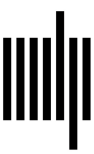
Gradient Expectations

Structure, Origins, and Synthesis of Predictive Neural Networks

By: Keith L. Downing

Citation:

Gradient Expectations: Structure, Origins, and Synthesis of Predictive Neural Networks
By: Keith L. Downing
DOI: 10.7551/mitpress/14723.001.0001
ISBN (electronic): 9780262374675
Publisher: The MIT Press
Published: 2023



The MIT Press

1 Introduction

It's tough to make predictions, especially about the future.
—Yogi Berra (famous American baseball player and coach)

Aside from the classic prerequisites to evolutionary success—survival and fecundity—the ability to predict clearly tips the Darwinian scales like few other cognitive faculties. Having just an inkling of what lies around the bend, behind the bush, or over the horizon can spell the difference between feast and famine, pleasure and pain, life and death—or, in Yogi Berra's world, strikeout and round-tripper.[1] People who know what lies ahead can amass fame, fortune, and a long line of eager followers and envious competitors.

But despite those common shortcomings that render the bulk of us losers in Las Vegas casinos, suckers for poor investments, and benchwarmers in baseball, most of us possess the predictive apparatus that aided man's ascent to the top of the food chain; and that, if nothing more spectacular or profitable, does help us snake our way to a speedy checkout at Food Lion. We may not know whether pork bellies will trade higher or lower tomorrow, but we can easily surmise that the three teenagers, each with a single bag of pork rinds at register 3, will file out long before the guy with a full cart and a screaming baby in the express line, or the partygoer in line 7 with only a few items but no visible means of differentiating a debit card from a driver's license. When paying attention, we make some quick, predictive calculations, jump to aisle 3, and never look back. We hardly even recognize that we've done any complex thinking. After all, it's not rocket science . . . unfortunately.

In 1969, when Neil Armstrong took his giant leap for mankind, would anyone have ventured that, a half century later, the final frontier would be between our own ears? Intelligence exemplifies Churchill's "riddle, wrapped in a mystery, inside an enigma," and somewhere buried deep in that tangled mess lies prediction, not as a disjoint, free-floating feature but as the featured attraction.

In the past few decades, many prominent cognitive scientists (Llinas 2001; Hawkins 2004; Clark 2016; Buzsaki 2006) have begun to hail prediction as the hallmark of intelligence. For example, Buzsaki sets a predictive tone with the first sentence of *Rhythms of the Brain* (Buzsaki 2006, vii): "The short punch line of this book is that brains are foretelling devices, and their predictive powers emerge from the various rhythms they perpetually generate."

Equally convinced is Llinas, who writes, in *i of the Vortex* (Llinas 2001, 21), "The capacity to predict the outcome of future events—critical to successful movement—is likely, the ultimate and most common of all global brain functions."

Complementing those two prominent neuroscientists, the cognitive scientist and philosopher Andy Clark prefaces his popular *Surfing Uncertainty* (Clark 2016, xiv) with his usual elegance:

> The mystery is, and remains, how mere matter manages to give rise to thinking, imagining, dreaming, and the whole smorgasbord of mentality, emotion and intelligent action. . . . But there is an emerging clue. . . . The clue can be summed up in a single word: prediction. To deal rapidly and fluently with an uncertain and noisy world, brains like ours have become masters of prediction—surfing the waves of noisy and ambiguous sensory stimulation by, in effect, trying to stay just ahead of them.

No less emphatic is tech entrepreneur and cognitive scientist Jeff Hawkins, who writes, in *On Intelligence* (Hawkins 2004, 89), "The cortex is an organ of prediction. If we want to understand what intelligence is, what creativity is, how your brain works, and how to build intelligent machines, we must understand the nature of these predictions and how the cortex makes them. Even behavior is best understood as a by-product of prediction."

