

# **Plato's Camera**

**How the Physical Brain Captures a Landscape of Abstract Universals**

**Paul M. Churchland**

**The MIT Press  
Cambridge, Massachusetts  
London, England**

© 2012 Massachusetts Institute of Technology

All rights reserved. No part of this book may be reproduced in any form by any electronic or mechanical means (including photocopying, recording, or information storage and retrieval) without permission in writing from the publisher.

For information about special quantity discounts, please email [special\\_sales@mitpress.mit.edu](mailto:special_sales@mitpress.mit.edu)

This book was set in Stone Sans and Stone Serif by Toppan Best-set Premedia Limited. Printed and bound in the United States of America.

Library of Congress Cataloging-in-Publication Data

Churchland, Paul M., 1942–.

Plato's camera : how the physical brain captures a landscape of abstract universals / Paul M. Churchland.

p. cm.

Includes bibliographical references (p. ) and index.

ISBN 978-0-262-01686-5 (hardcover : alk. paper)

1. Philosophy of mind. 2. Universals (Philosophy). 3. Cognitive neuroscience—Philosophy. I. Title.

B945.C473P63 2012

128'.2—dc23

2011021046

10 9 8 7 6 5 4 3 2 1

# Contents

Preface vii

## **1 Introduction: A Fast Overview 1**

- 1 Some Parallels and Contrasts with Kant 1
- 2 Representations in the Brain: Ephemeral versus Enduring 4
- 3 Individual Learning: Slow and Structural 11
- 4 Individual Learning: Fast and Dynamical 16
- 5 Collective Learning and Cultural Transmission 25
- 6 Knowledge: Is It True, Justified Belief? 30

## **2 First-Level Learning, Part 1: Structural Changes in the Brain and the Development of Lasting Conceptual Frameworks 35**

- 1 The Basic Organization of the Information-Processing Brain 35
- 2 Some Lessons from Artificial Neural Networks 38
- 3 Motor Coordination 45
- 4 More on Colors: Constancy and Compression 50
- 5 More on Faces: Vector Completion, Abduction, and the Capacity for 'Globally Sensitive Inference' 62
- 6 Neurosemantics: How the Brain Represents the World 74
- 7 How the Brain Does *Not* Represent: First-Order Resemblance 78
- 8 How the Brain Does *Not* Represent: Indicator Semantics 90
- 9 On the Identity/Similarity of Conceptual Frameworks across Distinct Individuals 104

## **3 First-Level Learning, Part 2: On the Evaluation of Maps and Their Generation by Hebbian Learning 123**

- 1 On the Evaluation of Conceptual Frameworks: A First Pass 123
- 2 The Neuronal Representation of Structures Unfolding in Time 139
- 3 Concept Formation via Hebbian Learning: Spatial Structures 157
- 4 Concept Formation via Hebbian Learning: The Special Case of Temporal Structures 165
- 5 A Slightly More Realistic Case 170

6	In Search of Still Greater Realism	174
7	Ascending from Several Egocentric Spaces to One Allocentric Space	180
<b>4</b>	<b>Second-Level Learning: Dynamical Changes in the Brain and Domain-Shifted Redeployments of Existing Concepts</b>	<b>187</b>
1	The Achievement of Explanatory Understanding	187
2	On the Evaluation of Conceptual Frameworks: A Second Pass (Conceptual Redeployments)	196
3	On the Evaluation of Conceptual Frameworks: A Third Pass (Intertheoretic Reductions)	204
4	Scientific Realism and the Underdetermination of Theory by Evidence	215
5	Underdetermination Reconceived	223
<b>5</b>	<b>Third-Level Learning: The Regulation and Amplification of First- and Second-Level Learning through a Growing Network of Cultural Institutions</b>	<b>251</b>
1	The Role of Language in the Business of Human Cognition	251
2	The Emergence and Significance of Regulatory Mechanisms	255
3	Some Prior Takes on This Epicerebral Process	261
4	How Social-Level Institutions Steer Second-Level Learning	268
5	Situated Cognition and Cognitive Theory	274
	Appendix	279
	References	281
	Index	287

## Preface

That the *eye* is a camera is now a commonplace. It most surely is a camera, and we understand in great detail how it works. That the *brain* might be a camera is much less obvious. Indeed, the suggestion is likely to furrow one's forehead in sheer incomprehension. What could be the point of such a metaphor?

Its point is very different from the case of the eye, but the metaphor is no less apt. The eye constructs a representation, or 'takes a picture,' of the landscape or configuration of the objective *spatiotemporal particulars* currently displayed before its lens. This picture-taking process is completed in milliseconds, and the eye does it again and again, because its target landscape is typically in constant flux. The learning brain, by contrast, very slowly constructs a representation, or 'takes a picture,' of the landscape or configuration of the *abstract universals*, the *temporal invariants*, and the *enduring symmetries* that structure the objective universe of its experience. That process takes months, years, decades, and more, because these background features take time to reveal themselves in full. Moreover, the brain of each creature typically undergoes this 'picture taking' process only once, so as to produce the enduring background conceptual framework with which it will interpret its sensory experience for the rest of its life. And yet the brain manages to pull that abstract information—about the universe's timeless categorical and dynamical structure—from its extended sensory encounters with that universe, no less than the eye manages to pull a representation of the fleeting here-and-now from its current luminous input. For this reason, it is appropriate to think of the biological brain as *Plato's Camera*. This marvelous organ has the power to gain a lasting grip on those much more fundamental dimensions of reality, those dimensions that are timeless, changeless, and still.

But what manner of 'pictures' might these decidedly more abstract representations be? On this score, think *maps*. Not the two-dimensional

maps that grace your automobile's glove compartment, but *high-dimensional* maps—maps with three, or a hundred, or even a million distinct dimensions, maps with extraordinary resolution and structural detail. Such maps reside—hundreds and perhaps even thousands of distinct maps—inside the brains of animals in general, and inside the brains of humans in particular. They are maps not of any geographical realities; their high-dimensionality takes them out of that comparatively meager realm. Rather, they are maps of abstract *feature domains*. They are maps of families of complex *universals* and the often intricate similarity-and-difference relations that unite and divide them. They are maps, that is, of the timeless and invariant *background structure* of the ever-changing, ever-unfolding physical universe in which all brains are constrained to operate. They are maps that constitute the 'conceptual frameworks' so familiar to us from the philosophical tradition, and so vital to any creature's comprehension of the world in which it lives.

However, and contrary to tradition, these frameworks are *not* families of predicate-like elements, united by a further family of sentence-like general commitments in which those elements appear. They are not Quinean 'webs of belief,' nor any other classical 'system-of-sentences.' Indeed, they are not *linguaformal* at all. Instead, these high-dimensional maplike frameworks typically consist of a large family of high-dimensional prototype points and prototype-trajectories united and mutually configured by an intricate family of similarity (i.e., proximity) and difference (i.e., distality) relations. The full range of such prototype points and trajectories represent the full range of possible *kinds of things*, and possible *kinds of processes* and *behavioral sequences*, that the creature expects, or is conceptually prepared, to encounter.

Their high-dimensionality and abstract subject-matters notwithstanding, such maps represent the world in the same fashion that any successful map represents its target domain. Specifically, there is a homomorphism or conformal relation—that is, a similarity-of-internal-structure relation—between the configuration of map-elements within the neuronal activation-space that embodies the map at issue, on the one hand, and the configuration of objective similarity-and-difference relations that structure the abstract domain of objective features or processes thereby represented, on the other. In short, the inner conceptual map 'mirrors' the outer feature-domain. Imperfectly, no doubt. But possession of the map allows the creature to anticipate at least some of the real-world features that it encounters, at least some of the time. The interactive library of such maps, possessed by any creature, constitutes its background knowledge, or better, its

background understanding (for better or for worse), of the world's lasting abstract structure. The creature's subsequent pragmatic adventures depend utterly on that background understanding—on its extent and on its accuracy. Despite the urgings of some prominent Pragmatists, knowledge is not *just* a representation-free matter of being *able* or knowing *how* to behave. Our knowledge is richly representational, as we shall see, and our motor behavior depends on it profoundly. It's just that the relevant representations are not remotely propositional or linguaformal in character. Rather, they are high-dimensional geometrical manifolds. They are maps.

Maps, as we know, can be 'indexed.' That is, a point within the map can be specified, perhaps with a reaching fingertip, as the map-indexer's *current location*, within the larger range of locational possibilities comprehended by the map as a whole. The indexer's fingertip may assume a specific position on the two-dimensional highway map, a position with a unique  $\langle x, y \rangle$  pair of coordinates, to indicate "you are here." The abstract feature-domain maps within the brain can also be 'indexed,' this time by the activity of our *sense organs*, to indicate to the brain "you are *here*" in the space of possible objective situations. A sense organ's current activity can cause a signature *pattern*,  $\langle x_1, x_2, \dots, x_n \rangle$ , of  $n$  simultaneous activation-levels, across some population of  $n$  neurons that embodies the relevant map. And that activation-pattern will specify a unique position within the coordinates of that  $n$ -dimensional space, a position that represents the abstract feature currently confronted in the creature's perceptual environment. In this way, our several sense organs are continuously indexing our many feature-space maps to provide us with an unfolding understanding of our unfolding objective world-situation. Without such internal maps for our senses to index, there would be no understanding at all. We may call this the Map-Indexing Theory of Perception, and it gives a contemporary voice to Plato's further claim that a prior grasp of 'universals' is necessary for any particular case of perceptual understanding.

How those feature-domain maps are actually embodied in one's neuronal populations, thence to be indexed by subsequent sensory activity, is one of the central elements of the account of cognition to be explored in this book, as is the story of how those intricate feature-domain maps are generated or learned in the first place. This is the story of Basic-Level or First-Level Learning, to be pursued in chapters 2 and 3. (As you may imagine, Plato's own fantastical story, of a prebirth visit to a nonphysical realm, has no place in the account to be proposed.)

And there are two further kinds of learning, quite different but no less important, to be pursued in later chapters. Still, I will bring these

introductory remarks to a close here. Their function has been to provoke your imagination, and to prepare your understanding, for an account of cognition that deploys an importantly new set of kinematical and dynamical resources—those of Cognitive Neuroscience—to address the questions that have long defined the discipline of Epistemology. What is knowledge? How is it acquired? How is it, or how should it be, evaluated? How does it make progress? What is its long-term destiny? By the end of this book, you will be confronting a new and unfamiliar set of answers. Whether those answers are worth embracing, I leave for you to judge. But however strange they may be, you will recognize them as brain-focused answers to long-familiar questions.



