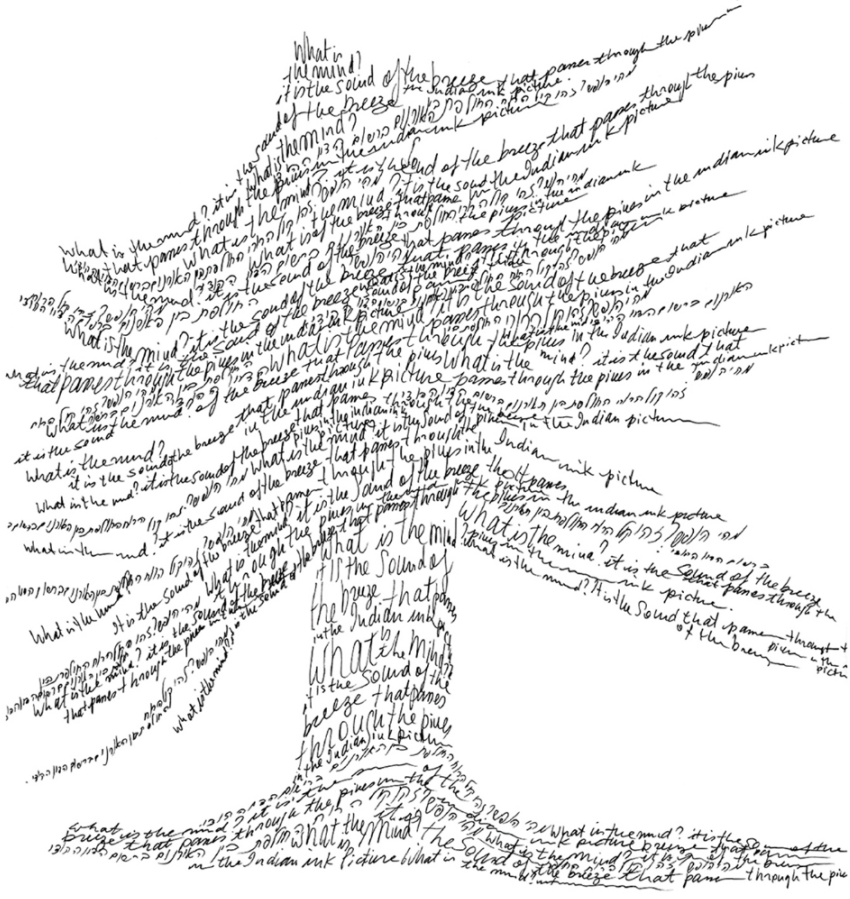
THE EVOLUTION OF THE SENSITIVE SOUL

LEARNING AND THE ORIGINS OF CONSCIOUSNESS



SIMONA GINSBURG AND EVA JABLONKA

ILLUSTRATIONS BY ANNA ZELIGOWSKI

The Evolution of the Sensitive Soul

Learning and the Origins of Consciousness

Simona Ginsburg and Eva Jablonka

with illustrations by Anna Zeligowski

The MIT Press
Cambridge, Massachusetts
London, England

This book was set in Stone by Westchester Publishing Services. Printed and bound in the United States of America.

Library of Congress Cataloging-in-Publication Data.

Names: Ginsburg, Simona, author. | Jablonka, Eva, author.
Title: The evolution of the sensitive soul : learning and the origins of consciousness / Simona Ginsburg and Eva Jablonka.
Description: Cambridge, MA : MIT Press, 2019. | Includes bibliographical references and index.
Identifiers: LCCN 2018019048 | ISBN 9780262039307 (hardcover : alk. paper)
Subjects: LCSH: Consciousness. | Consciousness--Physiological aspects. | Evolution (Biology)
Classification: LCC BF311 .G5325 2019 | DDC 153--dc23
LC record available at https://lccn.loc.gov/2018019048

10 9 8 7 6 5 4 3 2 1

Contents

Preface

There is nothing more intimate and cherished, and nothing more elusive, than subjective experiencing. In this book we use an evolutionary approach to explore the biological basis of such experiencing or, as it is usually called, "consciousness." We argue that consciousness emerged in the context of the evolution of learning, and we maintain that by figuring out how the evolutionary transition to subjectively experienced living occurred, we can gain an insight into the nature of this mode of being. We are therefore interested in consciousness-as-we-know-it—in animal consciousness, the only type of consciousness that we know exists, rather than in hypothetical machine consciousness. This is reflected in the title of our book: the "sensitive soul" is the apt term used by Aristotle to describe the ability of animals to subjectively experience percepts and feelings.

We build our evolutionary account on what we have learned from studies of neurobiology, cognitive science, animal learning, philosophy of mind, and evolutionary biology. Some of our interpretations and conclusions are likely to be proven wrong, but we think that the framework we use—embedding subjective experiencing in the biological, evolutionary processes of neural animals rather than in medium-free, multirealizable computations—will remain useful.

Our approach focuses on the *evolutionary transition to minimal consciousness* and is based on both the teleological framework developed by Aristotle twenty-four hundred years ago and on our current understanding of biological evolution, the twenty-first-century developmentally informed theory that has its origins in the nineteenth-century work of Jean-Baptiste Lamarck, Charles Darwin, and their followers. It is a theory that is constantly updated as new developments occur in molecular and developmental biology, paleontology, ecology, and other biological disciplines. Evolutionary theory is, for us, the most general framework for understanding the biological world. It is a conceptual bottleneck through which any theory of

life and mind must pass. If a biological (or psychological, or sociological) theory fails to pass through this bottleneck, it is likely there is something seriously wrong with it.

Because evolution is so central to biological investigations, it is natural to assume that it has been incorporated into the framework of consciousness studies, both as a yardstick for measuring the validity of new theories and as a source of insights. But in fact, until very recently there has been a strange lacuna in the field. Although most scientists and philosophers who write about consciousness are now convinced that it is a biological process that is a product of evolution, its evolutionary origins are rarely central to their discussions. Indeed, it was not until the first decade of the twenty-first century, after more than one hundred years of academic silence about the evolution of consciousness, that serious attempts to understand it began to emerge again. It seems that the skepticism about the possibility of conducting a scientific study of consciousness that prevailed until the 1990s, coupled with conceptual difficulties, were important reasons for this neglect.