Although the advantages of explicit predictive skills in everyday life seem obvious, the proposals by the above scientists are much more radical, as they argue for prediction's centrality in the workings of the brain. This book investigates that claim by examining a host of neural networks, both natural and artificial, with a special focus on the internal flow of signals that seems to embody expectations.

1.1 Data from Predictions

From a practical machine learning (ML) perspective, the ability to predict provides an invaluable service to *data-hungry*, supervised-learning algorithms, such as most conventional neural networks. The well-known fuel of these algorithms is data—as in the popular phrase, *data is the new oil*. For a supervised-learning system, data constitutes pairs of input patterns and their corresponding target output patterns (often called *labels*). For example, the data pairs for a facial recognition system consist of pixel images labeled with names, social-security codes, or other unique identifiers. In today's online world, there is no shortage of images, text, and other unstructured data, but labeling often requires time-consuming human analysis.

This same human bottleneck applies to supervised sequence-completion tasks, in which a system receives several words, phonemes, images, or other unstructured elements of a sequence and must *predict* the next item. In these cases, the next item constitutes the target, and when these targets are full images, acoustic patterns, and the like, the labeling chore for humans becomes all the more arduous.

The beauty of predictive algorithms is that if they operate somewhat autonomously in a *situated setting* (i.e., they have direct access to a physical or virtual environment in which they can move about and sense their surroundings), then they generate their own targets by merely *waiting one timestep*. Any prediction, made at time T, of the environmental state at time T+1, will receive its target when the agent observes the world at time T+1. Thus, the data item is the state at time T (plus possibly other states prior to T) paired with the (target) state at T+1.

By moving around, observing, and predicting, the agent generates its own data set, which it can then use in a supervised-learning fashion to improve its own predictive abilities. And improving one's competence at predicting future states of the world often equates with building an accurate model of both the world and the effects of one's actions on the world, which are, in turn, two core aspects of intelligence.

1.2 Movement and Prediction

The most convincing argument for prediction as the primary prerequisite to intelligence revolves around movement and its pivotal contribution to survival. Consider the agent of figure 1.1 and its ability to respond to environmental change. In this diagram, the time required to sense and then act generally undercuts the duration of an environmental state, i.e., the environment's timescale (τ_e) exceeds that of the agent (τ_a). The sequential processes of reading world state X and responding to it fall within X's time window, thus making the response appropriate.

Conversely, when the agent operates at a lower frequency than the environment (as in figure 1.2), the delay between sensing and acting causes obvious problems: the agent senses state X and responds to it, but only after the state has changed to Y, yielding the response inappropriate.

Faced with an environment that changes faster than the latencies of sensing and acting allow it to respond ($\tau_a > \tau_e$), the (perpetually confused) agent's prospects for a long and happy life seem grim.[2]