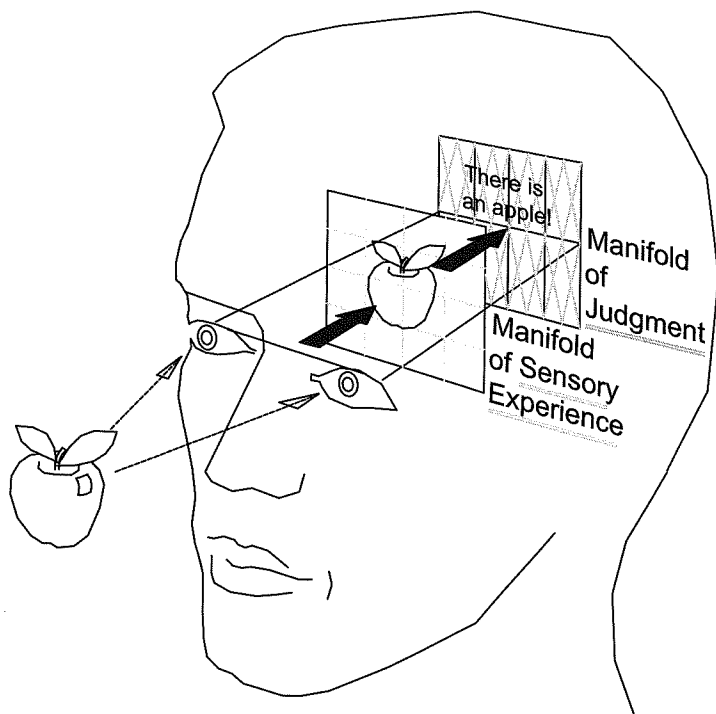
# 1 Introduction: A Fast Overview

## 1 Some Parallels and Contrasts with Kant

A novel idea is sometimes best introduced by analogy, or contrast, with ideas already familiar. Let us begin, then, with the Kantian portrait of our epistemological situation, and more specifically, with his characterization of the two faculties of empirical *intuition* and rational *judgment*. Kant argued that both of these faculties constitute a human-specific canvas on which the activities of human cognition are doomed to be painted. Space and time were claimed to be the 'pure forms' of all *sensory* intuition—the abstract background forms, that is, of all possible human sensory representations. And the organized family of the various 'pure concepts of the understanding' were claimed to provide the inevitable framework of expression for any of the *judgments* that we humans ever make about the empirical world. Accordingly, while the world-in-itself (the 'noumenal' world) is certainly not 'constructed' by us, the world-as-perceived-and-thought-by-us (the 'empirical' world of three-dimensional physical objects) does indeed display a substantial component that reflects precisely the peculiar contributions brought to the business of cognition by our own internal cognitive machinery (fig. 1.1).

Kant, of course, had an agenda that we moderns need not share: the desire to vindicate an alleged class of *synthetic a priori* truths (e.g., geometry and arithmetic), and to explain in detail how such truths are possible. Beyond this, he had commitments that we moderns may wish to deny, such as the innateness of the 'pure forms and concepts' at issue, and the implasticity of the cognitive activities that they make possible. But his portrait still constitutes a useful starting point from which a competing, and importantly different, portrait can be quickly sketched and readily grasped.

Consider, then, the possibility that human cognition involves not just *two* abstract 'spaces'—of possible human *experiences* on the one hand, and

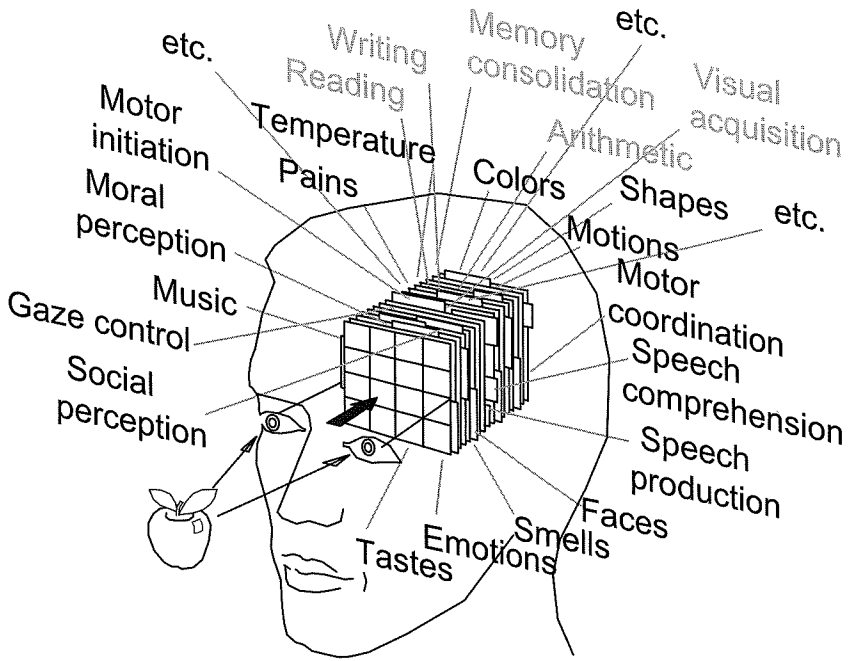


**Figure 1.1**

Kant's portrait of human cognition.

possible human *judgments* on the other—but rather many *hundreds*, perhaps even *thousands*, of internal cognitive ‘spaces,’ each of which provides a proprietary canvas on which some aspect of human cognition is continually unfolding (fig. 1.2). Consider the possibility that each such figurative cognitive space is physically embodied in the very real space of the possible *collective* activities of some proprietary population of topic-specific *neurons* within the human or animal brain.

Suppose also that the internal character of each of these representational spaces is not *fixed* by some prior decree, either divine or genetic, but is rather shaped by the extended experience of the developing animal, to reflect the peculiar empirical environment and practical needs that it encounters, and the peculiar learning procedures embodied in the brain's ongoing business of synaptic modification. These internal spaces may thus be *plastic* to varying degrees, and may hold out the promise of an enormous *range* of conceptual and perceptual possibilities for one and the same



**Figure 1.2**  
The multiple-maps portrait of human cognition.

species of creature, in stark contrast to the frozen conceptual prison contrived for us by Kant.

Consider also the possibility that the human brain devotes every bit as much of its cognitive activity to the production and administration of coherent *motor behavior* (walking, swimming, speaking, piano playing, throwing strikes, making a dinner, running a meeting) as it does to the perceptual and judgmental activities of typical concern to traditional philosophies. And note the possibility that such proprietary cognitive spaces—whose neuronal basis is located, for example, in the frontal cortex, the motor cortex, and the cerebellum—can successfully *represent* the complex motor procedures and action procedures listed above, by means of the muscle-manipulating *trajectories* of the collective neuronal activities within those spaces. Note also that such trajectories and limit-cycles within sundry *sensory* spaces can equally well represent complex *causal* processes and *periodic* phenomena, as externally encountered in perceptual experience rather than as internally generated to produce bodily behavior.

So we begin by expanding the number of representational spaces, into the hundreds and thousands, far beyond the Kantian pair. We locate them in discrete anatomical parts of the brain. We make each one of them plastic and pluripotent in its semantic content and its conceptual organization. And we reach out to include motor cognition and practical skills, along with perceptual apprehension and theoretical judgment, as equal partners in our account of human knowledge.

## 2 Representations in the Brain: Ephemeral versus Enduring

But we have not yet confronted the single largest contrast that we need to draw, between the Kantian portrait above and the account to be pursued in this book. For Kant, there is no question but that the fundamental unit of human cognition is the *judgment*, a unit that lives in a space of sundry logical relations with other actual and possible judgments, a unit that displays the characteristic feature of truth or falsity. However, on the account proposed in this book, there is no question but that the 'judgment'—as conceived by Kant and by centuries of other logicians—is *not* the fundamental unit of cognition, not in animals, and not in humans either. Instead, the fundamental unit of cognition—strictly, of occurrent or ephemeral cognition—is the *activation pattern* across a proprietary *population* of neurons. It is the activated *point* within any one of the many hundreds of representational *spaces* urged above.

This fundamental form of representation is one we share with all other creatures that possess a nervous system, and it does roughly the same job in every space and in every case. Such a representation lets the animal know—better, it constitutes the animal's knowledge—that its current location, in the space of possible situations comprehended by the underlying population of neurons, is *here* on the cognitive map embodied in that population. That activation point is rather like the brilliant dot of a laser pointer illuminating one tiny spot on an otherwise unlit highway map, a moving dot that continuously updates one's current position in the space of geographic possibilities represented by that map.

Since these thousands of spaces or 'maps' are all connected to one another by billions of axonal projections and trillions of synaptic junctions, such specific locational information within one map can and does provoke subsequent pointlike activations in a *sequence* of downstream representational spaces, and ultimately in one or more *motor*-representation spaces, whose unfolding activations are projected onto the body's muscle systems, thereby to generate cognitively informed behaviors.

On this view, Kantian-style 'judgments,' though entirely real, constitute an extremely peripheral form of representational activity, marginal even for adult humans, and completely absent in nonhuman animals and pre-linguistic children. How we humans manage to generate a 'language space,' to sustain our speech production and speech comprehension, is an engaging scientific question to which we shall return later in the book. For now, let me announce that, for better or for worse, the view to be explored and developed in this book is diametrically opposed to the view that humans are capable of cognition precisely because we are born with an innate 'language of thought.'

Fodor has defended this linguaformal view most trenchantly and resourcefully in recent decades, but of course the general idea goes back at least to Kant and Descartes. My own hypothesis is that all three of these acute gentlemen have been falsely taken in by what was, until recently, the only *example* of a systematic representational system available to human experience, namely, human language. Encouraged further by the structure of our own dearly beloved Folk Psychology,<sup>1</sup> they have wrongly read back *into* the objective phenomenon of cognition-in-general a historically *accidental* structure that is idiosyncratic to a single species of animal (namely, humans), and which is of profoundly secondary importance even there. We do of course use language—a most blessed development we shall explore in due course—but language-like structures do not embody the basic machinery of cognition. Evidently they do not do so for animals, and not for humans either, because the human neuronal machinery, overall, differs from that of other animals in various small degrees, but not in fundamental kind.