For an evolutionary account of the origins of experiencing, biologists must agree on what the first type of subjectively experiencing animal was like. This means that one has to try and characterize minimal consciousness or, alternatively, find some good criteria (or markers) that indicate that it is in place. This is a difficult task because it is not clear how to identify consciousness in animals very different from us. Nevertheless, such difficulties have not stopped evolutionary biologists from trying to solve comparable problems. Inquiries into the origin of life presented similar conceptual and theoretical difficulties, but scientists did not shy away from the problem, and an evolutionary approach proved to be extremely fruitful. Because questions about the origin of life, the origin of minimal consciousness, and the origin of human abstract values are all concerned with the emergence of new types of goal-directed systems and confront similar conceptual and methodological challenges, we modeled our approach to the evolution of minimal consciousness on the best-developed research program of the three, research on the origin of life.

The benefits of the evolutionary, origin-focused approach are obvious: if we can locate when and how in evolutionary history the transition from an organism that lacked consciousness to one with minimal consciousness occurred, it becomes possible to explore the processes and organizational principles involved without being misled by later derived dissociations and integrations that mask the fundamental properties of subjective experiencing. We approach this task by characterizing the essential features and dynamics of consciousness and singling out a diagnostic, tractable,

biological capacity that was necessary for its presence—the evolutionary transition marker of consciousness. Studying the evolutionary transition to conscious animals by tracing the evolutionary history of this diagnostic capacity allows us to identify and reconstruct (i.e., reverse engineer) the type of system of which it is part. Our focus is therefore on *minimal animal consciousness*—human reflective consciousness is not the subject of this book, although we believe (and argue) that our approach has important implications for its study.

As we have discovered, trying to understand minimal consciousness is a mammoth undertaking. When we started our work, we were blissfully ignorant of its real dimensions. This was a good thing, for had we been aware of the magnitude of the task, we would probably not have proceeded. As we became immersed in the project, we realized that we (and our prospective readers) needed a lot of background knowledge in order to have a foundation on which to build an evolutionary account. The result is this rather fat book. The size of the book led us to divide it into two parts, with part I (chapters 1–5) providing the historical, biological, and conceptual foundations on which we build, and part II (chapters 6–10) developing the evolutionary arguments. We know that long science books are not fashionable, and we are aware that some people may not be interested in all of the background chapters. For example, people who are wary of history can skip chapter 2, those who are allergic to philosophy can skim over chapter 4, and nonbiologists may hum through the neurobiological and biochemical details. We do hope, however, that we have managed to convey some of the excitement and humility that we felt as we researched and wrote our book—and that Anna Zeligowski's illustrations will engage our readers' aesthetic sensibilities, deepening the sense of inquiry and wonder that the subject elicits.

Figure Credits

1.2. Yeshayahu Leibowitz. Portrait by Bracha L. Ettinger ©. 14

1.3. The autopoietic system. Figure 1 (Views on agency), p. 15 in E. Di Paolo, "Extended Life," *Topoi* 28 (2009): 9–21. Reproduced by permission of Springer. 21

2.1. Jean-Baptiste Pierre Antoine de Monet, Chevalier de Lamarck. Reproduced by kind permission of Wellcome Library, London. 46

2.4. Charles Darwin and his son. Reproduced by kind permission of Cambridge University Library. Classmark: Dar 225:129. 64

3.6a. Neural selectionist schemes. Loosely based on figure 65, p. 303 in J.-P. Changeux, *L'Homme neuronal* (Paris: Fayard, 1983). 121

3.6b. Neural selectionist schemes. Figure 5, p. 40 in G. M. Edelman, *Wider than the Sky: The Phenomenal Gift of Consciousness* (New Haven, CT: Yale University Press, 2004). Reproduced by permission of Yale University Press. 121

3.7. Reentrant pathways leading to primary consciousness. Figure 1, p. 5522 in G. M. Edelman, "Naturalizing Consciousness: A Theoretical Framework," *Proceedings of the National Academy of Sciences USA* 100 (2003): 5520–5524. Reproduced by permission of The National Academy of Sciences, U.S.A. 125

3.11. The global neural workspace (GNW) model. Figure 1, p. 14530 in S. Dehaene, M. Kerszberg, and J.-P. Changeux, "A Neuronal Model of a Global Workspace in Effortful Cognitive Tasks," *Proceedings of the National Academy of Sciences USA* 95(24) (1998): 14529–14534. Adapted by permission of The National Academy of Sciences, U.S.A. 131

4.5. The "reality space" model. Figure 5, p. 73 in B. Merker, "Consciousness without a Cerebral Cortex: A Challenge for Neuroscience and Medicine," *Behavioral and Brain Sciences* 30(1) (2007): 63–81. Adapted by permission of Cambridge University Press. 178

5.5. Lateral views of the brains of some extant vertebrates. Figure 1, p. 744 in R. G. Northcutt, "Understanding Vertebrate Brain Evolution," *Integrative and Comparative Biology* 42 (2002): 743–756. Reproduced by permission of Oxford University Press on behalf of The Society for Integrative and Comparative Biology. 210

5.9. Panskepp's evolutionary view of the emotional organization of the brain. Figure 3, p. 6 in J. Panksepp, "Cross-Species Affective Neuroscience: Decoding of the Primal Affective Experiences of Humans and Related Animals," *PLoS One* 6(9) (2011): e21236. Reproduced by permission (unrestricted open access). 217

II.1. An imaginary exploration-stabilization system. Figure 11.8, p. 451 in E. Jablonka and M. J. Lamb, *Evolution in Four Dimensions*, 2nd ed. (Cambridge, MA: MIT Press, 2014). Reproduced by permission of The MIT Press. 246

6.4b. Two scenarios of evolving neurons. Loosely based on Figure 2a, p. 918 in G. Jékely, "Origin and Early Evolution of Neural Circuits for the Control of Ciliary Locomotion," *Proceedings of the Royal Society B* 278 (2011): 914–922. 264

7.1. The flight simulator constructed to study conditioning. Adapted (from photographs) by permission of B. Brembs. 299

7.2. Instrumental conditioning of pigeons. Figure 4.1, p. 47 in S. F. Walker, *Learning and Reinforcement* (Essential Psychology Series) (London: Methuen, 1976). Adapted by permission of Routledge. 305

7.5 a–c. Sensitization in *Aplysia*. Figure 1, p. 2 in M. Mayford, S. A. Siegelbaum, and E. R. Kandel, "Synapses and Memory Storage," *Cold Spring Harbor Perspectives in Biology* 4 (2012): a005751. Modified by permission of Cold Spring Harbor Laboratory Press. 310

7.5d. Classical conditioning in *Aplysia*. Figure 2, p. 4 in R. D. Hawkins and J. H. Byrne, "Associative Learning in Invertebrates," *Cold Spring Harbor Perspectives in Biology* 7(5) (2015): a021709. Modified by permission of Cold Spring Harbor Laboratory Press. 310

