Areas: Roles in Natural Behavior." *Trends in Neurosciences* 43 (8): 581–595.

Pasnau, Robert. 2011. *Metaphysical Themes: 1274–1671*. Oxford: Oxford University Press.

Patton, Lydia. 2021. "Abstraction, Pragmatism, and History in Mach's Economy of Science." In *Interpreting Mach: Critical Essays*, edited by John Preston, 142–163. Cambridge: Cambridge University Press.

Pauly, Philip. 1987. *Controlling Life: Jacques Loeb & the Engineering Ideal in Biology*. Oxford: Oxford University Press.

Pavlov, Ivan. 1960. *Conditioned Reflexes: An Investigation of the Physiological Activity of the Cerebral Cortex*. New York: Dover Publications.

Originally published in 1927.

Pecere, Paolo. 2020. *Soul, Mind and Brain from Descartes to Cognitive Science*. Cham, Switzerland: Springer Nature.

Peterman, Alison. 2021. "The World Soul in Early Modern Philosophy." In *World Soul: A History*, edited by James Wilberding, 186–222. Oxford: Oxford University Press.

Peterson, Erik L. 2016. *The Life Organic: The Theoretical Biology Club and the Roots of Epigenetics*. Pittsburgh: Pittsburgh University Press.

Phemister, Pauline, and Lloyd Strickland. 2015. "Leibniz's Monadological Positive Aesthetics." *British Journal for the History of Philosophy* 23 (6): 1214–1234.

Piccinini, Gualtiero. 2020. *Neurocognitive Mechanisms: Explaining Biological Cognition*. Oxford: Oxford University Press.

Piccinini, Gualtiero, and Carl F. Craver. 2011. "Integrating Psychology and Neuroscience: Functional Analyses as Mechanism Sketches." *Synthese* 183: 283–311.

Pickering, Andrew. 1995. *The Mangle of Practice*. Chicago: University of Chicago Press.

Pickering, Andrew. 2010. *The Cybernetic Brain: Sketches of Another Future*. Chicago: University of Chicago Press.

Pippin, Robert B. 1982. *Kant's Theory of Form*. New Haven, CT: Yale University Press.

Pippin, Robert B. 1987. "Kant on the Spontaneity of Mind." *Canadian Journal of Philosophy* 17 (2): 449–475.

Plato. 1961. "Cratylus." In The Collected Dialogues of Plato, edited by E. Hamilton and H. Cairns, 421–474. Princeton, NJ: Princeton University Press.

Plutynski, Anya. 2020. "Cancer Modeling: The Advantages and Limitations of Multiple Perspectives." In *Understanding Perspectivism: Scientific Challenges and Methodological Prospects*, edited by Michela Massimi and Casey McCoy. New York: Routledge.

Poellner, Peter. 2000. *Nietzsche and Metaphysics*. Oxford: Oxford University Press.

Poincaré, H. J. 1890. "Sur le problème des trois corps et les équations de la dynamique." *Acta Mathematica* 13: 1–270.

Polger, Thomas W. 2006. *Natural Minds*. Cambridge, MA: MIT Press.

Polger, Thomas W., and Lawrence A. Shapiro. 2016. *The Multiple Realization Book*. Oxford: Oxford University Press.

Potochnik, Angela. 2017. *Idealization and the Aims of Science*. Chicago: University of Chicago Press.

Price, Huw. 2011. *Naturalism without Mirrors*. Oxford: Oxford University Press.

Price, Huw, and Richard Corry. 2007. "A Case for Causal Republicanism?" In *Causation, Physics, and the Constitution of Reality: Russell's Republic Revisited*, edited by Huw Price and Richard Corry, 1–10. Oxford: Oxford University Press.

Prinz, Jesse J. 2012. *The Conscious Brain: How Attention Engenders Experience*. Oxford: Oxford University Press.

Psillos, Stathis. 1999. *Scientific Realism: How Science Tracks Truth*. London: Routledge.

Putnam, Hilary. 1975. "The Nature of Mental States. Philosophical Papers, Vol. 2." In *Mind, Language and Reality*, 429–440. Cambridge: Cambridge University Press.

Putnam, Hilary. 1988. *Representation and Reality*. Cambridge, MA: MIT Press.

Putnam, Hilary. 1997. "Philosophy and Our Mental Life." In *The Philosophy of Mind: Classical Problems/Contemporary Issues*, edited by Brian Beakley and Peter Ludlow, 91–99. Cambridge, MA: MIT Press.

Originally published in 1973.