An account of cognition that locates us on a continuum with all of our evolutionary brothers and sisters is thus a prime desideratum of any responsible epistemological theory. And the price we have to pay to meet it is to give up the linguaformal 'judgment' or 'proposition' as the presumed unit of knowledge or representation. But we need no longer make this sacrifice in the dark. Given the conceptual resources of modern neurobiology and cognitive neuromodeling, we are finally in a position to pursue an alternative account of cognition, one that embraces some highly specific and very different units of representation. What I hope to show in what follows is that this initial investment in an unorthodox

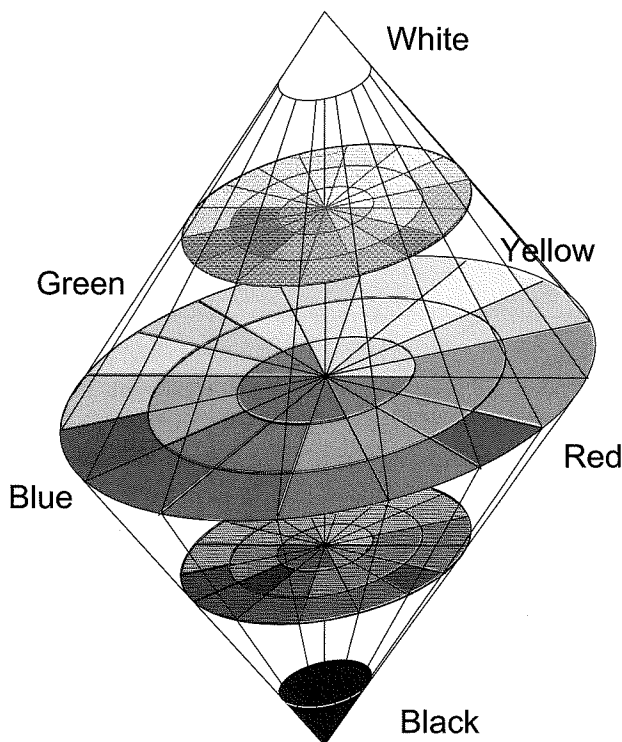
1. Folk psychology is the ordinary conceptual framework—comprehending notions like "desires that *p*," "believes that *p*," "fears that *p*," "intends that *p*"—used by lay persons to explain and predict the behavior of other humans.

assumption, concerning the nature of our primary units of representation, will yield extraordinary dividends as we proceed.

One of those dividends emerges very early in our story, for the portrait of knowledge held out to us draws a fundamental distinction between the ephemeral vehicles of our knowledge of the fleeting-here-and-now on the one hand, and the comparatively stable and enduring vehicles of our *background* knowledge of the world's-general-structure-in-space-and-time on the other. As suggested, the former vehicles are the fleeting activation patterns across a given population of neurons; they are the ever-moving, ever-jumping points of here-and-now activation within their proprietary conceptual subspaces. Think again of the moving, darting laser-dot on the otherwise unlit road map.

The latter or 'background' vehicles are entirely different. At this level of *general* knowledge, the vehicle or unit of representation is *the entire conceptual framework*. It is the *entire activation space* for the relevant population of neurons, a space that has been sculpted by months or years of learning, a space that encompasses all of the possible instances of which the creature currently has any conception. Indeed, that space is precisely the background canvas on which each fleeting *instance* of any category gets 'painted,' and that painting consists in nothing more nor less than an activation at a specific location within that sculpted space of *possible* activations. (To reprise the metaphor of the previous paragraph, the background conceptual framework is the entire *road map* at issue, the waiting space of all *possible* positions that the laser-dot might at some time illuminate.)

To illustrate with a simple and concrete example, consider the spindle-shaped space of possible *color-experiences* portrayed in figure 1.3 (plate 1). That space embodies every possible color-qualia of which the human visual system is normally capable, and that space is organized in a very specific way, a way common to all humans with normal color vision. Barring some form of color-blindness, we all share the *same* family of distance-relations and betweenness-relations that collectively locate each color-representation within that space, relative to all of its sibling color-representations. To possess this roughly double-coned space is to have the most rudimentary human knowledge of the general structure of the domain of objective colors. And to have a current activation vector at a specific point within this internal conceptual space (halfway up the central axis, for example) is to have a current representation or *experience* of a specific color (in that case, middle-gray). Figure 1.3 (plate 1) displays a color-coded sampling of representational positions within that space, but of course the space itself is continuous and includes positions for all of the various hues and shades



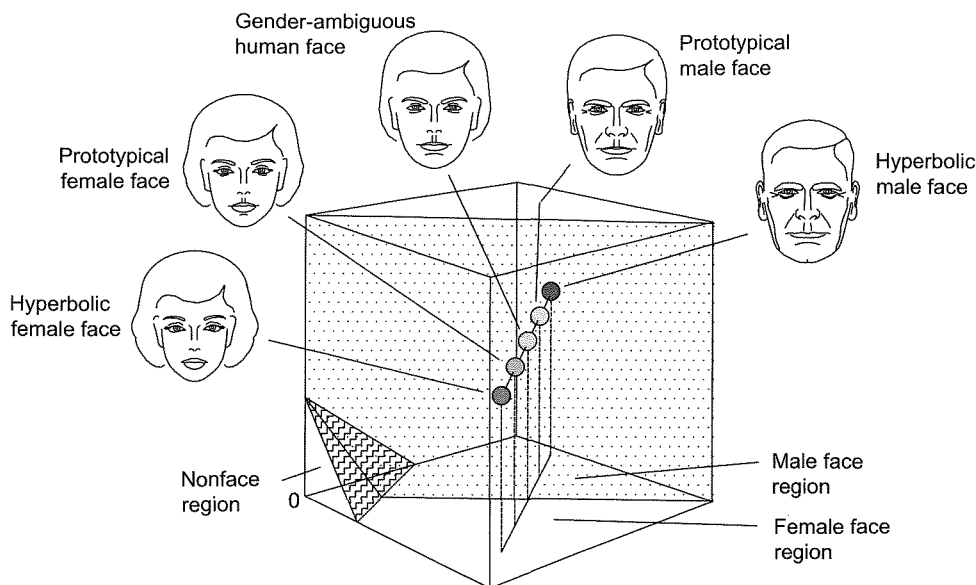
**Figure 1.3**

A map of the space of possible colors. See plate 1.

that occupy the various positions between the sample colors actually displayed in the diagram.

A second example concerns one's internal conceptual space for representing human *faces*, as (speculatively) portrayed in the example of figure 1.4. In fact, that three-dimensional space is a schematic of the activation space of a specific population of neurons within an *artificial* neural network, one that attempts to model the gross structure of the primary visual pathway in humans. That network was trained to discriminate faces from nonfaces, male faces from female faces, and to reidentify the faces of various named individuals across diverse photographs of each.<sup>2</sup>

2. See Cottrell 1991. The artificial network actually had eighty neurons at the crucial representational layer, not three, as my three-dimensional diagram would suggest. I deploy this low-dimensional fiction to help make visual sense of what is happening in the more complex case of an eighty-dimensional space.



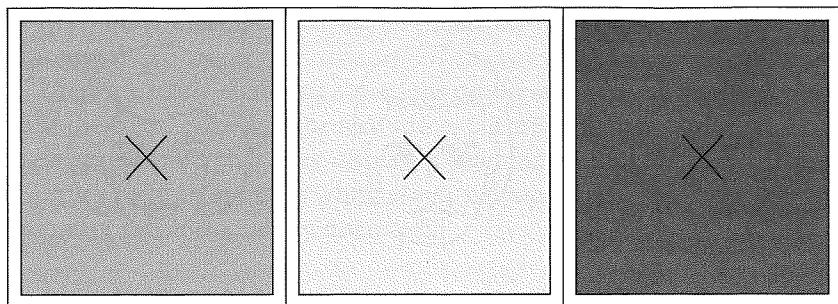
**Figure 1.4**

A map of the space of possible human faces.

As you can see immediately, the training process has produced a hierarchical structure of distinct representational regions within the space as a whole. Sundry nonface images presented to the network's sensory neurons get represented by sundry activation points close to the *origin* of this downstream conceptual space. By contrast, various face images, presented as sensory input, get represented by various activation points within a much larger 'face region' away from the origin. That complementary region is itself split in two, into roughly equal regions for male faces and female faces, respectively. Within each gender-subspace lies a scatter of much smaller subspaces (not shown), each of which comprehends a cluster of closely proximate activation points for representing distinct sensory presentations of the face of a unique individual.

Of course, *every* point in the larger region represents a face of *some* kind or other, not just the five points illustrated. But as one moves progressively farther away from the solid straight line between the average or prototypical male-face point and the average or prototypical female-face point (see again fig. 1.4), one encounters representations of faces that are increasingly nonstandard or hyperbolic in various ways. Since such outlying points





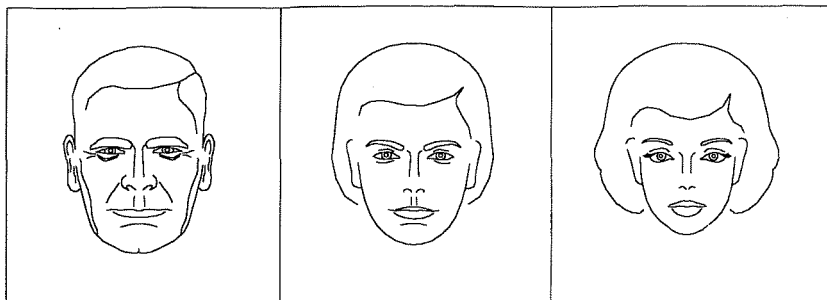
**Figure 1.5**

Colored afterimages due to neuronal fatigue. See plate 2.

contrast maximally with the more central points, they represent faces that ‘stand out’ from the crowd, faces that differ maximally from what we expect in an average, humdrum face. And just as the color space has a middle-gray at its central point, so our face space has a gender-neutral face at its central point.

The evident structure of these two spaces—for colors and for faces—reveals itself anew in some simple perceptual illusions. Fix your gaze on the × at the center of the reddish square on the right side of figure 1.5 (plate 2) and hold it for ten seconds. Then fixate on the × in the central square immediately to its left. For a few brief seconds you will perceive that hue-neutral square as being faintly green. A parallel experiment, beginning with the genuinely green square at the left of figure 1.5, will make that same central square look faintly red. If you look back to the color space of figure 1.3 (plate 1), you will notice that red and green are coded at opposite sides of that space, with maximal distance between them. Forcing oneself into a protracted period of steady neural activity, in order to code either one of these extremal colors, produces a specific pattern of short-term ‘fatigue’ and/or ‘potentiation’ across the three kinds of color-coding neurons that underlie our three-dimensional color space.