7.7. Cell memory and cell heredity through self-sustaining loops. Figure B.3, p. 418 in S. B. Gissis and E. Jablonka, eds., *Transformations of Lamarckism: From Subtle Fluids to Molecular Biology* (Cambridge, MA: MIT Press, 2011). Reproduced by permission of The MIT Press. 317

7.8. Cell memory and cell heredity through three-dimensional templating. Figure B.4, p. 420 in S. B. Gissis and E. Jablonka, eds.,

Transformations of Lamarckism: From Subtle Fluids to Molecular Biology (Cambridge, MA: MIT Press, 2011). Reproduced by permission of The MIT Press. 318

7.9. Cell memory and cell heredity through the maintenance of DNA methylation patterns. Figure B.1, p. 415 in S. B. Gissis and E. Jablonka, eds., *Transformations of Lamarckism: From Subtle Fluids to Molecular Biology* (Cambridge, MA: MIT Press, 2011). Reproduced by permission of The MIT Press. 319

7.10. Three mechanisms of cell memory and cell heredity mediated by small regulatory RNAs. Figure B.2, p. 417 in S. B. Gissis and E. Jablonka, eds., *Transformations of Lamarckism: From Subtle Fluids to Molecular Biology* (Cambridge, MA: MIT Press, 2011). Reproduced by permission of The MIT Press. 320

7.13. Phylogenetic relationships of major animal phyla. Figure 1, p. R877 in M. J. Telford, G. E. Budd, and H. Philippe, "Phylogenomic Insights into Animal Evolution," *Current Biology* 25(19) (2015): R876–R887. Reproduced by permission of Elsevier. 330

8.1 a, c. Interactions in the primate ventral visual system. Figure 1, p. 2 in M. Manassi, B. Sayim, and M. H. Herzog, "When Crowding of Crowding Leads to Uncrowding," *Journal of Vision* 13(13) (2013): 10. Adapted by permission of The Association for Research in Vision and Ophthalmology. 358

8.6. A comparison between some midbrain and higher brain structures. Figure 3, p. 4 in T. Mueller, "What Is the Thalamus in Zebrafish?," *Frontiers in Neuroscience* 6 (2012): 64. Adapted by permission of Thomas Mueller. 387

8.7. A section through the insect brain. Reproduced by permission of www.cronodon.com. 388

8.8. The octopus brain. Figure 23, p. 103 in J. Z. Young, *A Model of the Brain* (Oxford: Clarendon Press, 1964). Reproduced by permission of Oxford University Press. 389

9.2. Phylogenetic relationships among major animal phyla. Figure 1, p. R877 in M. J. Telford, G. E. Budd, and H. Philippe, "Phylogenomic Insights into Animal Evolution," *Current Biology* 25(19) (2015): R876–R887. Reproduced by permission of Elsevier. 408

10.1. Comparison between the structures of the ocelloid and the vertebrate eye. Figure S1 (A and F) in S. Hayakawa, Y. Takaku, J. S. Hwang,

T. Horiguchi, H. Suga, W. Gehring, K. Ikeo, and T. Gojobori, "Function and Evolutionary Origin of Unicellular Camera-Type Eye Structure," *PLoS One* 10(3) (2015): e0118415. Adapted by permission (unrestricted open access). 459

10.2. Richard Semon. Reproduced by permission of University of Zurich, Archives of the history of medicine, PN 31.01.1066:65. 464

Introduction to Part I Rationale and Foundations

> When the answer cannot be put into words, neither can the question be put into words. The riddle does not exist. If a question can be framed at all, it is also possible to answer it.
>
> —Wittgenstein 1922, 6.5

In this book we try to answer the question: How did minimal animal consciousness originate during animal evolution? We argue that this is an answerable question if one can uncover a capacity that is a good marker of the evolutionary transition from preconscious to conscious animals. This can be, we maintain, an Archimedean point to explore the biological nature of consciousness, of sentience. We present our evolutionary transition–oriented account of consciousness in part II of the book. Part I provides background information about the explanatory framework we use, the history of the evolutionary approach to mentality, and current neurobiological and philosophical approaches to consciousness. Part I ends with a chapter that may be viewed as a bridge to part II, as it deals with present ideas, including our own, about the distribution of consciousness in the animal world. The background chapters, particularly 2 (history) and 4 (philosophy), are somewhat idiosyncratic: since ours is an evolutionary transition–oriented perspective, we highlight those facets of history and philosophy that have an evolutionary orientation or that make evolutionary sense. We do not, therefore, engage with philosophers who hold dualistic positions with regard to the mind-body problem or with those who do not regard consciousness as the product of biological evolution.

But how can one study the biological nature and evolution of minimal consciousness? Consciousness is a goal-directed process, and the study of subjective experiencing, which is the hallmark of animals with "sensitive souls," has conceptual challenges similar to those presented by other

1 Goal-Directed Systems: An Evolutionary Approach to Life and Consciousness

How can consciousness be studied? Our point of departure is Aristotle's "soul," the organizational dynamics of living beings, and its different manifestations in different types of organisms: the "nutritive and reproductive soul," which involves self-maintenance and reproduction and is present in all living things; the "sensitive soul," which is equated with the living organization of sentient, subjectively experiencing beings; and the "rational soul," which is special to reasoning humans. Our main interest is in the sensitive soul and we ask: Is it possible for scientists to study the sensitive souls of bees, of dogs, and of humans? More generally, we inquire whether it is possible to relate teleological and mechanistic causations or whether there is an unbridgeable explanatory gap between them. We start with the life gap—the gap between inanimate matter and animate beings with nutritive souls—and the study of the origin of life, a research program with a long and successful history. Although we cannot yet construct living organisms from inanimate matter, the evolutionary transition to a living organization is no longer seen as a mystery. We then ask whether the investigations of the life gap can illuminate the qualia gap—the enigma of how living matter gives rise to subjective experiencing, to sensitive souls. Adopting Daniel Dennett's evolutionary hierarchy of goal-directed systems, which parallels Aristotle's teleological hierarchy of souls, we suggest that an evolutionary, transition-oriented approach not only may lead to biological insights but also may settle some thorny philosophical problems.