Quine, W. V. O. 1948. "On What There Is." *Review of Metaphysics* 2 (5): 21–38.

Rabinbach, Anson. 1990. *The Human Motor: Energy, Fatigue, and the Origins of Modernity*. Berkeley, CA: University of California Press.

Ramsden, Edmund. 2021. "Behavioral Engineering and the Problems of Animal Misbehavior." In *Nature Remade: Engineering Life, Envisioning Worlds*, edited by Luis Campos, Michael R. Dietrich, Tiago Saraiva and Christian C. Young, 89–102. Chicago: University of Chicago Press.

Ramsey, William M. 2007. *Representation Reconsidered*. Cambridge: Cambridge University Press.

Ramsey, William M. 2020. "Defending Representation Realism." In *Mental Representations*, edited by Joulia Smortchkova, Krzysztof Dolega and Tobias Schlicht, 54–78. Oxford: Oxford University Press.

Rashevsky, Nicolas. 1934. "Foundations of Mathematical Biophysics." *Philosophy of Science* 1 (2): 176–196.

Rashevsky, Nicolas. 1938. *Mathematical Biophysics: Physicomathematical Foundations of Biology*. Chicago: University of Chicago Press.

Remmert, Volker R. 2005. "Galileo, God and Mathematics." In *Mathematics and the Divine: A Historical Study*, edited by T. Koetsier and L. Bergmans, 347–360. Amsterdam: Elsevier.

Requarth, Tim. 2015. "The Big Problem with "Big Science" Ventures—Like the Human Brain Project." *Nautilus*. https://nautil.us/the-big-problem-with-big-science-ventureslike-the-human-brain-project-235387/

Rescorla, Michael. 2013. "Against Structuralist Theories of Computational Implementation." *British Journal of Philosophy of Science* 64: 681–707.

Rice, Collin. 2020. "Universality and the Problem of Inconsistent Models." In *Understanding Perspectivism: Scientific Challenges and Methodological Prospects*, edited by Michela Massimi and Casey McCoy, 85–108. New York: Routledge.

Riskin, Jessica. 2016. *The Restless Clock*. Chicago: University of Chicago Press.

Ritchie, J. Brendan, and Gualtiero Piccinini. 2018. "Computational Implementation." In *Routledge Handbook of the Computational Mind,* edited by Mark Sprevak and Matteo Colombo, 192–204. London: Routledge.

Rodieck, R. W., and J. Stone. 1965. "Response of Cat Retinal Ganglion Cells to Moving Visual Patterns." *Journal of Neurophysiology* 28 (5): 819–832.

Rokni, Uri, Andrew G. Richardson, Emilio Bizzi, and H. Sebastian Seung. 2007. "Motor Learning with Unstable Neural Representations." *Neuron* 54: 653–666.

Rorty, Amélie, and Martha Craven Nussbaum, eds. 1992. *Essays on Aristotle's De Anima*. Oxford: Oxford University Press.

Rosenberg, Alex. 2014. "Can Naturalism Save the Humanities?" In *Philosophical Methodology: The Armchair or the Laboratory*, edited by Matthew C. Haug, 39–42. London: Routledge.

Rosenblatt, Frank. 1958. "The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain." *Psychological Review* 65 (6): 386–408.

Rosenblatt, F. 1962. *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*. New York: Spartan.

Rosenblueth, Arturo, and Norbert Wiener. 1945. "The Role of Models in Science." *Philosophy of Science* 12 (4): 316–321.

Rosenblueth, Arturo, and Norbert Wiener. 1950. "Purposeful and Non-purposeful Behavior." *Philosophy of Science* 17 (4): 318–326.

Rosenblueth, Arturo, Norbert Wiener, and Julian Bigelow. 1943. "Behavior, Purpose and Teleology." *Philosophy of Science* 10 (1): 18–24.

Roskies, Adina L. 2021. "Representational Similarity Analysis in Neuroimaging: Proxy Vehicles and Provisional Representations." *Synthese* 199: 5917–5935. doi: 10.1007/s11229-021-03052-4.

Rossi, Paolo. 1956. "Sulla valutazione delle arti meccaniche nei secoli XVI e XVII." *Rivista Critica di Storia della Filosofia* 11 (2): 126–148.

Rossi, Paolo. 2002. *I filosofi e le macchine*. Milan: Feltrinelli.

Originally published in 1962.

Roux, Sandrine. 2013. "L'ennemi cartésien. Cartésianisme et anti-cartésianisme en philosophie de l'esprit et en sciences cognitives." *Astérion* 11. http://asterion.revues.org/2419.