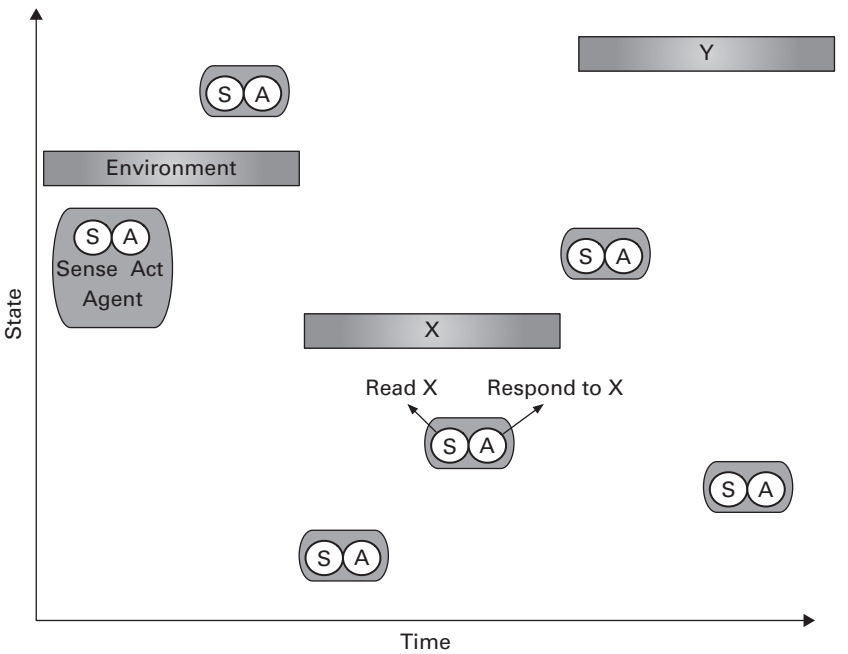


Figure 1.1
Comparing the natural timescale of a sensing-and-acting agent (τ_a) to that of its environment (τ_e) when $\tau_a < \tau_e$. Horizontal bars denote relatively stable environmental states, while the agent's vertical position denotes its own state. Larger time constants entail longer stable states.

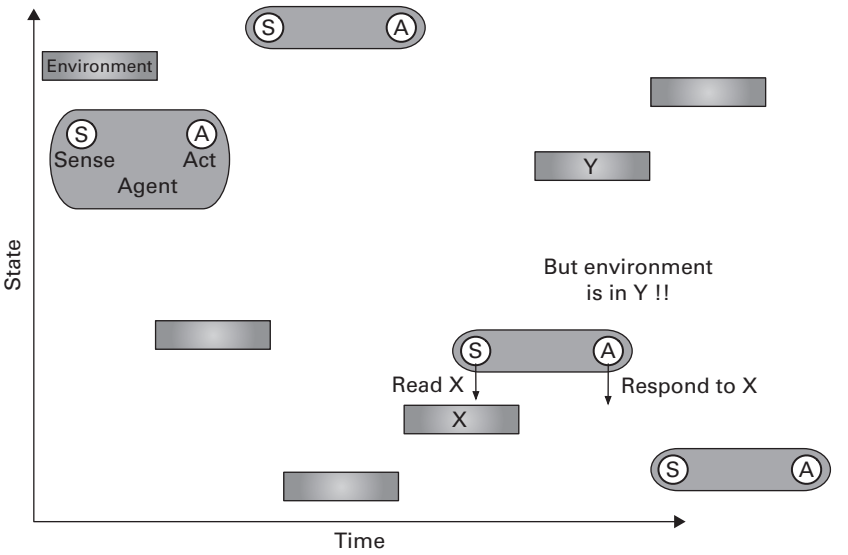


Figure 1.2
Comparing the natural timescale of a sensing-and-acting agent (τ_a) to that of its environment (τ_e) when $\tau_a > \tau_e$.

Prediction provides the perfect antidote to this temporal mismatch and indecision. As shown in figure 1.3, a predictive agent can sense state X, use its model of the world to predict that X leads to Y*, and then respond to Y*. As the figure indicates, Y* only approximates Y, but anything close to Y supports better preparation than the assumption that X will persist. It is probably better for a gazelle to run away from a rustling bush with lingering doubt as to the exact angle from which the tiger will pounce than to ignore the possibility of an abrupt and violent change of state.

Interestingly enough, the gazelle's innate speed compounds its predictive challenges. The faster an agent moves, the more quickly its environment changes. This motion-induced environmental timescale (τ_m) essentially supersedes τ_e in estimating the agent's evolutionary fitness. The faster it moves, the smaller τ_m becomes, and thus the greater the need for low sense-and-act latency and/or accurate prediction. Basic properties of biochemical signaling and neural circuitry place strict lower bounds on latency (τ_a) such that the only feasible solution for almost all mobile organisms involves predictive mechanisms. As Llinas (2001) argues, these predictive abilities are the brain's main function, and to such a convincing degree that primitive sessile organisms need no brain at all. He uses the (now popular) example of the sea squirt, which begins life as a free-swimming larva before permanently attaching itself to a fixed location and digesting its own brain, which would apparently serve as nothing more than an energy sink during its adult stage.