That fatigue induces those neurons, when finally relieved of their extremal burden, to fall back *past* their normal neutral points, that is, to relax back into a momentary activation pattern that briefly *misrepresents* a *hue-neutral* perceptual input as if it were an input from the side of the color space exactly *opposite* from the original, fatigue-inducing stimulus. Hence the brief illusion of its exactly complementary color.



**Figure 1.6**

Gender-switch effect due to neuronal fatigue.

You may observe a similar short-term illusion concerning the *gender* of the face portrayed in the middle position of figure 1.6. That deliberately gender-ambiguous, vector-average, neutral human face is flanked by a hyperbolic female face on the right, and a hyperbolic male face on the left. Use your hand to cover all but the male face, and fixate on the bridge of his nose for ten seconds (do not move your eyes). Then slide your hand to the right so as to reveal (only) the neutral face, shift your fixation suddenly to the bridge of *its* nose, and make a snap judgment as to the *gender* of that (objectively neutral) face. Since hyperbolic male and female faces (like red and green colors) are coded at opposite sides of the relevant activation space, the preceding color-experiment suggests that, thanks to a comparable fatigue or saturation effect, the neutral face will now look, if anything, to be faintly *female*. If your reaction is like mine, it does. And as with the color illusion, you can also perform the opposite experiment. Fixate for at least ten seconds on the isolated hyperbolic female face, and then suddenly refixate on the neutral face to see if it subsequently looks, if only for a second or so, to be distinctly *male*.<sup>3</sup> Be your own judge. (You may also notice an intriguing *age* effect. The central face appears as two entirely *distinct individuals* in the two conditions. When seen as male, it looks quite young—no more than twenty. When seen as female, it looks

3. A more elaborate and carefully controlled experiment of this sort is reported by O'Toole and company (see Blanz et al. 2000). The distortions induced in normal face perception, by various hyperbole-induced profiles of neuronal fatigue, extend to selectively *facilitating* the recognition of specific test-faces. This happens when the test subject is 'primed' by face-images deliberately constructed to be face-space opposites, or '*anti-faces*,' of the face to be identified.

to be a much older person—no less than thirty-five. I'm still pondering that one.)<sup>4</sup>

There are many charming variations on these experimental themes, for both the color space and the face space, but we must here move on. That these spaces *exist*, that they display a determinate internal *structure*, and that they have an at least slightly *plastic* neuronal basis, are the suggestive lessons of this introductory discussion.

### 3 Individual Learning: Slow and Structural

Still, one might ask of these introductory spaces if they are perhaps *innate*—that is, somehow specified in the human genome. Whatever the correct answer here, it is surely empirical. As for the human color space, the answer may well be positive. As we shall see later in this book, the probable neuronal and synaptic basis for the three-dimensional color solid discussed above seems to be anatomically simple, repetitive, and highly uniform across normal human individuals. It is therefore a fair candidate for being somehow specified in the human genome.

On the innateness of the human representational-space for *faces*, however, our judgment must incline to the negative, for several reasons. First, and unlike the three-dimensionality of both your color space and the (schematic) face space of figure 1.2, the dimensionality of your brain's face space lies almost certainly in the thousands and perhaps in the tens of thousands. Second, the configuration of the synaptic connections that structure that face space must therefore include at least a million synaptic connections.<sup>5</sup> Unlike our color system, none of these connections are repetitive in strength or placement, so their genetic specification would be

4. These two examples portend a multitude. Readers may remember the three buckets of water—one tepid, flanked by one hot and one cold—touted by Locke and Berkeley. The tepid one feels warm, or cool, depending on which of the two flanking buckets initially fatigued the thermal receptors within the testing hand. The waterfall illusion provides a fourth example. Given three textured squares, a down-flowing motion in the left square will eventually induce an illusory *upward* motion in the motionless central square, and an upward-flowing motion in the right square will induce the opposite illusion. What is intriguing about the color and face illusions, as distinct from the two illusions just cited, is that they occur in a representational space of more than one dimension—three in the case of colors and probably hundreds in the case of faces.

5. On average, each neuron in the brain enjoys perhaps 1,000 synaptic connections with other neurons.

costly. And third, those connections are not uniform across individuals in any case. We already know that the metrical structure of an individual person's face space *varies* substantially as a function of which culture she grew up in—specifically, as a function of which racial or ethnic group happened to exhaust or dominate her childhood experience with human faces. (It is sometimes called, inaccurately, the *other-race effect*.) In sum, the basis of our face space is complex, not recursively specifiable, and highly variable across individuals. It is therefore a most unlikely candidate for being coded in the genome.

Of course, the sheer *existence* of a neuronal population primed to take on the job of parsing faces is likely something that is genetically specified. After all, each of us has an inferotemporal cortex (the presumed region of the human brain that supports facial recognition), and all normal infants fixate on facelike stimuli from birth. But the adult structure of that space, its idiosyncratic dimensions, and its internal similarity-metric are all features that are *epigenetically* determined. By far the greater part, and perhaps even all, of what anyone knows about faces has its origins in one's *postnatal* experience with faces.<sup>6</sup>

Such an antinativist conclusion, concerning the neuronal and synaptic basis of the brain's representational spaces, is appropriate for almost all of the many spaces that the brain comprehends. The only likely exceptions will be those occasional spaces—such as our color space, perhaps—that directly code the behavior of our various sensory neurons. There is a rule here that cannot be too far wrong. Specifically, the greater the *distance* between a given neuronal population and the body's sensory neurons—as measured by the number of distinct synaptic connections that have to be traversed for an axonal message to get from the one population to the other—the exponentially greater is the likelihood that the target population embodies a representational space that has been structured by *learning*.

The fact is, the modification, extinction, and growth of new *synaptic connections* is the single most dramatic dimension of structural change within the brain, from birth onward. Creating and adjusting the precious

6. One's prenatal *motor* and *proprioceptive* experience with one's own face, lips, and tongue may give one a leg up on the postnatal *visual* processing of postnatally encountered faces—for those 'visual' faces are homomorphic structures subject to homomorphic behaviors—but the 'knowledge' involved here is still epigenetic, even if acquired while still in the womb. The fact is, infants are highly active in the late stages of pregnancy: they move their tongues and mouths, and even suck their thumbs and fists.

configuration of one's  $10^{14}$  synaptic connections is the very essence of learning in one's infant and childhood stages, for it is the collective configuration of the synaptic connections onto any neuronal population that *dictates* the family of categories embodied in that population's proprietary activation space. As our artificial neural-network models reveal, that collective configuration of synaptic connections is what *shapes* the web of similarity and difference relations that unite and divide the categories within the resulting activation space. And that same collective configuration of synaptic connections is what *transforms* any incoming activation pattern (from, for example, an earlier population of sensory neurons) into a new activation pattern located at a specific location within this carefully sculpted secondary space. Those assembled connections are thus central to *both* the relatively enduring conceptual framework that is acquired by the learning individual in the first place, *and* to its subsequent moment-to-moment activations by ephemeral sensory inputs. These synaptic connections are, simultaneously, the brain's elemental information *processors*, as well as its principal *repository of general information* about the world's abstract structure.

Accordingly, the establishment of one's myriad synaptic connections, and the fine-tuning of their individual strengths or 'weights,' constitutes the focal and primary process of learning that anyone's brain undergoes in its first twenty years of life, and especially in the first ten, and most especially in the first five. It is during these periods that one's background conceptual framework is slowly established, a framework that is likely, with only minor modifications, to remain with one for the rest of one's life.

Notably, and despite its primacy, that synapse-adjusting/space-shaping process is almost wholly ignored by the traditions of academic epistemology, even into these early years of our third millennium. This is perhaps not too surprising. The profound inaccessibility of the brain's microprocesses, the absence of compensatory computer models thereof, the focus of professional concern on normative matters, and the tyrannical primacy of folk-psychological conceptions of cognitive activity collectively contrive for *past* theorists an acceptable excuse for this monumental and crippling oversight. But the validation-dates on these several excuses (as in "use before 01/01/2001") have all run out.

Beyond conventional microscopy, an ever-growing armory of experimental techniques and instruments—such as the selective staining of neurons and their connecting pathways, electron microscopy, single-cell microelectrodes, multiple-array microelectrodes, genetically modified mice, CAT scans, PET scans, MRI scans, fMRI scans, activity-sensitive

fluorescent dyes—now provides us with an overlapping set of windows onto the brain's physical structure and its neuronal activities, from the subcellular details of its molecular activities up to the molar-level behavior of its brainwide neuronal networks. The brain is no longer an inaccessible black box. On the contrary, the steady stream of experimental data just mentioned provides an ever-expanding set of empirical *constraints* on responsible cognitive theorizing. It also makes possible the invaluable sorts of back-and-forth interactions, between theoretical suggestions on the one hand and experimental probings on the other, that have proven so successful in other scientific endeavors. Theories can suggest and motivate experiments that might never have been tried, or even conceived, in their absence. Equally, experimental results can force needed modifications in the theories being tested. And the cycle begins anew.

Artificial neural networks, as realized directly in electronic hardwares, or as modeled at one remove in conventional computers, provide us with a second way of formulating and testing theories of how cognitive activities can arise from interconnected sets of neuron-like elements. Unlike biological networks, the artificial networks can be as simple and as well controlled as we wish. We can monitor their every connection, their every twitch, and their every change, without killing, damaging, or interfering with the network elements involved. Since, being electronic, they also run much faster than do their biological namesakes, we can also perform and complete experiments on their 'learning' activities within *hours*, experiments that would take months or years to perform on a biological brain. This means, of course, that we can often learn very swiftly that our initial attempt to model some brain function or other is just *wrong*. We can thus go back to the empirical brain for renewed theoretical inspiration, and then go back to the drawing board in hopes of constructing a more faithful artificial model.

This activity has already yielded the rough outlines of a competing conception of cognitive activity, an alternative to the "sentential" or "propositional-attitude" model that has dominated philosophy for the past 2,500 years. This newer conception is important for a number of reasons, not least because it finally (finally!) makes some explanatory contact with the physical and functional details of the biological organ—the brain—that actually sustains our cognitive activity. But there are many other reasons as well, perhaps the first of which is the novel account it provides of the *origins* of any individual's conceptual framework.