Rudin, Cynthia, and Joanna Radin. 2019. "Why Are We Using Black Box Models in AI When We Don't Need To? A Lesson from an Explainable AI Competition." *Harvard Data Science Review* 1 (2). doi: 10.1162/99608f92.5a8a3a3d.

Rule, Michael E., Timothy O'Leary, and Christopher D. Harvey. 2019. "Causes and Consequences of Representational Drift." *Current Opinion in Neurobiology* 58: 141–147.

Russell, Bertrand. 1913. "On the Notion of Cause." *Proceedings of the Aristotelian Society* 13: 1–26.

Rust, N. C., and J. A. Movshon. 2005. "In Praise of Artifice." *Nature Neuroscience* 8 (12): 1647–1650.

Sadeh, Sadra, and Claudia Clopath. 2022. "Contribution of Behavioural Variability to Representational Drift." *eLife* 11:e77907. doi: https://doi.org/10.7554/eLife.77907.

Sadtler, Patrick T., Kristin M. Quick, Matthew D. Golub, et al. 2014. "Neural Constraints on Learning." *Nature* 512: 423–426.

Sahasrabuddhe, Kunal, Aamir A Khan, Aditya P Singh, et al. 2021. "The Argo: A High Channel Count Recording System for Neural Recording in Vivo." *Journal of Neural Engineering* 18:015002. doi: https://doi.org/10.1088/1741-2552/abd0ce.

Saleem, A. B., E. Mika Diamanti, Julien Fournier, Kenneth D. Harris, and Matteo Carandini. 2018. "Coherent Encoding of Subjective Spatial Position in Visual Cortex and Hippocampus." *Nature* 562: 124–127.

Salmon, Wesley. 1984. *Scientific Explanation and the Causal Structure of the World.* Princeton, NJ: Princeton University Press.

Sarma, Gopal P., Chee Wai Lee, Tom Portegys, et al. 2018. "OpenWorm: Overview and Recent Advances in Integrative Biological Simulation of Caenorhabditis elegans." *Philosophical Transactions of the Royal Society, B* 373: 20170382.

Saxe, Andrew, Stephanie Nelli, and Christopher Summerfield. 2020. "If Deep Learning Is the Answer, What Is the Question?" *Nature Reviews Neuroscience* 22: 55–67.

Saxena, Shreya, and John P Cunningham. 2019. "Towards the Neural Population Doctrine." *Current Opinion in Neurobiology* 55: 103–111.

Schaffer, Simon. 1994. "Babbage's Intelligence: Calculating Engines and the Factory System." *Critical Inquiry* 21 (1): 203–227.

Schaeffer, Rylan, Brando Miranda, and Sanmi Koyejo. 2023. "Are Emergent Abilities of Large Language Models a Mirage?" https://arxiv.org/abs/2304.15004.

Schneider, Susan. 2019. *Artificial You*. Princeton, NJ: Princeton University Press.

Schoonover, Carl E., Sarah N. Ohashi, Richard Axel, and Andrew J. P. Fink. 2021. "Representational Drift in Primary Olfactory Cortex." *Nature* 594: 541–546.

Schrimpf, Martin, Jonas Kubilius, Ha Hong, et al. 2020. "Brain-Score: Which Artificial Neural Network for Object Recognition Is Most Brain-Like?" https://www.biorxiv.org/content/10.1101/407007v2.

Schuhl, Pierre-Maxime. 1938. *Machinisme et Philosophie*. Paris: Librairie Félix Alcan.

Scott, Stephen H. 2008. "Inconvenient Truths about Neural Processing in Primary Motor Cortex." *Journal of Physiology* 586 (5): 1217–1224.

Searle, John. 1980. "Minds, Brains, and Programs." *Behavioral and Brain Sciences* 3: 417–457.

Searle, John. 1992. *The Rediscovery of the Mind*. Cambridge, MA: MIT Press.

Sebestik, Jan. 2011. "Otto Neurath's Epistemology and Its Paradoxes." In *Otto Neurath and the Unity of Science*, edited by John Symons, Olga Pombo and Juan Manuel Torres, 41–57. Dordrecht, Netherlands: Springer.

Secord, James A. 1981. "Nature's Fancy: Charles Darwin and the Breeding of Pigeons." *Isis* 72 (2): 162–186.