Llinas extends his argument to frame cognition as internalized motion control. Early in embryonic development (as well as in the mature stages of very primitive organisms), the activity of muscles is controlled locally, in a very emergent (but limited) manner: active muscles stimulate neighboring muscles, often in a rhythmic manner, which serves many useful purposes in both movement and digestion. As motor neurons arise and their axons migrate and connect to muscles, that control moves upward in the neural hierarchy. Two key

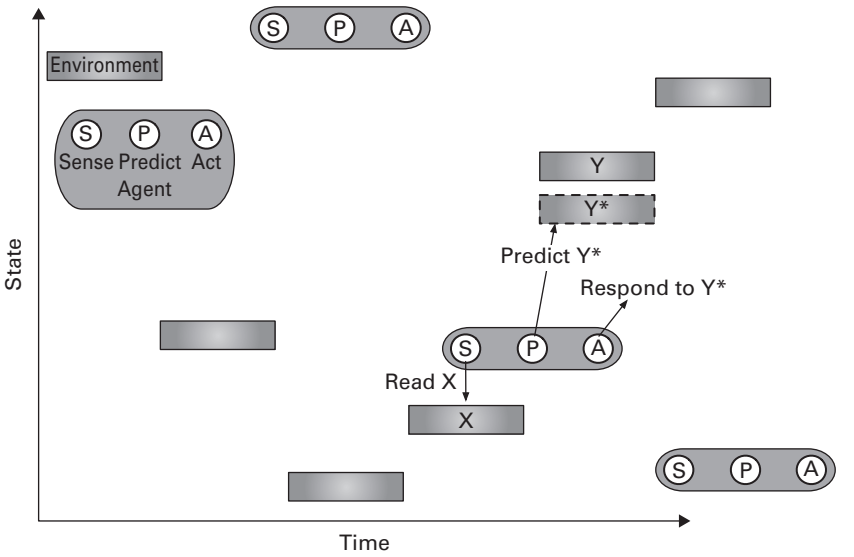


Figure 1.3
Adding prediction to the agent's cognitive repertoire and thus combating problems associated with timescale discrepancies: $\tau_a > \tau_e$.

advantages of this *encephalization* are (1) coordination of more intricate muscle activity by neurons that are both intraconnected in complex patterns (more than are the muscles themselves) and linked to multiple muscles, and (2) creation of avenues for integrating diverse sensory inputs from throughout the body into the decision making of the premotor neurons and higher-level circuitry. The price paid for this complexification is latency, which rises by tens to hundreds of milliseconds per neural layer. Hence, the need for sophisticated sensing becomes acute, as a slow decision process needs a more accurate *picture of the world* and, even more important, a model of the agent-world coupling, in order to make reasonably accurate predictions of future states and to coordinate a platoon of muscles to operate in an elaborate environment. The complexities of sensing, acting, and predicting must increase in lockstep. Thus, to the extent that evolution bootstrapped motion control up the neural hierarchy, prediction needed to join the ascent.

1.3 Adaptation and Emergence

Adaptation is a system's ability to change in response to (or anticipation of) changes in its environment. The above gazelle example fits this definition, as does the ability of herbivores to evolve faster reaction times in their arms race against carnivores on the African savanna. Adaptation spans many timescales.

In this book, the term will be restricted to internal *structural* change of some significant duration (typically in a neural network). So the change in a gazelle's speed and direction at any given moment will not fall under this more-restrictive definition, but a change in neural synapses (their number and/or strength) to modify the animal's conception of *dangerous bush-stirring sounds* would qualify as a relatively short-term adaptation.