Life and consciousness seem to be the very core of what it means to be a sentient biological being. It is therefore not surprising that both life and consciousness are notoriously difficult to define and analyze and that they have long frustrated the philosophers and biologists who have attempted to account for them in naturalistic terms. The Cartesian view that the living body is a material machine, whereas the mind is nonmaterial, deeply influenced Western thought from the seventeenth century onward and gave rise to the infamous mind-body problem. However, by the dawn of the twentieth century, life and consciousness—body and mind—seemed to

many people to be intimately related. This was not just because consciousness or subjective experiencing (we shall be using these terms interchangeably) could be understood as a product of the evolution of living organisms but also because life and consciousness were both seen as ongoing, self-organizing processes. This communality is probably why Henri Bergson, for example, equated the two notions, suggesting that life is creative becoming and charged with consciousness:

> The evolution of life, from its early origins up to man, presents to us the image of a current of consciousness flowing against matter, determined to force for itself a subterranean passage, making tentative attempts to the right and to the left, pushing more or less ahead, for the most part encountering rock and breaking itself against it, yet in one direction at least succeeding in piercing its way through into the light. That direction is the line of evolution which ends in man. (Bergson 1920, pp. 27–28)

There is beauty in this "current" metaphor, as in so much of Bergson's prose, but although the continuity between life and consciousness is self-evident because all known conscious beings are alive, we do not endorse Bergson's position that consciousness and life are identical. It is not surprising that this nebulous view did not lead to a scientific approach to the subject. On the contrary, it reinforced the generally shared feeling at that time that the nature of life and consciousness would remain forever elusive, forever inaccessible to scientific inquiry.[1] This impasse started to be overcome later in the twentieth century, however, and today the nature of life, though recognized as a very difficult problem, is no longer seen as scientifically impenetrable. With consciousness this is not yet the case, but more and more biologists, psychologists, and philosophers believe that the increasing understanding of the nervous system, the insights into the biology of cognition and affect, the progress in computational biology and in brain imaging, and the advances in the naturalistic philosophy of mind all point in the same hopeful direction.

We have already indicated that alongside the term "consciousness," which is widely used and therefore unavoidable, we will be using the term "subjective experiencing." "Subjective experiencing" is a self-explanatory, intuitive term encompassing the paradigmatic processes that we identify with conscious experiences: it refers to what happens to us and in us when we have not eaten for a few days, when we trip over a rock and sprain our ankle, when we taste a ripe banana, when we watch the starry night sky in the desert, when we hold and smell a baby, when our beloved mother dies, when we are attacked at night, when we have a nightmare, or when we

solve a difficult mathematical problem. It is also what disappears when we fall into dreamless sleep or into a coma. People often refer to humans and animals who subjectively experience as "sentient," and that is how we use the words "sentient" and "sentience" here.

In addition to the intuitive appeal of "subjective experiencing," we like the term because "experiencing" is a verbal noun, so the dynamic nature of the processes involved is explicit. We realize that this casual characterization of subjective experiencing may irritate some of our readers, and we can only repeat the argument of Patricia Churchland, who maintained that starting a discussion of consciousness by defining it is not necessary because "we use the same strategy here as we use in the early stages of any science: delineate the paradigmatic cases, and then bootstrap our way up from there."[2] However, for those who want more, we offer a provisional characterization (not a definition), which we will expand on later. We suggest that consciousness is not a property or a capacity of a system such as having sight, nor is it a processes such as metabolism. We see subjective experiencing *as a mode of being* that involves activities that generate temporally persistent, dynamic, integrated, and embodied neurophysiological states that ascribe values to complex stimuli emanating from the external world, from the body, and from bodily actions. Although perceptual consciousness (e.g., seeing a red poppy) and affective consciousness (e.g., feeling pain or fear) can be distinguished, and it seems that the first (perception) can occur in the absence of the second (feeling), they are a unified aspect of experience, something that is evident when the evolutionary history of consciousness is addressed. Inevitably, this characterization is, at this point, rather opaque and clumsy, and certainly it is lacking in poetry, but we hope that as we proceed some flesh will be put on its dry bones.

Our term "subjective experiencing" is thus equivalent to both sentience and consciousness, but consciousness researchers have qualified the latter term in many different ways, some of which are overlapping and often confusing.[3] It is important to stress here that human consciousness, which laypeople usually associate with the term "consciousness" and which we discuss from an evolutionary perspective in the last chapter, is not the main topic of this book. Our book is about the origins and evolution of sentience, of *minimal animal consciousness*—the ability to have basic subjective experiences—rather than the ability to reflect about those subjective experiences, which seems to be the peculiar gift and curse of humans.

Interpretive problems plague not only the notion of consciousness but also the related concepts of awareness, mind, soul, self, mentality, and cognition. "Awareness" usually refers to a state of wakeful attention and precludes

the subjective, nonreflective experiencing that occurs when we learn implicitly or when our thoughts are just roaming, whereas "self-awareness" is similar to self-consciousness but commonly has more affective connotations, as in shyness. "Mind" and "cognition" are usually, though inconsistently, used in a very broad way. Cognition, in the broad sense, refers to any information processing that involves interactions between sensors and effectors; it is used not only when referring to all types of neural processing in animals but also for describing processes involving flexible sensor-effector interactions and signal transduction networks in nonneural organisms like bacteria, paramecia, fungi, and plants.[4] The commonly used terms "mind" and "mentality" usually refer to intellectual faculties and to thinking but are sometimes used more generally—for example, as in "the mind-body problem." "Spirit" is used for a nonmaterial, mental, psychological "something" that is separated from the body, while "self" refers to a subjectively felt distinction between the subject and the world.

The term "soul" is also ambiguous, referring, in most of the monotheistic theological texts that followed the rise of Christianity, to something that is usually separated from the body; something that is responsible for morality and that remains after death. This usage, however, was not universally shared in the ancient world. In Genesis 1:25, God is said to have created the animals. All animals, beginning with the creatures swarming in the seas and ending with man, are what the Hebrew biblical text calls a "living soul" (נפש חיה) or an "animal soul" (חיה is both "animal" and "living" in Hebrew), rendered in the King James translation as "the moving creature that hath life." Significantly, "living soul" is not an attribute attached to plants, which were created much earlier, on the third day; plants grow and reproduce after their own kind but are not said to be living souls, so a clear distinction between plants and animals is made, with only the latter being ensouled. Some pre-Socratic philosophers were more liberal, granting both life and soul to plants and even to magnets. Nevertheless, both the ancient Hebrews and the pre-Socratic philosophers seem to have regarded the soul as an intrinsic part of the entity in question, in contradistinction to Plato and his school, who attributed an autonomous existence to the soul and regarded it as a separable entity. This latter notion had a profound influence on theological and philosophical reflections in the Western world.