Sellars, Wilfrid. 1956. "Empiricism and the Philosophy of Mind." In *Minnesota Studies in the Philosophy of Science, vol. I,* edited by H. Feigl and M. Scriven, 253–329. Minneapolis: University of Minnesota Press.

Serre, Thomas. 2019. "Deep Learning: The Good, the Bad, and the Ugly." *Annual Review of Vision Science* 5: 399–426.

Shagrir, Oron. 2010. "Brains as Analog-Model Computers." *Studies in History and Philosophy of Science* 41: 271–279.

Shannon, Claude E. 1951. "Prediction and Entropy of Printed English." *Bell System Technical Journal* 30: 50–64.

Shea, Nicholas. 2018. *Representation in Cognitive Science*. Oxford: Oxford University Press.

Shenoy, K. V. 2015. "Recording from Many Neurons Simultaneously." In *The Future of the Brain*, edited by Gary Marcus and Jeremy Freeman, 78–89. Princeton, NJ: Princeton University Press.

Shenoy, K. V., M. Sahani, and M. M. Churchland. 2013. "Cortical Control of Arm Movements: A Dynamical Systems Perspective." *Annual Review of Neuroscience* 36: 337–359.

Shepherd, Gordon M. 1991. *Foundations of the Neuron Doctrine*. Oxford: Oxford University Press.

Sherrington, C. S. 1906a. *The Integrative Action of the Nervous System*. New York: Charles Scribner's Sons.

Sherrington, C. S. 1906b. "Observations on the Scratch-Reflex in the Spinal Dog." *Journal of Physiology* 34 (1–2): 1–50.

Sherrington, C. S. 1909. "A Mammalian Spinal Preparation." *Journal of Physiology* 38 (5): 375–383.

Shmueli, G. 2010. "To Explain or to Predict?" *Statistical Science* 25: 289–310.

Silberstein, Michael. 2021. "Constraints on Localization and Decomposition as Explanatory Strategies in the Biological Sciences 2.0." In Neural Mechanisms: New Challenges in the Philosophy of Neuroscience, edited by Fabrizio Calzavarini and Marco Viola. Berlin: Springer.

Silver, David, Julian Schrittwieser, Karen Simonyan, et al. 2017. "Mastering the Game of Go without Human Knowledge." *Nature* 550: 354–359.

Simmons, Alison. 2011. "Re-humanizing Descartes." *Philosophic Exchange* 41 (1): 53–67.

Simon, Herbert. 1962. "The Architecture of Complexity." *Proceedings of the American Philosophical Society* 106 (6): 467–482.

Simon, Herbert. 1969. *The Sciences of the Artificial*. Cambridge, MA: MIT Press.

Sinz, Fabian H., Xaq Pitkow, Jacob Reimer, Matthias Bethge, and Andreas S. Tolias. 2019. "Engineering a Less Artificial Intelligence." *Neuron* 103: 967–979.

Skidelsky, E. 2003. "From Epistemology to Cultural Criticism: Georg Simmel and Ernst Cassirer." *History of European Ideas* 29: 365–381.

Skinner, B. F. 1938. *The Behavior of Organisms: An Experimental Analysis*. New York: D. Appleton-Century Company.

Skinner, B. F. 1940. "Review of The Organism by Kurt Goldstein." *Journal of Abnormal and Social Psychology* 35 (3): 462–465.

Skinner, B. F. 1961. "The Concept of the Reflex in the Description of Behavior." In *Cumulative Record*, edited by B. F. Skinner, 319–346. East Norwalk, CT: Appleton-Century-Crofts.

Originally published in 1931.

Skinner, B. F. 1961. *Cumulative Record*. New York: Appleton-Century-Crofts.

Skinner, B. F. 1961. "Current Trends in Experimental Psychology." In *Cumulative Record*, edited by B. F. Skinner, 223–241. East Norwalk, CT: Appleton-Century-Crofts.

Originally published in 1947.

Smart, J. J. C. 1959. "Sensations and Brain Processes." *Philosophical Review* 68 (2): 141–156.

Smetters, D. K., A. Majewska, and Rafael Yuste. 1999. "Detecting Actionpotentials in Neuronal Populations With Calcium Imaging." *Methods* 18: 215–221.

Smith, Justin E. H. 2011. *Divine Machines: Leibniz and the Sciences of Life*. Princeton, NJ: Princeton University Press.

Smith, Laurence D. 1995. "Inquiry Nearer the Source: Bacon, Mach, and The Behavior of Organisms." In *Modern Perspectives on B. F. Skinner and Contemporary Behaviorism*, edited by J. T. Todd and E. K. Morris, 39–50. Westport, CT: Greenwood Press.