This is an issue that is mostly ignored, submerged, or finessed with a stick-figure theory by the philosophical tradition. Two styles of 'solution'

have dominated. The writings of Plato, Descartes, and Fodor illustrate the first great option: since one has no idea how to explain the origin of our concepts, one simply pronounces them innate, and credits either a prior life, almighty God, or fifty million years of biological evolution for the actual lexicon of concepts we find in ourselves. The works of Aristotle, Locke, and Hume illustrate the second great option: point to a palette of what are taken to be sensory 'simples,' such as the various tastes, smells, colors, shapes, sounds, and so forth, and then explain our base population of simple concepts as being faint 'copies' of the simple sensory originals, copies acquired in a one-shot encounter with such originals. Nonsimple or 'complex' concepts are then explained as the recursively achieved concatenations and/or modulations of the simple 'copies' (and never mind the origins of *that* constructive machinery).

Both options are hopeless, and for interesting reasons. If we put Almighty God and Plato's Heaven aside as nonstarters, the preferred (i.e., the evolutionary) version of the first option confronts the difficulty of how to code for the individual connection-places and connection-strengths of fully  $10^{14}$  synapses—so as to sculpt the target conceptual framework—using the resources of an evolved genome that contains only 20,000 genes, 99 percent of which (all but a paltry 300 of which) we share with *mice*, with whom we parted evolutionary company some fifty million years ago. The problem here is not so much the nine-orders-of-magnitude gap between the number of available genes and the number of synaptic connections (in principle, a recursive procedure can bridge a gap of any size), although that yawning gap does give pause for much thought. The real difficulty is the empirical fact that each person's matured synaptic configuration is radically different from anyone else's. It is utterly unique to that individual. That synaptic configuration is thus a hopeless candidate for being recursively specifiable as *the same* in all of us, as it must be if the numbers gap just noted is to be recursively bridged, and if the same conceptual framework is thereby to be genetically recreated in every normal human individual. (More on this later.)

The second option is little or no better than the first. Empirical research on the neuronal coding strategies deployed in our several sensory systems reveals that, even in response to the presumptively 'simplest' of sensory stimuli, the sensory messages sent to the brain are typically quite complex, and their synapse-transformed offspring—that is, the downstream conceptualized representations into which they get coded—are more complex still, typically *much* more complex. (More on this later, also.) The direct-inner-copy theory of what concepts are, and of how we acquire them, is

a joke on its face, a fact reflected in the months and years it takes any human infant to acquire the full range of our discriminatory capacities for most of the so-called 'simple' sensory properties. (It takes *time* to configure the brain's  $10^{14}$  synapses, even to comprehend the 'simple' properties at issue.) Additionally, and as anyone who ever pursued the matter was doomed to discover, the recursive-definitions story suggested for 'complex' concepts was a crashing failure in its own right. Try to construct an explicit definition of "electron," or "democracy"—or even "cat" or "pencil," for that matter—in terms of concepts that plausibly represent sensory simples.

Perhaps the strongest argument that either side—blanket nativism versus concatenative empiricism—could adduce in its own favor was the evident poverty of the opposing view. Given the depth of their poverties, each had a nontrivial point, so long as these alternatives were seen as exhausting the possibilities. But of course they do not. As we shall see in the next chapter, we already possess a workable story of how a neuronal activation space can be slowly sculpted, by experience, into a coherent and hierarchical family of prototype regions. This story also accounts for the subsequent context-appropriate *activation* of those concepts, activations made in response to sensory instances of the categories they represent. And, as we shall see in subsequent chapters, this same neurostructural and neurofunctional framework sustains penetrating explanations of a wide variety of perceptual and conceptual phenomena, including the profile of many of our cognitive *failures*. It forms the essential foundation of the larger epistemological theory to be outlined in the chapters to come.

#### 4 Individual Learning: Fast and Dynamical

But *only* the foundation. Adjusting trillions of synaptic connections is not the only way to engage in a process worthy of the term "learning," and boasting a well-tuned configuration of such connections is not the only way to embody systematic 'knowledge' about the world. To this basic dimension of learning—the dimension of *structural* changes in the brain—we must add a second dimension of learning: the dimension of *dynamical* changes in the brain's typical or accustomed modes of operation. These dynamical changes can take place on a much shorter time scale (seconds and less) than the structural changes we have been discussing (weeks, months, and years), and they typically involve no structural changes whatever, at least in the short term. But this dimension of cognitive development is at least as important as its structural precursor, as we shall see.



One can conceive of the brain's dynamical activity as a single moving point in the brain's 'all-up' neuronal activation space, a point in ceaseless motion, a point that spends its time, marble-like, rolling merrily around the almost endless hills and valleys of the conceptual landscape that the basic or structural learning process has taken so long to sculpt. This landscape analogy is accurate enough, in that it rightly suggests that one's unfolding cognitive state tends to favor the 'valleys' (the acquired prototype regions or categories) and to slide off the 'hills' (the comparatively improbable ridges between the more probable valleys). But it fails adequately to suggest the spectacular *volume* of the brain's all-up activation space (the assembled sum of the thousands of vaguely Kantian *subspaces*). Let us do a brief accounting.

If we assume, very conservatively, that each neuron in the brain admits of only ten different functionally significant levels of activity—ten steps between a minimum spiking frequency of 0 Hz and a maximum of 90 Hz, for example—then, since the brain has  $10^{11}$  neurons, we are looking at a space of 10 to the  $10^{11}$  power or  $10^{100,000,000,000}$  functionally distinct, a priori possible global activation states. (For comparison, the accessible universe contains only about  $10^{87}$  cubic meters.) It is within this almost incomprehensible volume of distinct activational possibilities that a given individual's moving activation-point must craft its one-shot, three-score-and-ten years, idiosyncratic cognitive excursion—that individual's conscious (and unconscious) life.

That space is far too large to explore any significant portion of it in a human lifetime ( $\approx 2 \times 10^9$  seconds). If one's activation point were to streak through the possible points at a rate of 100 per second, a lifetime's exploration would touch only  $2 \times 10^{11}$  distinct points, leaving fully  $5 \times 10^{99,999,999,988}$  points unvisited.

This accounting is roughly accurate, but it concerns only the a priori volume of the human neuronal activation space, that is, its potential volume if each neuron's activities were independent of every other's. By contrast, its *a posteriori* volume, if still mind-boggling, is substantially smaller, for a reason the reader will likely already appreciate. The whole point of the synapse-adjusting learning process discussed above was to make the behavior of neurons that are progressively higher in the information-processing hierarchy profoundly and systematically *dependent* on the activities of the neurons below them. Accordingly, that learning process shrinks the space of (empirically) possible global activation points dramatically.

More specifically, it shrinks the original space to a set of carefully contrived internal subspaces, each of which is an attempt to represent—via its acquired internal structure—some proprietary aspect or dimension of the external world's enduring structure. Being devotedly general in their representational significance, these subspaces therefore represent the brain's conception of the full range of *possible* ways in which the actual world may present itself to us in our singular, ongoing perceptual experience. For example, the color space of figure 1.3 attempts to represent the range of all possible colors. The face space of figure 1.4 attempts to represent the range of all possible human faces. A third space might represent the range of all possible projectile motions. A fourth, the range of all possible voices. And so on. Collectively, these many subspaces specify a set of 'nominally possible' worlds—worlds that instantiate the same categories and share the enduring causal structure of our own world, but differ in their initial conditions and ensuing singular details.

Accordingly, those spaces hold the key to a novel account of both the semantics *and* the epistemology of modal statements, and of counterfactual and subjunctive conditionals. The account envisioned here does not require us to be swallow-a-camel 'realists' about the objective existence of possible worlds, nor does it traffic in listlike state-descriptions. In fact, the relevant representations are entirely nondiscursive, and they occupy positions in a space that has a robust and built-in *probability metric* against which to measure the likelihood, or *unlikelihood*, of the objective feature represented by that position's ever being instantiated.

Once again, the acquired microstructure of the brain holds out some much-needed explanatory resources, but the price of using them is a reconception of our modal, subjunctive, and counterfactual knowledge, a reconception that moves us away from the linguaformal framework of Folk Psychology and classical logical theory, and toward the activation-vector-space framework appropriate to the brain's basic modes of operation.

This shift in focus will be a gain in one further respect. Nonhuman animals demonstrate their possession of modal, subjunctive, and counterfactual knowledge no less than do humans. But nonhuman animals do not traffic in sentences or in propositional attitudes at all, let alone in their evaluation across an infinity of possible worlds. The classical accounts, therefore, cannot explain the great bulk of modal knowledge throughout the animal kingdom. They cannot get beyond the idiosyncratic case at which they were initially aimed, namely, language-using humans. The present approach, by contrast, holds out a uniform account, of broadly 'modal' knowledge, for all terrestrial creatures.

As stated earlier, one's inner cognitive narrative is a specific trajectory through such an antecedently sculpted space. But even given such a well-informed space, the path of that cognitive trajectory is not dictated by one's sensory experience alone. Far from it. At any point in time, your next activation point, within your global activation space, is always dictated (1) partly by your current sensory inputs, (2) partly by the already acquired profile of your background conceptual framework (that is, by the lasting configuration of your synaptic connections), but also, and most importantly in the present context, (3) *by the concurrent activation-state of your entire neuronal population*, a complex factor that reflects your cognitive activity immediately preceding the present computational interaction. This arrangement makes the brain a genuine *dynamical system*, a system capable of a great range of possible behaviors, much of which is unpredictable even in principle.

Once again, a Kantian parallel may help to illuminate the claims here being made, by the contrasts required as well as by the similarities in place. It is arguable that Kant embraced a third great canvas of cognitive activity, distinct from Intuition and Judgment. This was the faculty of Imagination, whose hallmark feature was the *spontaneity* of the cognitive activities there displayed. Unlike Kant, I shall not postulate a distinct space or canvas to sustain the activities of the imagination. On my view, those activities take place in the very same family of neuronal activation spaces discussed earlier. What distinguishes imaginative activity from other forms of cognition is not its location, but its cause. Imaginative cognitive activity arises not from ascending inputs from the several sensory modalities, but from descending or recurrent inputs from neuronal populations higher up in the brain's information-processing hierarchy. It is initiated and steered by brain activity from above rather than by sensory activity from below.

On the matter of spontaneity, however, I line up with Kant. And with good reason. The brain is a dynamical system of unparalleled complexity. It is a continuously varying physical system with many billions of degrees of freedom—the current activation levels of its billions of neurons, for starters. It is a system whose dynamics are decidedly nonlinear, which means that, for many regimes of activity, infinitesimally small differences in one's current cognitive state can snowball exponentially into very large differences in the brain's subsequent cognitive state. This puts inescapable limitations on the degree to which we, or any other conceivable physical device for that matter, can *predict* any brain's unfolding cognitive activities, even on the assumption that the brain's behavior is rigorously deterministic. The problem is that, for such a system, effective prediction requires,

first, infinitely perfect information about the brain's current structural and dynamical state, and second, infinitely accurate computations concerning its law-governed development into subsequent states. Neither requirement can be met in this world, nor even relevantly approximated.