We have used the problematic term "soul" in the title of this book. However, our usage will not follow the Platonic or theological traditions in the Western world. We use the term as a tribute to Aristotle, the greatest-ever philosopher of living things and the founder of the life sciences, and to his great treatise *De Anima* (*On the Soul*). In *De Anima* Aristotle carved the living

it is the person who can be said to be unconscious; her cells are neither conscious nor unconscious—they are nonconscious. Similarly, living cells taken from an animal such as a human may be grown in culture, differentiate into neurons, and form interesting neural networks. However, to say that neural networks are conscious and have subjective experiences would imply that they can be unconscious, and this seems to us to be a completely vacuous statement. Since we are interested in living beings who can *lose consciousness*, we cannot attribute sentience or consciousness to a motile bacterium or a ripe tomato.[9] The distinction between conscious and unconscious makes no sense with bacteria and tomatoes. We therefore agree with Aristotle that plants have splendid nutritive souls, but they *do not have* (losable) sensitive souls like those of humans and cats.

But are complex adaptive behaviors and nervous systems, the hallmarks of animals, sufficient for rendering them conscious? It is clear that complex adaptive responses can occur without subjective experiencing. We can build robots that exhibit adaptive behaviors, but these robots are not deemed sentient because they do not satisfy the list of characteristics considered necessary and sufficient for minimal consciousness.[10] Similarly, cells form complex networks in petri dishes, and they can also learn. Organized neural systems, such as severed spinal cords, can exhibit learning too, but as has been found with some unfortunate victims of terrible accidents, spinal learning is not associated with subjective experiencing, so such a network cannot confer consciousness.[11] Adaptive behaviors and learning, even when involving neurons, are therefore not sufficient criteria for identifying consciousness. If we do not want to render the distinction between neutrally instantiated conscious and unconscious states unintelligible, we have to qualify the kind of nervous system and the kind of neural dynamics that generate subjective experiencing.

Although they are not sufficient, there are reasons for thinking that in the biological world, a *nervous system* and a *brain* are necessary for subjective experiencing to occur. First, since it is the whole organism that experiences rather than only a part of an organism, subjective experiencing must involve a *systemic reaction*. The response to a stimulus must be integrated with the overall state of the organism in a way that preserves the specificity of the response in terms of its location, modality, or strength yet leads to whole-organism subjective experiencing and a particular coordinated action, which depend on multiple reciprocal interactions. Second, in a multicellular body, different types of stimuli, in different locations and acting on different sense organs, must be able to elicit integrated yet specific subjective experiencing—something that requires both elaborate connectivity and a

common "language" for communication. Third, as integration occurs, the stimulus and the integrated state must persist in order for feeling to occur, and this seems to require both spatial convergence and temporal synchronization in a center of communication. The rapid transmission of signals by electrical impulses, neural connectivity, temporal binding, and a brain in which signals are combined and where rapid and persistent feedback loops can occur fulfill all of these requirements. Since bacteria and tomatoes do not have such central transmission-integration systems, most people would agree they seem to be unendowed with the phenomenal consciousness of animals—with felt needs and perceptions.

For all these and other reasons that we shall come to in later chapters, we follow Aristotle and maintain that only animals (and, as we argue later, not *all* animals) have a sensitive soul, the second level of living organization in Aristotle's soul hierarchy. Of course, organisms endowed with rationality (Aristotle's third level of soul) also have sensitive souls, but as we suggest in chapter 10, their sensitive souls are radically different from those of nonrational animals because the evolution of rationality involved profound changes in subjective experiencing.

Although we adopt the Aristotelian hierarchy and find Aristotle's intrinsically teleological stance and his emphasis on the functional unity of the organism inspiring and useful,[12] our evolutionary approach to consciousness is, in a way, very non-Aristotelian because the historical-evolutionary perspective came into use only two hundred years ago. Aristotle's nonevolutionary approach does not tell us much about the temporal changes and gradations between the different soul levels. Is there continuity between the nutritive (plant) soul and the sensitive (animal) soul, or is the difference between them qualitative? Are there gray areas that are particularly difficult to categorize? Aristotle's position here is not entirely clear: most of his writings suggest that he saw the differences as qualitative.

From an evolutionary point of view, understanding the transitions that resulted in the three Aristotelian goal-directed systems is enormously challenging. The first problem, understanding *the transition to the first living system*, to the nutritive soul, is still not fully solved, although great strides have been made in this domain. Very little is known about the second, understanding *the transition to subjective experiencing*, the evolutionary origin of the sensitive soul. The third, understanding *the transition to rationalizing, symbolizing animals*, to the rational (human) soul, is one of the hottest topics in present-day evolutionary-cognitive biology, and progress is being made. All of these goal-directed systems are the products of chemical and biological evolution, and there is an evolutionary continuity between them. Studying

the transitions that led to their emergence may therefore provide valuable insights into the dynamic organization of the systems and also tell us something about the way in which they are related.

But can the evolutionary approach really tell us what life, consciousness, and rationality are, or will these aspects of being evade biological explanation? Since all are inherently dynamic systems driven by goals, and goals presuppose a criterion that enables evaluation, are they amenable to conventional scientific investigation? Can we find only the correlates of life, consciousness, and rationality but never attain a full explanation of such systems in biological terms? Is there an unbridgeable explanatory gap? Or more than one? The claim for such a gap was forcibly made by Yeshayahu Leibowitz, an Israeli philosopher, biologist, and theologian who, from a point of view diametrically opposed to our own, made us recognize the special difficulties and challenges inherent in understanding goal-directed systems.

The Leibowitz Challenge: The Kantian Epistemological Gap

> For it is quite certain that in terms of merely mechanical principles of nature we cannot even adequately become familiar with, much less explain, organized beings and how they are internally possible. So certain is this idea that we may boldly state that it is absurd for human beings even to attempt it, or to hope that perhaps some day another Newton might arise who would explain to us, in terms of natural laws unordered by any intention, how even a mere blade of grass is produced.
>
> —Kant 1790/1987, pp. 282–283

> Modern science is based on the concept of cause that Aristotle called the mechanical cause, and thus it [modern science] creates a rift between science and the theory of values. No deep philosophical reflection is needed to clearly recognize that the concept of goal is connected to that of value, which is the meaning that we attribute to things.
>
> —Leibowitz 1985, p. 27

Yeshayahu Leibowitz (figure 1.2) is little known outside the Israeli political and cultural scene, but he was probably the most outstanding and controversial Israeli intellectual during the second half of the twentieth century. His intellectual authority was based on his acerbic eloquence and vast knowledge of matters both secular and religious: he was a physician, had a doctorate in chemistry, was an eminent scholar of Judaism and an interpreter of

Figure 1.2
Yeshayahu Leibowitz (1903–1994). Portrait by Bracha L. Ettinger ©.