Smith, Laurence D. 1996. "Knowledge as Power: The Baconian Roots of Skinner's Social Meliorism." In *B. F. Skinner and Behaviorism in American Culture*, edited by Laurence D. Smith and William Ray Woodward, 56–82. Bethlehem, PA: Lehigh University Press.

Smith, Roger. 1973. "The Background of Physiological Psychology in Natural Philosophy." *History of Science* 11: 75–123.

Smith, Roger. 1992. *Inhibition: History and Meaning in the Sciences of Mind and Brain*. Berkeley: University of California Press.

Smyth, Darragh, Ben Willmore, Gary E. Baker, I. D. Thompson, and D. J. Tolhurst. 2003. "The Receptive-Field Organization of Simple Cells in Primary Visual Cortex of Ferrets under Natural Scene Stimulation." *Journal of Neuroscience* 23 (11): 4746–4759.

Snir, I. 2020. *Education and Thinking in Continental Philosophy, Contemporary Philosophies and Theories in Education*. Cham: Switzerland: Springer Nature.

Snyder, Jason. 2019. "Recalibrating the Relevance of Adult Neurogenesis." *Trends in Neurosciences* 42 (3): 164–178.

So, K., A. C. Koralek, K. Ganguly, M. C. Gastpar, and J. M. Carmena. 2012. "Assessing Functional Connectivity of Neural Ensembles Using Directed Information." *Journal of Neural Engineering* 9 (2):026004. doi: 10.1088/1741–2560/9/2/026004.

Sonkusare, Saurabh, Michael Breakspear, and Christine Guo. 2019. "Naturalistic Stimuli in Neuroscience: Critically Acclaimed." *Trends in Cognitive Sciences* 23 (8): 699–714.

Soteriou, Matthew. 2016. *Disjunctivism*. London: Routledge.

Soteriou, Matthew. 2020. "The Disjunctive Theory of Perception." In *Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/archives/sum2020/entries/perception-disjunctive/.

Spade, Paul Vincent. 1999. "Ockham's Nominalist Metaphysics: Some Main Themes." In *Cambridge Companion to Ockham*, edited by Paul Vincent Spade, 100–117. Cambridge: Cambridge University Press.

Sprevak, Mark. Unpublished manuscript. "Physical Computations Are Idealisations."

Sprevak, Mark. 2013. "Fictionalism about Neural Representations." *The Monist* 96: 539–560.

Sprevak, Mark. 2018. "Triviality Arguments about Computational Implementation." In *Routledge Handbook of the Computational Mind*, edited by Mark Sprevak and Matteo Colombo, 175–191. London: Routledge.

Sprevak, Mark. 2019. "Review of The Language of Thought: A New Philosophical Direction by Susan Schneider." *Mind* 128: 555–564.

Sprevak, Mark. 2021 draft. "Predictive Coding IV: The Implementation Level." https://marksprevak.com/publications/predictive-coding-iv-the-implementation-level-cfb2/

Sprevak, Mark, and Matteo Colombo. 2019. "Introduction." In *Routledge Handbook of the Computational Mind*, edited by Mark Sprevak and Matteo Colombo, 1–6. Abingdon, UK: Routledge.

Srinivasan, M., S. Laughlin, and A. Dubs. 1982. "Predictive Coding: A Fresh View of Inhibition in the Retina." *Proceedings of the Royal Society of London B: Biological Sciences* 216: 427–459.

Stadler, Friedrich. 2021. "Ernst Mach and the Vienna Circle: A Re-evaluation of the Reception and Influence of His Work." In *Interpreting Mach: Critical Essays*, edited by John Preston, 184–207. Cambridge: Cambridge University Press.

Standing, L. 1973. "Learning 10,000 Pictures." *Quarterly Journal of Experimental Psychology* 25: 207–222.

Stein, Howard. 1989. "Yes, but . . . Some Skeptical Remarks on Realism and Anti-Realism." Dialectica. 43(1-2): 47-65.

Steinmetz, Nicholas A., Christof Koch, Kenneth D. Harris, and Matteo Carandini. 2018. "Challenges and Opportunities for Large-Scale Electrophysiology with Neuropixels Probes." *Current Opinion in Neurobiology* 50: 92–100.

Sterling, P., and S. Laughlin. 2015. *Principles of Neural Design*. Cambridge, MA: MIT Press.

Stevenson, Ian H., and Konrad P. Kording. 2011. "How Advances in Neural Recording Affect Data Analysis." *Nature Neuroscience* 14 (2): 139–142.