Carving the timescales into three general pieces (short, medium, long) results in three standard classes of adaptation: learning, development, and evolution, with some overlap among them, especially the former two. However, given our focus on the brain and other (artificial) neural networks, a reasonable working distinction is possible: learning involves changes to existing interneural connection strengths, while development involves the formation of entire network topologies, including the formation of connections. Granted, in real brains, new synapses grow and die throughout life (as part of learning, disuse, and so on), but the net result is the change in one neuron's effect on another. Those changes of influence versus changes in topology will be the main distinction between learning and development in these chapters.

Evolutionary change is easier to distinguish, as it requires inheritable genomic modifications (for the most part): structural change is to the genetic material that gets passed on to the next generation. An evolutionary change is thereby recognized as a structural difference between an organism and its descendants.

Although changes to the immediate firing levels of individual neurons typically reflect responses to those of other neurons or to environmental stimuli, that timescale of adaptation will typically not fall under our working definition, since the normal changes inherent in neuronal firing will not be considered structural.

The processes directing adaptive change have great significance in this book. Those involving a central control algorithm that analyzes many or all components before adjusting each such unit are of only peripheral interest. My focus is on emergent systems in which local activity leads to global patterns, which may eventually exert some influence on the local behavior. But those global influences are not hardwired into the system; they must arise from the local dynamics.

My earlier book, *Intelligence Emerging* (Downing 2015), delves deeply into emergent mechanisms underlying learning, development, and evolution with respect to natural and artificial neural networks. In that work, I posit that each emergent adaptive process in nature appears to be driven by relatively random trial-and-error search for appropriate synaptic strengths, stable and efficient network topologies, and high-fitness genomes (which are genomes that, among other things, encode good recipes for development and learning). Crucially, the trial-and-error processes at the slower timescales produce landscapes for the search performed by the faster adaptive processes. Even the finely tuned (by learning) network of synapses sculpts basins of attraction that strongly bias the dynamics of reasoning.

Although *Intelligence Emerging* includes some limited material on predictive mechanisms in the brain, the focus is on the trial-and-error processes by which neural motifs may learn to encode predictions. Explanations of how prediction *grounds out* in stochastic search removes any serious contradictions between that book and this one, but the tension between those two underlying mechanisms of intelligence deserves careful scrutiny.

When I have expectations about possible futures and leverage them to choose actions, clearly I am not following a pure trial-and-error process. The (albeit uncertain) lookahead governs actions that are far from random. Information about how tweaking parameter A will affect component B gives any agent an advantage over those who will randomly change any parameter in the quest for improvement.

Clearly, the emergence of intelligence, from bacteria to humans, cannot be explained by purely random activity, but at each temporal scale, processes governed by seemingly

arbitrary *choices* play a strong role. At the evolutionary timescale, the unpredictable (though not purely random) mutations and recombinations of genetic material produce new genotypes, which are then subjected to the decidedly nonarbitrary forces of natural selection. Over the generations, the genetic material figurately *learns* the effective adaptations for a given environment. In infants, random movements dominate as the brain gradually sorts out what works and does not. In general, all species in all phases of life strike a balance between *exploration* and *exploitation*: relatively random activities versus those known to be productive, where the former can be viewed as *reconnaissance* expeditions, with the accrued information used to guide later exploitation.

Intelligence Emerging puts extra emphasis on exploration, due largely to my fascination with the nondeterministic (yet doggedly persistent) nature of so many biological processes. This book digs deeper into exploitation, via prediction. However, emergence remains a dominant force in the chapters that follow.