Maimonides, and was also an ordained rabbi belonging to the rationalistic Orthodox Jewish tradition. An outspoken critic of Israeli politics, a staunch opponent of retaining any of the territories seized during the Six-Day War, a supporter of conscientious objection to military service in the Occupied Territories and Lebanon, and an advocate of the separation of religion and state, Leibowitz was seen by many Israelis as the modern incarnation of a fierce biblical prophet. He also looked like one—tall, thin, and stooping, with a high forehead, bright eyes, and a slight Eastern European accent accentuating his formidable Hebrew, he was universally admired and feared. Professor Leibowitz, as he was referred to (he was a professor of biochemistry at the Hebrew University), was also the first philosopher of biology in Israel. The three issues on which he focused his interest in this field included the nature of life, the relationship between genetics and embryological development, and the mind-body (or as he called it, "the psycho-physical") problem. Goal-directedness was the common denominator of all three, and this was what passionately interested him. Leibowitz, a great admirer of Immanuel Kant, followed Kant's argument that there is an epistemological, explanatory gap between the mechanismic (mechanism-based) and teleological descriptions of natural and psychological processes.

He fully endorsed Kant's view that the teleological nature of the processes has to be *assumed* to permit analysis: it is not itself amenable to scientific (mechanism-based) study. The origin of life and the goal-directed processes of embryological development cannot be fully described in terms of molecular biology and genetics, he argued. He vehemently denied that the mind—willing, feeling, and thinking, which are all purposeful processes—can ever be understood in terms of neural mechanisms, however sophisticated.

Like all aspiring Israeli intellectuals, we had to face up to Leibowitz's Kantian challenges. Although we agreed with his political views and admired his civic courage, we strongly objected to his epistemological dualism. However, a person's gut feelings and vague scientific optimism will not do when confronting Leibowitz: a flood of angry and learned arguments will crush the poor offender. We had to confront head-on the challenge of explaining goal-oriented systems in some kind of mechanismic terms. We had to explain how a goal, a term that implies some kind of evaluation, can be accounted for within the framework of science, which has no room for values.

Leibowitz's challenge, like Kant's, was a general claim about the incommensurability between mechanismic and teleological explanations, and although Leibowitz is now dead, the challenge is not: we are still engaged in lively discussions with his argumentative ghost. His unifying focus on value has encouraged us to adopt a comparable unifying approach: since orientation toward a goal is a hallmark of a living system such as a food-seeking bacterium, or an embryological process that seems to "strive" toward a steady state, or a subjectively experiencing state of being such as that of a thirsty cat or a moralizing prophet, we believe that what scientists reveal by studying one goal-directed system may be worth exploring in others. The question is how best to study these types of system. What is the best approach for studying life, embryological development, or subjective experiencing?

We find an evolutionary approach focusing on the transition from a pre-teleological system to a teleological one particularly attractive and promising for four related reasons. First, if we can identify the evolutionary transitions to unicellular organisms, to multicellular organisms, to embryologically developing organisms, and to conscious, subjectively experiencing organisms, and describe them in terms of the changes in the systems' organization, it can help to characterize the mechanisms and dynamics that make them goal-directed. Second, the goal-directed system that appears immediately after a transition does not carry the baggage of later evolved structures and processes and will therefore enable us to recognize the most fundamental

an account of living beings from an evolutionary perspective, enumerated their properties:

> All these [living beings] possess individuality, either simple or compound; have a shape peculiar to their species; are born at the moment life begins to exist in them or when they are separated from the body whence they spring; are permanently or temporarily animated by a special force which stimulates their vital movements; are only preserved through nutrition which more or less restores their losses of substances; grow for a limited period by internal development; form for themselves the compound substances of which they are made; reproduce and multiply so as to carry on the species like themselves; lastly, all reach a period when the state of their organization no longer permits of the maintenance of life within them. (Lamarck 1809/1914, p. 195)

Individuality, metabolism, growth, reproduction, some form of heredity, and death were the characteristics of life, according to Lamarck. The special force of life that he listed as one of his characteristics was a *physical* force that resulted from the flux of subtle fluids in the self-organizing material body, the subtle fluids being electricity and heat fluxes. Lamarck was a committed and sophisticated materialist and, as we shall describe later, abhorred any kind of nonphysical incursion into the study of life (and also of mind).

There are other catalogs of life characteristics that are similar to Lamarck's list. They share many features, and this consensus has been important for investigations into the origin of life.[16] In table 1.1 we present several representative twentieth-century lists that emphasize the biochemical-metabolic, genetic-molecular, and evolutionary aspects of living systems. Although the motivations and emphases of these compilations are different, they are clearly related and describe the basic processes of living systems as biologists perceive them. Other compilations have a slightly different focus, many stressing the more abstract and holistic properties of living systems, such as emergence and self-organization. Margaret Boden, who analyzes the relations between life and mind from the perspectives of artificial intelligence (AI) and artificial life (AL), claims that in almost all lists of properties, self-organization, autonomy, emergence, development, adaptation, responsiveness, evolution, reproduction, growth, and metabolism loom large.[17]

Lists reflect history-laden scientific ideas, but they can be used as pointers and guides for research. Of course, they are only the first step in a very long journey. To convince the skeptic that a naturalistic explanation of life is feasible, additional questions must be answered. What are the organizational principles and the dynamics of a system that *generates* the above

Table 1.1
Representative twentieth-century lists of characteristics of minimal living systems.

Author	Characteristics	Emphasis
Gánti 1971/1987	(1) Inherent unity; (2) metabolism; (3) inherent stability; (4) information-carrying subsystem; (5) program control; (6) growth and multiplication; (7) hereditary system enabling open-ended evolution; (8) mortality	Self-organization, evolution
Orgel 1973	(1) Functionally complex organization; (2) subject to natural selection; (3) replication of genetic material; (4) information for specifying the living system stored in stable chemical molecules	Information, evolution
Maturana and Varela 1980	(1) Individuality (closure); (2) self-production; (3) responsiveness; (4) regulation and selectivity	Self-organization
Mayr 1982	(1) Complexity and organization; (2) chemical uniqueness (living organisms are composed of large polymers); (3) quality (some relations between aspects of the living world can only be described qualitatively); (4) uniqueness and variability; (5) possession of a genetic program; (6) historical nature; (7) subject to natural selection; (8) indeterminacy (biological systems have emergent properties)	Evolution
de Duve 1991	(1) Manufacturing its own constituents; (2) extracting energy and converting it to work for the system; (3) catalyzing the system's reactions; (4) having information systems, enabling reproduction; (5) closure (individuality); (6) regulation; (7) multiplication	Metabolism

characteristics? How did such a system emerge during evolutionary history? We look at possible answers to these questions in the next two sections.