Stosiek, Christoph, Olga Garaschuk, Knut Holthoff, and Arthur Konnerth. 2003. "In Vivo Two-Photon Calcium Imaging of Neuronal Networks." *Proceedings of the National Academy of Sciences* 100 (12): 7319–7324.

Stringer, Carsen, Marius Pachitariu, Nicholas Steinmetz, Charu Bai Reddy, Matteo Carandini, and Kenneth D. Harris. 2019. "Spontaneous Behaviors Drive Multidimensional, Brainwide Activity." *Science* 364: eaav7893.

Sullivan, Emily. 2022. "Understanding from Machine Learning Models." *British Journal for the Philosophy of Science* 73: 109–133. doi: 10.1093/bjps/axz035.

Sullivan, Jacqueline A. 2009. "The Multiplicity of Experimental Protocols: A Challenge to Reductionist and Non-reductionist Models of the Unity of Neuroscience." *Synthese* 167: 511–539.

Sussillo, David, and Omri Barak. 2013. "Opening the Black Box: Low-Dimensional Dynamics in High-Dimensional Recurrent Neural Networks." *Neural Computation* 25: 626–649.

Suvrathan, Aparna, and Jennifer L. Raymond. 2018. "Depressed by Learning—Heterogeneity of the Plasticity Rules at Parallel Fiber Synapses onto Purkinje Cells." *The Cerebellum* 17: 747–755.

Taylor, Samuel D. 2021. "Causation and Cognition: An Epistemic Approach." *Synthese* 199: 9133–9160. https://doi.org/10.1007/s11229-021-03197-2.

Teller, Paul. 2021. "Making Worlds with Symbols." *Synthese* 198(Suppl 21): 5009–5013. doi: 10.1007/s11229-018-1811-y.

Titchener, E. B. 1914. "On "Psychology as the Behaviorist Views It"". *Proceedings of the American Philosophical Society*. 53(213): 1–17

Thompson, Evan. 2007. *Mind in Life: Biology, Phenomenology, and the Sciences of Mind.* Cambridge, MA: Belknap Press of Harvard University Press.

Thompson, Evan. 2009. "Life and Mind: From Autopoiesis to Neurophenomenology." In *Emergence and Embodiment: New Essays on Second-Order Systems Theory*, edited by Bruce Clarke and Mark B. N. Hansen. Durham, NC: Duke University Press.

Thomson, Eric, and Gualtiero Piccinini. 2018. "Neural Representations Observed." *Minds and Machines* 28: 191–235.

Todes, Daniel P. 2014. *Ivan Pavlov: A Russian Life in Science.* Oxford: Oxford University Press.

Tolhurst, David J., Michelle P. S. To, Mazviita Chirimuuta, Tom Troscianko, Pei-YingChua, and P. George Lovell. 2010. "Magnitude of Perceived Change in Natural Images May Be Linearly Proportional to Differences in Neuronal Firing Rates." *Seeing and Perceiving* 23: 349–372.

Tolman, Edward. C. 1938. "The Determiners of Behavior at a Choice Point." *Psychological Review* 45: 1–41.

Tong, Frank, Ken Nakayama, Morris Moscovitch, Oren Weinrib, and Nancy Kanwisher. 2000. "Response Properties of the Human Fusiform Face Area." *Cognitive Neuropsychology* 17: 257–279.

Trachtenberg, Joshua, Brian E. Chen, Graham W. Knott, et al. 2002. "Long-Term in Vivo Imaging of Experience-Dependent Synaptic Plasticity in Adult Cortex." *Nature* 420: 788–794.

Turing, Alan. c.1950. *Programmers' Handbook for Manchester Electronic Computer Mark II*, Computing Machine Laboratory, University of Manchester. https://www.alanturing.net/programmers_handbook/.

Turing, Alan. 2005. "Proposed Electronic Calculator." In *Alan Turing's Automatic Computing Engine*, edited by B. Jack Copeland, 369–454. Oxford: Oxford University Press.

Originally published in 1945.

Turner, Alexander Matt, Logan Smith, Rohin Shah, Andrew Critch, and Prasad Tadepalli. 2021. "Optimal Policies Tend To Seek Power." *35th Conference on Neural Information Processing Systems (NeurIPS 2021)*. https://arxiv.org/pdf/1912.01683.pdf.