The result is a system whose cognitive behavior, in general, simply cannot be predicted—not by itself, and not by anything else either. This need not mean that *no* regularities will display themselves to the casual eye. To the contrary, when the brain is in the midst of some prototypical activity, such as brushing one's teeth, dealing a deck of cards, or sipping a cup of coffee, its specific motor behavior can be reliably predicted for several seconds or so into the future. And if we look at behavior as reckoned over days and weeks, we can reliably predict that, if the environment remains normal, people will take dinner about six o'clock, go to bed around ten, and get up around six or seven. The details of such periodic behaviors may be beyond us (Will he have sausages, or fish? Will he wear the green pajamas, or the blue? Will he put on the right shoe first, or the left?) But even a nonlinear system can display roughly stable, if endlessly variant, orbits or cycles. Beyond these two exceptions, however (very short-term behaviors, and long-term patterns), a person's cognitive and motor behavior is deeply unpredictable. It displays a spontaneity that reflects its origins in an unmonitorable and ever-shifting mix of mercurial microprocesses.

But let us return to the topic of learning. Beyond a welcome measure of spontaneity, what the recurrent or descending neuronal pathways also make possible is the ongoing *modulation* of the brain's cognitive response to its unfolding sensory inputs. The details of that modulation reflect the ever-changing dynamical state of the brain as a whole; it reflects all of the acquired here-and-now contextual information embodied in the brain at the time of the sensory inputs in question. Most importantly, the current context into which each sensory input arrives is never exactly the same twice, not twice in a lifetime. For even with the brain's *synaptic* connection-weights fixed in their mature configuration, the brain's *dynamical* state—its all-up pattern of *current neuronal activation-levels*—provides an ever-moving, never-repeating cognitive context into which its every sensory input is interpretively received. Accordingly, one never has a truly identical cognitive response on two different occasions, even if the total sensory inputs should happen to be identical on those two occasions. Identical rocks, thrown sequentially into this internal Heraclitean river, will never make exactly the same splash.

To be sure, the resulting differences are usually small, and their downstream cognitive consequences are typically small as well. The brain, like

the solar system, is at least a quasi-stable dynamical system. But sometimes the downstream differences are substantial, and reflect a changed outlook on the world, as when a trusted friend suddenly mistreats an innocent colleague horribly. Though the friend may return to normal in the days and weeks that follow, his smiles and greetings and other social exchanges never look quite the same to you again. Your perceptions, anticipations, and interactions, at least where he is concerned, are permanently altered: you have learned something about his character.

More specifically, that experience has kicked your cognitive trajectory into an importantly different and hitherto unvisited region of your antecedently sculpted neuronal activation space. In the vast majority of that space's many dimensions, your trajectory remains in familiar territory. But in at least a handful of the available dimensions, that trajectory is now exploring new ground.

Strictly, it should be said, that trajectory is *always* exploring novel territory, since it never intersects itself *exactly*. (For an isolated dynamical system, such a perfect return would doom the system to an endless and unchanging periodicity.) But sometimes the novelty of one's activation-space position is not minor: it is substantial. Occasionally, one's redirected trajectory takes one out of a familiar and much-toured basin of dynamical attraction, over a local ridge of relative improbability, and into a new and quite different basin of attraction, a basin in which *all* sensory inputs of a certain kind now receive an importantly different regime of conceptual interpretation. If that new regime happens to yield an increased capacity for anticipating and manipulating one's environment, or some specific aspect of it, then it is appropriate to credit the creature with a new insight into the world. Though no *structural* changes have taken place in one's nervous system, such a case is still a clear instance of learning—we may call it *dynamical learning*.

The example of the misapprehended and then reapprehended friend is a deliberately mundane example of the process. That process will loom larger in importance when one appreciates that most cases of major scientific insight or so-called 'conceptual revolution' are also instances of dynamical learning. Consider, for example, Newton's famously sudden realization (the falling apple incident at Woolsthorpe) that the Moon's orbit is just another case of *projectile motion*, governed by the same laws as a stone thrown here on Earth. Consider Darwin's realization that the origin of distinct species might be owed to an entirely natural analogue of the *artificial selection* long practiced in animal husbandry. Consider Torricelli's insight that we are all living at the bottom of an *ocean of air*, and the test

of that hypothesis with a (steadily falling) barometer lugged manfully up a mountainside. Or consider Bernoulli's, Maxwell's, and Boltzmann's conjecture that a gas is just a cloud of *tiny ballistic particles* in rebounding collision with each other, and with the walls of whatever container might confine them.

In all of these cases, and in many others, the initial cognitive change effected lay not in the reconfiguration of anyone's synaptic weights—the changes at issue happened much too swiftly for that molasses-like process to provide the explanation. Rather, the change consisted in the *dynamical redeployment* of conceptual resources already in place, resources learned years ago and in other contexts entirely, resources *originally* learned by way of that slower synaptic process here found wanting. What was novel in the historical examples cited above was not the concepts deployed (*inertial projectile, selective reproduction, deep ocean, and swarm of ballistic particles*). These concepts were already in place. The novelty lay in the unusual target or circumstance of their deployment—namely, the *Moon, wild animals, the atmosphere, and confined gases*, respectively. In each case, an old and familiar thing came to be understood as an unexpected instance of a quite different category, a category hitherto employed in quite different circumstances, a category that makes new and systematic sense of the old phenomenon. To borrow a notion from biology, we are here looking at a variety of *cognitive exaptations*—that is, at cognitive devices initially developed in one environment that turn out to serve unusually well for a different purpose in a different environment.

As the reader may begin to surmise, the account of dynamical learning to be outlined in chapter 4 offers a new account of what theoretical hypotheses are, of what explanatory understanding consists in, and of what explanatory unification or “intertheoretic reduction” consists in. Its classical adversaries are the *syntactic* account of theories (“a theory is a set of sentences”), with its appropriately *deductive* accounts of both explanation and intertheoretic reduction, and the *semantic* view of theories (“a theory is a family of homomorphic models”), with its appropriately *model-theoretic* accounts of explanation and reduction. Both of these classical accounts are inadequate, I shall argue, especially the older syntactic/sentential/propositional account. Among many other defects, it denies any theoretical understanding whatever to nonhuman animals, since they do not traffic in sentential or propositional attitudes. That classical account is deeply inadequate for humans also, since it wrongly attempts to apply *linguistic* categories, appropriate primarily at the social level, to the predominantly *nonlinguistic* activities of individual brains. These primordial activities are better described in terms of an *unfolding sequence*

of *activation-vectors* than in the procrustean terms of sentential or propositional attitudes. There is ample room for sentential representations within the epistemological story to be presented below, and important work for them to do. But their proper home lies in the social world—in the shared space outside of the human brain, in the space of public utterances and the printed page—not inside the individual human head.

The semantic view of theories is mistaken also, but of these two classical adversaries, it comes much closer to the truth of the matter, as we shall see in due course. Despite its esoteric and, to many, rather daunting standard formulation in set-theoretic terms, a specific theory can also be relevantly characterized, on the semantic view, as a *single complex predicate*, a predicate that may be true of a specified domain-to-be-explained.<sup>7</sup> Thus, the Moon is a *Newtonian projectile*; the evolution of animal species is a *process shaped by random variation and selective reproduction*; the particles of a gas constitute a *classical mechanical system with perfectly elastic collisions*; and so on. (This predicative characterization misleads slightly, since it returns us to linguistic items such as predicates. By contrast, the canonical formulation regards the theory proper as being an *extensional* entity—the set-of-all-models of which that predicate is true. But never mind.) What the *neurosemantic* account of theories does is rather similar to the semantic account, although it locates theories firmly inside the head. Simply replace “single complex predicate” with “single prototype point in high-dimensional activation space,” and you have the outlines of the view to be defended below.

The semantic view of theories is still a minority view, however, even among epistemologists, and so the reader may find the parallel just drawn to be neither illuminating nor very compelling. Let me therefore draw another parallel, which comes closer still to the truth of the matter. The view of theories to be defended here is a neurally grounded instance of the tradition represented by Mary Hesse, Thomas Kuhn, Ronald Giere, William Bechtel, Nancy Cartwright, and Nancy Nersessian. That tradition focuses on the role of models, metaphors, paradigms, mechanisms, and idealized ‘nomological machines’ in the business of scientific theorizing.<sup>8</sup>

7. Cf. van Fraassen 1980.

8. This list is confined specifically to philosophers of science, but the theoretical tradition here drawn upon reaches beyond the confines of that particular subdiscipline. ‘Cognitive Linguistics,’ as represented by the work of linguists such as J. Elman, E. Bates, R. Langacker, G. Lakoff, and G. Fauconnier, is a different but salient branch of the same tree. ‘Semantic-Field Theory,’ as represented by the work of philosophers such as E. Kittay and M. Johnson, is another. And so is the ‘Prototype Theory’ of conceptual organization explored by an entire generation of psychologists.

In some respects, this is a highly diverse group. Famously, Kuhn's focus lay at the social level; Giere's and Nersessian's focus lies firmly at the psychological level; Bechtel is himself an occasional neuromodeler; and Cartwright tends to focus on the metaphysical nature of objective reality. But they are all united in seeing our scientific endeavors as dependent on the artful *assimilation* of complex and problematic phenomena to some special phenomena that are familiar, tractable, and already well understood. It is the neurocomputational *basis* of precisely such assimilative processes that I aim to capture as instances of dynamical learning. In the chapters that follow, we shall explore the *consequences* of that capture for a range of standard issues in epistemology and the philosophy of science.

One may still hear echoes of the semantic account of theories in the above list. After all, Giere counts his own view as one (heterodox) instance of that account, and Sneed (1971) and Stegmüller (1976) have tried to put Kuhn's story into model-theoretic dress, with perhaps some success. Well and good. I own a residual sympathy for the semantic account, alien though it may seem. And I agree with the first part of van Fraassen's (1980) important claim that the proper medium for representing the essence of scientific theories is not the formalism of classical *first-order logic*, but rather the formalism of *set theory*. My disagreement concerns the second half of his claim. The proper formalism, I suggest, is not set theory either (or not *only* set theory). The proper formalism is *vector algebra* and *high-dimensional geometry*. And the proper deployment of this formalism is the story of how vector coding and vector processing are realized in the vast network of the biological brain. This allows us, among other things, to bring a *dynamical* dimension into our account of human scientific theorizing—it is, after all, a causal process unfolding in time—a dimension largely or wholly absent from the original semantic accounts.