Organizational Principles

> The soul is the first grade of actuality of a natural body having life potentially in it. The body so described is a body which is organized.
>
> —Aristotle 1984d, 412a, 27–28

In 1971, Humberto Maturana and Francisco Varela introduced a new concept into discussions about the nature of life.[18] This concept, "autopoiesis," was originally formulated to describe the dynamic organization of a

machine representing a minimal living system (the cell is the paradigmatic example):

> An autopoietic machine is a machine organized (defined as a unity) as a network of processes of production (transformation and destruction) of components which: (i) through their interactions and transformations continuously regenerate and realize the network of processes (relations) that produced them; and (ii) constitute it (the machine) as a concrete unity in space in which they (the components) exist by specifying the topological domain of its realization as such a network. (Maturana and Varela 1980, p. 78)

Autopoiesis has been a useful concept, aiding theorizing about both simple and extended manifestations of life. Maturana and Varela's focus was on the dynamic organization of an individual entity and its spatial and temporal persistence. One of the major motivations for their approach was to describe life in cognitive terms. A cognitive description of a very simple system requires that cognition is defined very broadly, as indeed it was: it was seen as the ability of the autopoietic entity to regulate its relations with the environment. The system achieves this through active sensor-effector relationships.[19] Figure 1.3 schematically describes a simple preautopoietic system (*A*), a basic autopoietic system (*B*), and an adaptive autopoietic system (*C*) that shows agency—the ability to act in the world in a goal-directed manner. The sensory components of the constitutive systems in figure 1.3*C* are coupled to effectors through feedback loops and enable the system to adaptively regulate its responses to a changing environment and to noise from within.[20]

In the same year (1971) that Maturana and Varela introduced the autopoiesis concept, the organic chemist Tibor Gánti published a book titled *The Principle of Life*, in which he developed a more concrete, chemical model of minimal life. Although less famous than Varela and Maturana's autopoietic model, which is deliberately very abstract and intended to capture the logical (dynamic, formal) structure of a living system, the chemical, cyclical-stoichiometric perspective of Gánti's minimal protocell model captures both the formal dynamics of living autopoietic processes and fleshes out their mechanical and chemical (material) facets. Although it is still idealized, it is extremely useful as a guide for theoretical and empirical approaches to the origin of life.[21]

Gánti started by enumerating the basic criteria for life (table 1.1, row 1). He regarded his first five criteria (individuality, metabolism, stability, a subsystem that carries information about the system as whole,[22] and regulation) as "absolute," by which he meant that these properties have to be

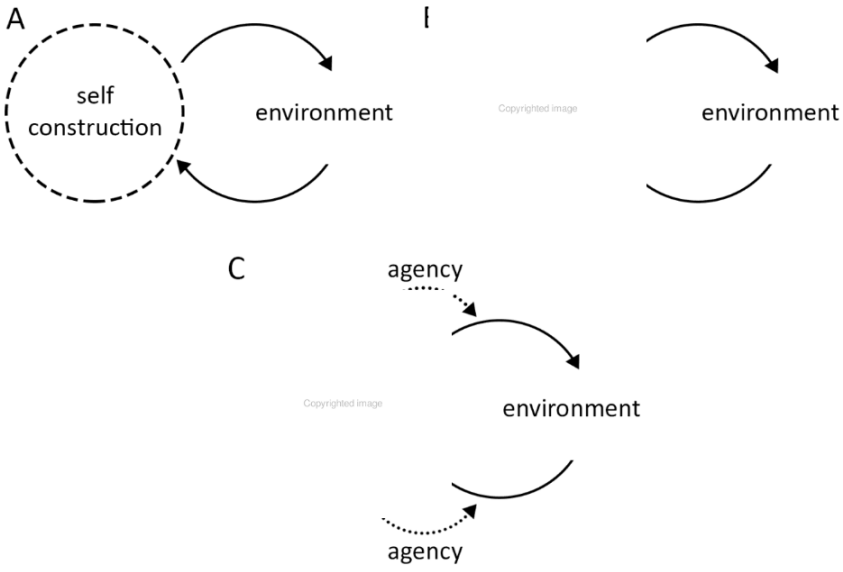


Figure 1.3
The autopoietic system: *A*, a preautopoietic system coupled to the environment, constituted by it and altering it; *B*, a self-sustaining and self-constructing system exhibiting elementary autopoiesis; *C*, a more advanced autopoietic system in which elements of the system adaptively control the way the system responds and constructs the environment. Modified by permission of Springer.

found in every living being and are jointly necessary and sufficient for life. His additional three criteria, which he called "potential" (the ability to grow and multiply, the capacity to exhibit hereditary variation and evolutionary change, and irreversible disintegration) are necessary for the ongoing, *long-term persistence* of the living state. He then constructed the simplest dynamic theoretical-chemical toy model that satisfied these criteria. He called his toy model the "chemoton" (figure 1.4). His chemoton is a chemical system made up of three indissolubly coupled autocatalytic subsystems that form a stable, functional entity. The links between the subsystems mean that they grow and reproduce in a regulated and coordinated manner. The "engine" of the chemoton is the autocatalytic metabolic cycle that transforms nutrients into the substances needed in the other two subsystems (the membrane and the information polymer) as well as for the cycle's own reproduction. Growth of the chemoton leads to the growth of the membrane, which, when it reaches a critical size, becomes