Vallor, Shannon. 2021. "The Thoughts the Civilized Keep." *Noema*. https://www.noemamag.com/the-thoughts-the-civilized-keep/

van den Belt, Henk 2009. "Playing God in Frankenstein's Footsteps: Synthetic Biology and the Meaning of Life." *Nanoethics* 3: 257–268.

van de Ven, Gido M., Hava T. Siegelmann, and Andreas S. Tolias. 2020. "Brain-Inspired Replay for Continual Learning with Artificial Neural Networks." *Nature Communications* 11: 4069.

Van Essen, David C., and Matthew F. Glasser. 2018. "Parcellating Cerebral Cortex: How Invasive Animal Studies Inform Noninvasive Mapmaking in Humans." *Neuron* 99: 640–663.

van Gelder, T. 1995. "What Might Cognition Be, If Not Computation?" *Journal of Psychology* 92 (7): 345–381.

VanRullen, R. 2017. "Perception Science in the Age of Deep Neural Networks." *Frontiers in Psychology* 8: 142. doi: 10.3389/fpsyg.2017.00142.

Varela, Francisco J., Evan Thompson, and Eleanor Rosch. 2016. *The Embodied Mind*. Cambridge, MA: MIT Press.

Originally published in 1991.

Vassányi, Miklós. 2011. *Anima Mundi: The Rise of the World Soul Theory in Modern German Philosophy*. Dordrecht, Netherlands: Springer.

Velliste, Meel, Sagi Perel, M. Chance Spalding, Andrew S. Whitford, and Andrew B. Schwartz. 2008. "Cortical Control of a Prosthetic Arm for Self-Feeding." *Nature* 453: 1098–1101.

Verkhratsky, Alexei, Margaret S. Ho, Robert Zorec, and Vladimir Parpura. 2019. "The Concept of Neuroglia." In *Neuroglia in Neurodegenerative Diseases*, edited by Alexei Verkhratsky, Margaret S. Ho, Robert Zorec and Vladimir Parpura, 1–13. Singapore: Springer Nature.

Vico, Giambattista. 1988. *On the Most Ancient Wisdom of the Italians Unearthed from the Origins of the Latin Language*. Translated by L. M. Palmer. Ithaca, NY: Cornell University Press.

Originally published in 1710.

Vilarroya, Oscar. 2017. "Neural Representation. A Survey-Based Analysis of the Notion." *Frontiers in Psychology* 8: 1458. doi: 10.3389/fpsyg.2017.01458.

Vintch, Brett, J. A. Movshon, and E. P. Simoncelli. 2015. "A Convolutional Subunit Model for Neuronal Responses in Macaque V1." *Journal of Neuroscience* 35 (44): 14829–14841.

von Neumann, John, and Arthur W. Burks. 1966. *Theory of Self-Reproducing Automata*. Urbana: University of Illinois Press.

Walker, Edgar Y., Fabian H. Sinz, Erick Cobos, et al. 2019. "Inception Loops Discover What Excites Neurons Most Using Deep Predictive Models." *Nature Neuroscience* 22: 2060–2065.

Walsh, Denis. 2015. *Organisms, Agency, and Evolution*. Cambridge: Cambridge University Press.

Walshe, Francis M. R. 1961. "Contributions of John Hughlings Jackson to Neurology." *Archives of Neurology* 5: 119–131.

Walter, W. Grey. 1953. *The Living Brain.* New York: W. W. Norton.

Ward, Dave, and Mog Stapleton. 2012. "Es Are Good. Cognition as Enacted, Embodied, Embedded, Affective and Extended." In *Consciousness in Interaction: The Role of the Natural and Social Context in Shaping Consciousness*, edited by Fabio Paglieri, 89–104. Amsterdam: John Benjamins.

Warren, Daniel. 2001. "Kant's Dynamics." In *Kant and the Sciences*, edited by Eric Watkins, 93–116. Oxford: Oxford University Press.

Watson, John B. 1913. "Psychology as the Behaviorist Views it." *Psychological Review* 20 (2): 158–177.

Weinberger, Naftali, and Colin Allen. 2022. "Static-Dynamic Hybridity in Dynamical Models of Cognition." *Philosophy of Science* 89 (2): 283–301. doi: https://doi.org/10.1017/psa.2021.27.

Weisberg, Michael. 2007. "Three Kinds of Idealization." *Journal of Philosophy*, 104 (12): 639–659.

Weisberg, Michael. 2013. *Simulation and Similarity: Using Models to Understand the World*. Oxford: Oxford University Press.

Westfall, Richard S. 1981. *Never at Rest: A Biography of Isaac Newton*. Cambridge: Cambridge University Press.