1.3.1 Gradients and Emergence in Neural Networks

The critical divide between the neural networks covered herein and those that have taken the AI world by storm since the early 2010s is the presence of a global control mechanism, and thus lack of convincing emergence, in those popular neural architectures. The vast majority of successful deep learning (DL) networks employ effective global controls powered by information with extensive spatial scope: causal knowledge of how tweaks to component A will affect the behaviors of other components, some of which may be quite distant from A in structural space, aka *gradients*. This spatial lookahead arises not by trial-and-error experience but by formal mathematical derivations across long causal chains. These provide computational templates into which the data of individual experiences nicely fit, yielding precise numerical gradients and thus well-founded lookaheads for intelligent action selection. For example, the question of how to most judiciously change one of the millions of weights (w) in a deep network so as to reduce the total error (E) on the output end of that network is answered in DL systems using the gradient $\frac{\triangle E}{\triangle w}$, which represents the expected change in E due to a unit change in w. A DL system computes one such gradient for each of the millions of weights. Thus, it amasses detailed information that helps predict how the change in any particular weight will contribute to the ultimate goal of reducing E.

Gradients play a central role in this book as well, but they have a much more local property: they link changes in one variable (e.g., the strength of a synapse) to *nearby* changes, such as to the error or *surprise* recorded by a neuron immediately downstream of that synapse. Other gradients capture differences in the firing rates of neurons across short expanses of space or time, again, local relationships and computations. Another variation of gradient records the changes in the predictions themselves, with larger differences manifesting *surprise* and stimulating learning. In each case, these local gradients and their usage exemplify emergence and often reflect current trends in neuroscience. Conversely, the *long-distance gradients* of contemporary deep learning have a much more tenuous relationship to biology, despite the fundamental biological inspiration of neural networks in general.

As a simple analogy, consider person X running for president of Land-O-Plenty. The ideal scenario for X is to know, for each citizen, c_i, how changes in c_i (with respect to acquired information, income, services, and the like) will ultimately affect the total number of presidential votes that X receives (V_x). That is, X would like to know $\frac{\triangle V_x}{\triangle c_i}$ for each

individual i. This is a huge request, but one that has, for better or worse, become reasonable in the social-media age. X can then target each voter with surgical precision, feeding them just the information (real or fake) that X can predict (with reasonable certainty) will nudge c_i toward voting for X (and broadcasting praise of X to her contacts). In short, X computes $\frac{\triangle V_x}{\triangle c_i}$ for all 350 million citizens, c_i, and then uses that as the basis for 350 million individual predictions as to how c_i will probably vote (and influence other voters) given a particular change $\triangle c_i$ that is tailored specifically for c_i.

Each of these gradients will be very complex, involving detailed reasoning about c_i's age, education, employment, lifestyle, voting history, and so on such that the logical connection between the tailored information and the desired support is long and winding. For example, telling c_i that *my opponent, Y, wants to guarantee funding for a nice new stretch of highway between Barleyburg and Soy City* will resonate poorly with c_i, since she (a) runs a popular general store on a scenic back road that is currently the only connection between Barleyburg and Soy City, (b) has no formal education beyond junior high school, and (c) is approaching retirement age. For the well-funded and tech-savvy political campaign, this type of information may lie within its reach. For the backpropagation algorithm that drives deep learning, it is a prerequisite, and one easily obtained with enough computing power.

In earlier days, before people eagerly divulged troves of personal information in public fora, X would have had to get by with knowledge of other, more general, gradients, such as $\frac{\triangle c_j}{\triangle c_i}$ for *all* friends, that is, the same relationship holds for any pair of friends, i and j. For example, as a heuristic (i.e., rule of thumb), X might assume that if c_i tells friend c_j something that c_i has just learned and now believes, then c_j will believe it too. In short, X understands how local changes in the beliefs of people come about: how information and belief spread *through the grapevine*. X can then plant the seeds for spreading belief by taking out attack advertisements in local newspapers against candidate Y, in the hopes that $\frac{\triangle c_j}{\triangle c_i}$ will kick in numerous times, producing widespread belief in Y's incompetence and evil intentions. Alternatively, X could promise (if elected) to pump government aid into the local economy, hoping that word of this favorable action would quickly reach all citizens of Barleyburg.