Before leaving this introductory discussion of dynamical or second-level learning, it is worth mentioning that one further major problem in the epistemological tradition is going to show up in a new and potentially more tractable guise, namely, the problem of the underdetermination of theory by evidence, and the status of broadly Realist versus broadly Instrumentalist interpretations of the enterprise of science. The underdetermination problem does not disappear—far from it—but it does assume a different form, and it heralds, I shall argue, a philosophical lesson somewhat different from that urged by either of these traditional adversaries. For one thing, as the reader may appreciate, the 'evidence' relation needs to be reconceived entirely, since the parties to it are no longer the 'theoretical' and the 'observational' sentences of the syntactic view,



nor the set-theoretic structures and their 'observational substructures' embraced by the semantic view. For another, we are going to find that underdetermination infects the domain of all possible *evidence* no less than the domain of all possible theories, with consequences we shall have to evaluate. And for a third, we shall rediscover another old friend—incommensurability (also reconfigured)—as we confront the practical infinity of neurocognitive alternatives potentially vying for human acceptance as the preferred vehicle of our global understanding. Despite these familiar bogeymen, it will be a recognizable version of Scientific Realism, as I see it, that best makes sense of the overall situation.

## 5 Collective Learning and Cultural Transmission

If it is important to distinguish a brain's *dynamical* adventures (in trying to apply its existing concepts to an ever-expanding experience of the world) from its more basic *structural* adventures (in slowly shaping a useful framework of concepts in the first place), then it is equally important to distinguish both of these originally individual activities from a third major level of learning—the level of *cultural* change and *collective* cognitive activity. For it is the institution of this third level of learning that most surely distinguishes the cognitive adventure of humans from that of any other species. This third-level activity consists in the cultural assimilation of individual cognitive successes, the technological exploitation of those successes, the transmission of those acquired successes to subsequent generations, and the ever-more-sophisticated *regulation* of individual cognitive activities at the *first two* levels of learning.

The existence and overwhelming importance of this third level of cognitive activity will be news to no one. But the proper characterization of that collectivized process is still a matter of substantial dispute. Is it the journey of *geist* toward complete self-consciousness, as Georg Hegel surmised? Is it the reprise of selective evolution at the level of linguistic items—a ruthless contest of selfish 'memes'—as Richard Dawkins has suggested? Is it the convergent march of science toward the Final True Theory, as some Pragmatists and Logical Empiricists dared to hope? Is it just the meandering and ultimately meaningless conflict between fluid academic fiefdoms competing for journal space and grant money, as some skeptical sociologists have proposed?

A wry answer might be that it is all of these. But a more considered and more accurate answer would be that it is none of the above. Nor will the true nature of this third-level process ever become clear until we appreciate

the manifold ways in which the various mechanisms of human culture serve to nurture, to regulate, and to *amplify* the cognitive activities of individual humans at the first two levels of learning, the levels we share with nonhuman animals.

As the offhand list of the preceding paragraph will attest, there is no shortage of philosophical theories about the structure, dynamics, and long-term future of cultural or third-level learning. They are many and various. But if the proposal here on the table—that the central *function* of these cultural mechanisms is the detailed exploitation and regulation of learning at the first two levels—then none of the familiar theories can hope to be anything more than incidentally or accidentally correct in their portrayals of the human epistemic adventure. For no epistemology or philosophy of science prior to the present period has had any interest in, or any clear conception of, these first two kinds of learning—namely, the generation of a hierarchy of prototype-representations via gradual change in the configuration of one's synaptic weights (first-level or 'structural' learning), and the subsequent discovery of successful redeployments of that hard-earned framework of activation-space representations, within novel domains of experience (second-level or 'dynamical' learning).

Indeed, those original and more basic levels of representation and learning have been positively *mischaracterized* by their chronic portrayal as just hidden, inward versions of the *linguistic* representations and activities so characteristic of cognitive activity at the third level. As noted earlier, Jerry Fodor is the lucid, forthright, and prototype perpetrator on this particular score, for his theory of cognitive activity is that it is explicitly language-like from its inception (see, e.g., Fodor 1975)—a view that fails to capture anything of the very different, *sublinguistic* styles of representation and computation revealed to us by the empirical neurosciences and by artificial neuromodeling. Those styles go wholly unacknowledged. This would be failure enough. However, the 'Language of Thought' hypothesis fails in a second monumental respect, this time, ironically enough, by *undervaluing* the importance of language. Specifically, it fails to acknowledge the extraordinary cognitive *novelty* that the invention of language represents, and the degree to which it has launched humankind on an intellectual trajectory that is impossible for creatures denied the benefits of that innovation, that is, for creatures confined to only the first and second levels of learning.

What I have in mind here is the following. With the emergence of language, the human race acquired a public medium that embodied—in its peculiar lexicon and in its accepted sentences—at least some of the acquired wisdom and conceptual understanding of the adults who share

the use of that medium. Not *all* of that acquired wisdom. Not by a long shot. But enough of it to provide an informed template to which the conceptual development and dynamical cognition of subsequent generations could be made to conform. These subsequent generations of language-learners and language-users are thus the heirs and beneficiaries of at least some of the cognitive achievements of their forebears. In particular, they do not have to sculpt a conceptual space entirely from scratch, as nonlinguistic animals do, and as prelinguistic humans must have. To the contrary, as human children learn their language, from their parents and from the surrounding community of conceptually competent adults, they can shape their individual conceptual developments to conform, at least roughly, to a hierarchy of categories that has already been proven pragmatically successful by a prior generation of cognitive agents.

At that point, the learning process is no longer limited to what a single individual can learn in a single lifetime. That collective medium of representation—language—can come to embody the occasional cognitive innovations of many different human individuals, and it can accumulate those innovations over hundreds and thousands of lifetimes. Most importantly, the conceptual template that the language embodies can slowly *evolve*, over historical periods, to express a different and more powerful view of the world than was expressed by its more primitive precursors.

It is important not to overstate this point. Almost *all* of anyone's acquired wisdom goes with him to the grave, including his inevitably idiosyncratic command of the resources of human language. There is no realistic hope of recording the specific configuration of anyone's  $10^{14}$  synaptic weights, and no realistic hope of tracing the dynamical history of anyone's brainwide neuronal activations, and thus no realistic hope of recreating, exactly, one's current brain state within the skull of another human. But one can hope to leave behind at least something of one's acquired understanding, if only a faint and partial digest thereof, through the communities of conversation and shared conceptual practice that one's speech behavior—whether live or printed—has helped to shape.

This said, it is equally important not to *understate* the importance of language. As a public institution whose current lexicon, grammar, and network of broadly accepted sentences are under no individual's exclusive personal control, a living language thereby constitutes a sort of 'center of cognitive gravity' around which individual cognitive activities may carve out their idiosyncratic but safely stable orbits. Moreover, as a cultural institution that long outlives the ephemeral individual cognizers that sequentially pass through it, a language embodies the incrementally added

wisdom of the many generations who have inevitably reshaped it, if only in small ways, during the brief period in which it was theirs. In the long run, accordingly, that institution can aspire to an informed structure of categories and conventional wisdom that dwarfs the level of cognitive achievement possible for any creature living outside of that transgenerational framework. Large-scale *conceptual* evolution is now both possible and probable.

Everyone will agree, of course, that a species with some mechanisms for historical recording can achieve more than a species with no such mechanisms. But I am here making a rather more contentious claim, as will be seen by drawing a further contrast with Fodor's picture of human cognition. On the Language of Thought (LoT) hypothesis, the lexicon of any *public* language inherits its meanings directly from the meanings of the innate concepts of each individual's innate LoT. Those concepts derive *their* meanings, in turn, from the innate set of causal sensitivities they bear to various 'detectable' features of the environment. And finally, those causal sensitivities are fixed in the human genome, according to this view, having been shaped by many millions of years of biological evolution. Accordingly, every normal human, at whatever stage of cultural evolution, is doomed to share the *same* conceptual framework as any other human, a framework that the current public language is therefore secondarily doomed to reflect. Cultural evolution may therefore *add* to that genetic heritage, perhaps considerably, but it cannot undermine it or supersede it. The primary core of our comprehensive conception of the world is firmly nailed to the human genome, and it will not change until that genome is changed.

I disagree. The lexicon of a public language gets its meanings not from its reflection of an innate LoT, but from the framework of broadly accepted or culturally entrenched sentences in which they figure, and by the patterns of inferential behavior made normative thereby. Indeed, the sublinguistic categories that structure any individual's thought processes are shaped, to a significant degree, by the official structure of the ambient language in which she was raised, not the other way around.

To raise an even deeper complaint, the meaning or semantic content of one's personal cognitive categories, whether innate or otherwise, derives not from any feature-indicating nomic relations that they may bear to the external world. Rather, it derives from their determinate place in a high-dimensional neuronal activation-space, a space of intricate and idiosyncratic similarity relations, a space that embodies a highly informed 'map' of some external domain of lasting properties. As we shall see in the next chapter, the correct account of first-level learning requires us to put aside

any form of atomistic, externalist, *indicator-semantics* in favor of a decidedly holistic, internalist, *domain-portrayal* semantics.

This means that both the semantic content of public languages and the semantic content of individual conceptual frameworks are not in the least bit 'nailed' to a fixed human genome. Both are free to vary widely, as a function of local epistemic circumstance and our individual and collective cognitive histories. But whereas each person's acquired conceptual framework is doomed to wink out after roughly three-score-and-ten years, the off-loaded, public structures of one's then-current language are fated to live on, in pursuit of an epistemic adventure that has no visible limits. Certainly this *third-level* world-representing process is not required to cleave to some Paleolithic or pre-Paleolithic conceptual framework somehow dictated by the human genome.

On the contrary, and given time, this third-level process opens the door to systematic reconstructions of our practical undertakings, and to systematic reconceptions of even the most mundane aspects of our practical and perceptual worlds. We can put out an initial public offering on the NYSE to support a company's plan to build a nuclear power plant to smelt aluminum from bauxite with megawatt applications of electric power. We can start a phone campaign to marshal registered Democrats to vote against the anti-choice initiative, Proposition 14, on the ballot of this year's state elections. We can aspire to write a thirty-two-bar tune in the key of G, based on the same chord sequence, appropriately transposed, as Harold Arlen's popular hit, "Stormy Weather." We can destroy the bacteria in a foul dishwashing sponge by boiling its residual water with klystron-generated microwaves in the countertop oven. We can marvel as we watch the spherical Earth rotate at fully fifteen degrees per hour, on its north-south axis, as the Sun 'sets' at the western horizon and the full Moon simultaneously 'rises' at the eastern horizon. Thoughts and undertakings such as these are simply beyond the conceptual resources of a Stone Age community of human hunter-gatherers. On the other hand, being the beneficiaries of their own third-level history, that Stone Age group has thoughts and undertakings—concerning such things as fire manipulation, food preparation, weapons technology, and clothing manufacture—that are equally inconceivable for the members of a baboon troop. We differ from the Stone Age humans by being at a very different rung on the long-term conceptual ladder that the institution of language provides. The baboons differ from us both, in having no such ladder to climb.