Whitehead, A. N. 1948. *An Introduction to Mathematics*. London: Oxford University Press.

Originally published in 1911.

Whitehead, A. N. 1967. *Science and the Modern World*. New York: Free Press.

Originally published in 1925.

Whiteway, Matthew R., and Daniel A. Butts. 2019. "The Quest for Interpretable Models of Neural Population Activity." *Current Opinion in Neurobiology* 58: 86–93.

Whiting, Jennifer. 1992. "Living Bodies." In *Essays on Aristotle's de Anima*, edited by Martha C. Nussbaum and Amélie Oksenberg Rorty, 75–92. Oxford: Oxford University Press.

Wiggers, Kyle. 2018. "Geoffrey Hinton and Demis Hassabis: AGI Is Nowhere Close to Being a Reality." *VentureBeat*, December 17.

Williamson, Jon. 2004. *Bayesian Nets and Causality: Philosophical and Computational Foundations*. Oxford: Oxford University Press.

Wilson, Jessica. 2006. "On Characterizing the Physical." *Philosophical Studies* 131: 61–99.

Wimsatt, William C. 2007. *Re-engineering Philosophy for Limited Beings*. Cambridge, MA: Harvard University Press.

Wolfe, Charles T., and Christopher Donohue, eds. 2023. *Vitalism in the 20th Century, History, Philosophy and Theory of the Life Sciences*. Cham, Switzerland: Springer.

Woodward, James F. 2003. *Making Things Happen*. Oxford: Oxford University Press.

Woodward, James F. 2007. "Causation with a Human Face." In *Causation, Physics, and the Constitution of Reality: Russell's Republic Revisited*, edited by Huw Price and Richard Corry, 66–105. Oxford: Oxford University Press.

Woodward, J. F. 2014. "A Functional Account of Causation; or, A Defense of the Legitimacy of Causal Thinking by Reference to the Only Standard That Matters—Usefulness (as Opposed to Metaphysics or Agreement with Intuitive Judgment)." *Philosophy of Science* 81 (5): 691–713.

Yamins, Daniel L. K., and J. J. DiCarlo. 2016. "Using Goal-Driven Deep Learning Models to Understand Sensory Cortex." *Nature Neuroscience* 19 (3): 356–365.

Yamins, Daniel L. K., H. Hong, C. F. Cadieu, E. A. Solomon, D. Seibert, and J. J. DiCarlo. 2014. "Performance-Optimized Hierarchical Models Predict Neural Responses in Higher Visual Cortex." *PNAS* 111 (23): 8619–8624.

Yarkoni, Tal, and Jacob Westfall. 2017. "Choosing Prediction over Explanation in Psychology: Lessons from Machine Learning." *Perspectives on Psychological Science* 12 (6): 1100–1122.

Yong, Ed. June 2021. "Neuroscientists Have Discovered a Phenomenon That They Can't Explain." *The Atlantic*. https://www.theatlantic.com/science/archive/2021/06/the-brain-isnt-supposed-to-change-this-much/619145/.

Young, Robert. 1990. *Mind, Brain, and Adaptation in the Nineteenth Century*. Oxford: Oxford University Press.

Yuille, Alan L., and Chenxi Liu. 2021. "Deep Nets: What Have They Ever Done for Vision?" *International Journal of Computer Vision* 129: 781–802.

Yuste, Rafael. 2015. "From the Neuron Doctrine to Neural Networks." *Nature Reviews Neuroscience* 16: 487–497.

Yuste, Rafael, and George Church, M. 2014. "The New Century of the Brain." *Scientific American,* March, 38–45.

Zammito, John H. 2018. *The Gestation of German Biology*. Chicago: University of Chicago Press.

Zeng, Hongkui, and Joshua R. Sanes. 2017. "Neuronal Cell-Type Classification: Challenges, Opportunities and the Path Forward." *Nature Reviews Neuroscience* 18: 530–546.

Zilsel, Edgar. 1942. "The Sociological Roots of Science." *American Journal of Sociology* 47 (4): 544–562.

Ziv, Y., L. D. Burns, E. D. Cocker, et al. 2013. "Long-Term Dynamics of CA1 Hippocampal Place Codes." *Nature Neuroscience* 16: 264.

Zuckerman, S. 1950. "The Mechanism of Thought: The Mind and the Calculating Machine." In *The Physical Basis of Mind,* edited by Peter Laslett, 25–35. Oxford, UK: Basil Blackwell.

# Index