Basically, X must rely on global or regional messaging (or money) and a general intuition about local causal relationships (gradients) in the hopes of producing a victorious global outcome: $V_x > V_y$. But X has no *direct, long-distance line* that links c_i to V_x: a single individual to a global outcome. This is politics the old-fashioned way, and learning the biological way.

Brain areas can broadcast global or regional signals (i.e., neuromodulators), but the only pinpoint messaging occurs between a neuron and one of the (possibly ten thousand) others to which it connects. The history of that local signaling plus any nascent or lingering neuromodulator then drive synaptic change, the cornerstone of learning. There exists no brainwide ledger that predicts how any synaptic change will affect overall behavior. Emergence, with all of its uncertainties, is the only known biological route to cognition and survival.

In summary, today's most powerful deep-learning systems rely on long-distance gradients and thus exhibit very little emergence, despite the fact that a researcher can rarely predict what such a network will learn from a bevy of examples. These nets are still surprising and normally frustratingly difficult to explain, but that complexity does not arise from purely local interactions. However, biological intelligence, with its inherent adaptability, does rely on local interactions and emergence across multiple timescales. And as shown

in the chapters that follow, many of these local mechanisms embody prediction when one carefully examines the nature of expectations in a neural system.

1.4 Overflowing Expectations

This book will not help you beat the stock market, prepare for tomorrow's weather, or move more efficiently through mega-store checkout lines. Sorry. The final, overt predictions produced by humans and machines are really only of peripheral relevance to the main story: expectations seems to be omnipresent in the brain, and our AI technologies could benefit by incorporating similar, widespread, local predictive mechanisms into neural networks.

At one time in our distant evolutionary history, the overt behavioral outputs were the main predictive achievement, but as nervous systems arose and evolved, the predictive capabilities ascended and proliferated throughout the neural circuitry. Today, many regions of the nervous systems of numerous animal species are amenable to useful predictive interpretations, often involving the collision of a *reality stream* and a *prediction stream*, the difference of which yields a *prediction error*, which constitutes a modified reality stream that continues upward in the neural hierarchy, while also contributing to a prediction of its own about activity at a lower level.

Under this interpretation, the brain is a flood of expectations and surprises (violated predictions) gushing down and up (respectively) the neural hierarchy. This view of neural processing has been around for more than seventy years and gained significant popularity over the past few decades. The intuitive advantage of such an arrangement seems pretty obvious: if information is *as expected* by some receiving region, then why should the sending region expend the energy to transmit it? When nothing unexpected happens in downtown Soy City, journalists resort to human-interest and nostalgia pieces to fill the local paper. The brain has no problem with slow news periods; it can use the time to consolidate some of its recent experiences while saving up some energy for effectively reporting future surprises.

The chapters that follow spotlight these theorized information pathways, as configured in various ways by psychology, neuroscience, and connectionism, in an attempt to further understand how brains predict, how they have evolved to do so, and how the neural mechanisms behind prediction might assist artificial intelligence researchers in building better, more adaptive, systems.

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 Times New Roman by Westchester Publishing Services.

Library of Congress Cataloging-in-Publication Data

Names: Downing, Keith L., author.
Title: Gradient expectations : structure, origins, and synthesis of predictive neural networks / Keith L. Downing.
Description: [Cambridge, Massachusetts] : The MIT Press, [2023] | Includes bibliographical references and index.
Identifiers: LCCN 2022037237 (print) | LCCN 2022037238 (ebook) | ISBN 9780262545617 (paperback) | ISBN 9780262374682 (epub) | ISBN 9780262374675 (pdf)
Subjects: LCSH: Deep learning (Machine learning) | Neural networks (Computer science) | Conjugate gradient methods.
Classification: LCC Q325.73 .D88 2023 (print) | LCC Q325.73 (ebook) | DDC 006.3/2—dc23/eng20230302
LC record available at https://lccn.loc.gov/2022037237
LC ebook record available at https://lccn.loc.gov/2022037238