The institution of language is, however, only the first of many powerful mechanisms at this third and supraindividual level of learning. My hope,

in the last chapter of this book, is to provide a novel perspective on all of them by exploring their roles in the regulation, amplification, and transmission of human cognitive activities at the first two levels of learning, as conceived within the neurostructural and neurodynamical frameworks outlined above. The payoff, if there is one, lies in the multilevel coherence of the portrait of human cognition that slowly emerges, and in the fertility of that portrait in grounding novel explanations of various aspects of human cognitive activity found problematic by our existing epistemological tradition. In short, I shall try to tell a new story about some old problems.

## 6 Knowledge: Is It True, Justified Belief?

The general motivation behind my rejection of the "true, justified belief" conception of knowledge should be roughly apparent already, even from the discussion of this introductory chapter. Let us examine each element of that wrongly probative analysis.

First, reflect on the required vehicle itself: the belief. This is an overtly *propositional* attitude, which requires for its specification a *declarative sentence*, a unit of representation not encountered in human cognition at least until the innovation of spoken language. It is not an element of either the first-level or the second-level processes of learning, as characterized above. To restrict knowledge to the domain of literal belief is therefore to deny any factual knowledge at all to animals and to prelinguistic human children, whose cognitive activities do not, or do not yet, involve the representational structures peculiar to language—not externally, and not internally, either.

Such a narrow conception will also ignore or misrepresent the great bulk of any *adult* human's factual knowledge, which reaches substantially beyond what little she can articulate in the comparatively clumsy vehicle of a declarative sentence. The relative poverty of such vehicles shows itself rather dramatically in the unexpected failures of the classical or programming research agenda within the field of artificial intelligence. A genuinely successful artificial intelligence, within any domain, requires a substantial knowledge base on which to ground its cognitive activities. But a long list of 'accepted sentences' proved to be a hopelessly inefficient means of storing information, and in addition, an exponentially *expanding* problem when we confront the task of retrieving the relevant bits of information as they are (unpredictably) needed. If you thought that the belief

is, or deserves to be, the basic unit of factual knowledge, you had best think again.

Consider now the requirement of justification. That there is something awkwardly procrustean about this venerable condition becomes evident as soon as one considers again the case of factual knowledge in animals and prelinguistic human children. The fact is, the practice of justifying one's cognitive commitments is a practice that arises only among adult humans, only at the social level, and it typically involves a multipersonal negotiation concerning the status of some declarative sentence or other. One can and does rehearse such negotiations privately (and a good thing, too). But such rehearsals are confined to adult humans, and they reflect a practice whose origins and guiding prototypes lie in the public domain even there. In all, justification is a business that has its natural home at the social or third level of human learning, and its focus, naturally enough, is on representational vehicles—declarative sentences—that we have already found unequal to the task of being the basic units of human and animal cognition. I hasten to add that I have no desire in this book to denigrate the importance of our public institutions for the evaluation and justification of factual claims. Quite the contrary. But the “justified-true-belief” formula at issue wrongly imposes a condition that is drawn from and appropriate to a third-level learning process as a condition on knowledge *in general*.<sup>9</sup> It is wrong because the justification requirement has no application whatever to first-level knowledge, and only problematic or metaphorical application to the results of second-level learning. How do I go about ‘justifying’ to you, or even to myself, a specific trajectory just carved out in my neuronal activation space? Not in any usual way. Especially when those spontaneous trajectories are only feebly under my control in the first place.

9. This second condition has proven awkward even in advance of the more fundamental complaints voiced above. For example, one's unfolding perceptual beliefs typically count as knowledge, but the requirement that each one of them be *justified*, second by second, is difficult even to make sense of. Alvin Goldman (1986) has proposed, plausibly, that this traditional condition should therefore be replaced by the weaker but more apt requirement that the belief at issue have been produced by a “generally reliable mechanism.” This is a step forward, but only a step. My complaint about the vehicles themselves remains unaddressed. And one wonders about those occasional cases where the generally reliable mechanism positively *malfunctions*, but, by the sheerest accident, happens to produce a true belief anyway. This meets the new condition, but it doesn't look like knowledge.

Finally, what about the requirement of truth? Surely, we are on firm ground in demanding truth as a condition on knowledge, at least if we restrict our attention to factual knowledge, as opposed to procedural knowledge or motor skills. But perhaps that ground is not so firm. An opening response to this plea for truth is that the neurocomputational perspective here to be explored makes it unexpectedly difficult to draw any clear or principled distinction between these two superficially different kinds of knowledge—factual knowledge versus practical skills—the one kind, truth-bearing; the other kind, not. A distinction that initially seems plain to uninformed common sense tends to disappear entirely when one confronts the realities of neuronal coding and world navigation, as we shall see.

A second and equally pregnant response concerns the sense of “truth” we are here supposed to impose. In its primary and literal sense it is a feature, once more, of declarative sentences. And its canonical explication in logical theory, the model-theoretic account of Alfred Tarski, leans entirely on the structural elements and grammatical features of human language.<sup>10</sup> That is an explication we have all learned to admire. But what are we supposed to say of unfamiliar representational vehicles such as a hierarchically structured, high-dimensional activation space—as a *conceptual framework* turns out to be, at bottom? And what are we supposed to say of representational vehicles such as a  $10^6$ -element neuronal activation vector, which is what a *perceptual representation* turns out to be? These vehicles are not the sorts of things to which the notion of Tarskian truth even applies. And yet, as we have seen, they constitute the bulk of any creature’s acquired factual knowledge. No doubt they can have, or lack, sundry representational virtues. But Tarski-style truth would seem not to be among them.

The “justified-true-belief” formula has, of course, received extensive critical attention in the literature. But the worries have focused almost exclusively on Gettier’s (1963) illustration that these conditions are not yet collectively *sufficient* for knowledge. Most of the resulting research has bent its efforts in search of a *fourth* condition to patch up the perceived hole.

The complaint leveled here is different and deeper. It is that not one of these three familiar conditions is even individually *necessary* for knowledge. Indeed, the “justified-true-belief” approach is misconceived from the

10. Strictly, it leans on the structural features of a very tiny and highly regimented *fragment* of human language, namely, the narrow formalism of the first-order predicate calculus. But the account is still salutary. Let us give credit where credit is due.



outset, since it attempts to make concepts that are appropriate only at the level of cultural or language-based learning do the job of characterizing cognitive achievements that lie predominantly at the *sublinguistic* level. There are indeed dimensions of success and failure that operate at these more basic levels. But their articulation will require that we temporarily put aside the peculiar structures of language and address those deeper forms of representation on their own terms. This need not require that we simply toss away the notion of truth, as some may fear. But it will require that we reconceive it, and generalize it, in some highly specific ways. Knowledge, then, is *not* justified true belief, at least, not in general. As we shall see in the chapters to follow, it is something rather more interesting than that, and a good deal more precious.

Before closing this introduction, let me try to summarize. We are looking at three distinct learning processes at three quite different levels. At the first level, we find a process of structural change, primarily in the micro-configuration of the brain's  $10^{14}$  synaptic connections. The product of this process is the metrical deformation and reformation of the space of possible activation patterns across each receiving neuronal population. This product is a configuration of *attractor* regions, a family of *prototype* representations, a hierarchy of *categories*—in short, a *conceptual framework*. In biological creatures, the time-scale of this unfolding process is no less than weeks, and more likely months, years, and even decades. It is slow, even molasses-like. It never ceases entirely, but by adulthood the process is largely over and one is more or less stuck with the framework it has created.

At the second level, we find a process of dynamical change in one's typical or habitual modes of neuronal activation, change that is driven not by any underlying synaptic changes, but by the recent activational history and current activational state of the brain's all-up neuronal population. Bluntly, the brain's neural activities are *self-modulating* in real time, thanks to the recurrent or feed-backward architecture of so many of its axonal projections. The product of this process is the systematic redeployment, into novel domains of experience, of concepts originally learned (by the first or basic-level process discussed above) in a quite different domain of experience. It involves the new use of old resources. In each case, this product amounts to the explanatory reinterpretation of a problematic domain of phenomena, a reinterpretation that is subject to subsequent evaluation, articulation, and possible rejection in favor of competing reinterpretations. The time-scale of such redeployments (in contrast to their subsequent evaluation and development) is much shorter than that of structural or basic-level learning—typically in the region of milliseconds

to hours. This is the process that comprehends sudden gestalt shifts, 'eureka' effects, and presumptive conceptual revolutions in an individual's cognitive take on some domain of phenomena.

At the third level, we find a process of cultural change, change in such things as the language and vocabulary of the community involved, its modes of education and cultural transmission generally, its institutions of research and technology, and its techniques of individual and collective *evaluation* of the conceptual novelties produced at the first two levels of learning. This is the process that most decisively distinguishes human cognition from that of all other species, for the accumulation of knowledge at this level has a time-scale of decades to many thousands, even hundreds of thousands, of years. Its principal function is the ongoing regulation of the individual and collective cognitive activity of human brains at the first two levels of learning.

A cautionary note should be entered right here in the opening chapter. A close examination of the processes at all three levels will reveal a fractionation into sundry subprocesses and substructures. Most obviously, the third-level process is going to display a hierarchy of interlocking regulatory mechanisms, depending on the particular human culture, and on the particular stage of its history, that we choose to examine. Less obviously, but no less surely, the first-level process is a knit of architectural, developmental, neurochemical, and electrophysiological activities, which neuroscientists and neuromodelers have only begun to unravel. Similarly, the dynamical or second-level process will reveal its own variations and intricacies. The tripartite division sketched above is a deliberate idealization whose function is to provide a stable framework for more detailed exposition at each level, and for an exploration of the major interactions between the three levels.

If plainly acknowledged, there need be no deception in this. Useful idealizations are the life-blood of scientific understanding. What remains to be determined is: just *how* useful is the idealization at issue? Without more ado, let us now try to find out.