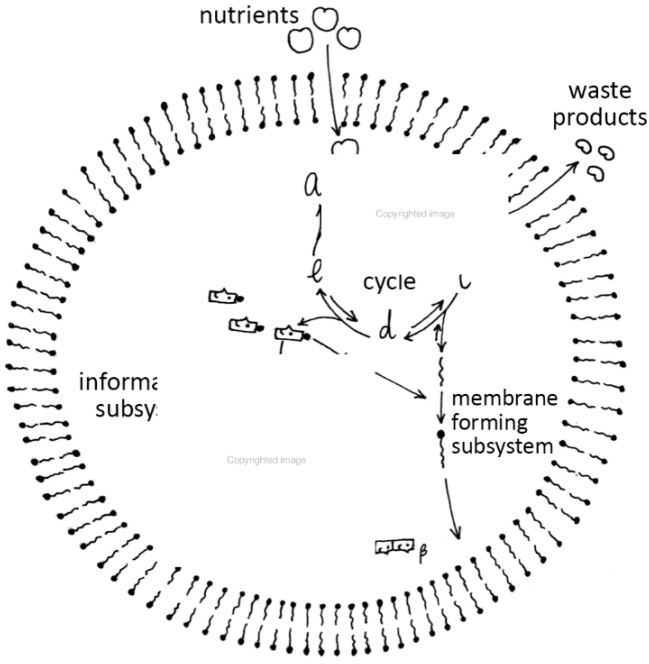


Figure 1.4
Gánti's chemoton, made up of three tightly coupled subsystems. At its core is the autocatalytic metabolic cycle, where molecule *a* combines with a substance formed spontaneously in the external environment and is transformed to *b*; *b* releases a waste product and forms *c*, which dissociates into *d* and a component of the membrane; *d* dissociates into *e* and a precursor of the polymer unit; and *e* dissociates to form two *a* molecules. Since the metabolic cycle forms more cycles like itself (because $a \rightarrow 2a$), it can be said to grow. One of the by-products of the metabolic cycle forms the hydrophobic part (*squiggle*) of the membrane unit, which combines with a hydrophilic component (*black dot*) to form the mature unit of the lipid membrane. The mature units spontaneously self-organize to form the membrane subsystem, which forms a boundary between the chemoton and its environment. In the third subsystem, the linear double-strand polymer reproduces by the template-based addition of units to the single "strands" of the initially double-strand structure, thus forming two double-strand polymers. Note that the membrane system and the linear polymer are functionally linked because the growth of the membrane depends on a by-product formed as the linear molecule polymerizes. This template-based polymerization occurs only when the units that are components of the polymer reach a critical concentration. When this happens, the original double-stranded polymer separates into two single structures (α and β), and the polymer components are consumed as they join to form new double structures (not shown). The length of the polymer therefore controls the number of turns of the metabolic cycle that are needed to produce the units necessary for its own reproduction and for the production of the by-product needed for forming membrane units. *Source*: Based on Gánti 2003, figure 1.1, p. 4, and the modification of this figure by Jablonka and Lamb 2006.

("reproduce"). Since different droplets differ in reproducible properties that affect their growth and fragmentation, evolution by chemical selection can occur. The droplets, claimed Oparin, are the forerunners of the first cellular organisms. Haldane's starting point was different. He began with self-replicating, virus-like elements, which formed spontaneously in the hot soup. These viral-like systems evolved through chemical-natural selection into more complex entities, leading eventually to a cell-like organism.

In the mid-twentieth century, the crystallographer John Desmond Bernal took up the challenge of understanding the transition from chemical to biological entities. He suggested that first we must have some idea of the geochemical conditions on the ancient Earth and then study—and try to simulate—three stages: (1) the emergence of complex monomers that can form the building blocks of biological beings (e.g., amino acids, pyrimidines); (2) the emergence from those monomers of more complex and more stable polymers and systems of interaction; and (3) the emergence from these latter systems of the first bona fide biological organisms. These three stages constitute what he described as the "first act," which led to the simplest, still-fragile life forms. The first act was followed by a "second act," in which the first living forms were stabilized, and by a "third act," which led to present-day complex organisms.[28]

Speculations like Bernal's led to great and heated debates among origin-of-life researchers. What kind of atmosphere did the ancient world have? Where did life emerge? At what temperature? What were the first monomers? What were the first polymers and self-sustaining systems? How did the jump to a protocell-like system occur? Soon, experiments began to provide answers. Here we can mention only a few of the landmarks in the history of this research, but they should be enough to give a sense of how the question of the origin of life moved into the realms of science.

The first demonstration that the basic monomers that characterize life can be formed in ancient Earth conditions was the famous Miller-Urey experiment. In 1952, Stanley Miller, a graduate student, and Harold Urey, his supervisor, took a mixture of water vapor and gases (methane, ammonia, and hydrogen), which they believed simulated the reducing conditions of the primeval atmosphere, and passed electric sparks ("lightning") through it. Organic monomers such as amino acids, which according to the warm pond scenario are some of the types of precursor molecules required for life, formed readily in these conditions. A few years later, in 1961, the Catalonian biochemist Joan Oró showed that the nucleic acid base adenine, as well as many amino acids, could be formed by heating an aqueous solution of ammonia and hydrogen cyanide. Although there has been no consensus

about the conditions that prevailed on ancient Earth, these experiments established the feasibility of the first of Bernal's stages, and they were followed by others showing that monomers can be formed under many different conditions. The second stage, the generation of self-sustaining complexes and cycles of reaction, is much more tricky, but there has been progress here too. For example, in line with some of Oparin's ideas, Sydney Fox showed in the late 1950s that amino acids spontaneously link to generate small peptides that form tiny closed spherical blobs that have many of the basic characteristics of life, demonstrating something like growth and a very crude form of reproduction through fragmentation.[29]

A more recent and very ingenious hypothesis, which combined Bernal's first two stages (the production of monomers and then polymers and self-sustaining systems of interactions), has been worked out in great chemical detail by the German chemist and patent lawyer Günter Wächterhäuser. It is rather different from the warm pond scenarios. Wächterhäuser suggested that life began in volcanic vents in the deep sea, in conditions of high pressure and high temperature, where pressurized hot water with dissolved volcanic gases (like carbon monoxide, ammonia, and hydrogen sulfide) flowed over catalytic solid surfaces (e.g., iron and nickel sulfides). As a result, organic compounds containing carbon were formed and bound to the catalytic surface, and the process became autocatalytic. Once such a primitive autocatalytic metabolism was established, it began to produce ever more complex organic compounds, ever more complex pathways, and ever more complex catalytic centers. The ancient system can be thought of as a scaffold for the formation of more complex systems that eventually came to have nucleic acids and the other familiar constituents of present-day biological systems. The scaffold was then discarded, although a few traces of its prior existence can still be discerned.

Many other scenarios stretch and challenge the imagination, and it is quite possible that several of them describe processes and products that occurred on ancient Earth. We still do not know which system or systems (they may have become combined) led to protocellular, chemoton-like life, and solving this problem is one of the biggest hurdles in origin-of-life research. Some hypotheses postulate that nucleic acids (RNA or something similar) came first ("genes first"); others suggest that biochemical reactions and pathways involving proteins came first ("metabolism first"); yet others stress the priority of a containing, enclosing (membrane) system.[30] Today, hybrid models that combine these three aspects seem most promising. Whatever their priorities and models, chemists and physicists today appreciate the complexity of a reaction system that can implement living

organization. The first living systems, as well as their precursors, were all chemically complex.[31]

The Transition Marker

One of the questions about the origin of life that has remained unanswered is: At what point in the evolutionary process do we decide that a system is living? Is a self-replicating, 10-base-pair-long RNA molecule living? Is the simple chemoton shown in figure 1.7 or even its precursors (maybe something like the autopoietic system shown in figure 1.5) really living, or do we need a more complex system? Intuitively, it seems that a small RNA molecule, a very simple autopoietic system, or even a protochemoton are not really alive—they are somehow too simple. But this may just be a prejudice, an unjustified intuition. Even if most people have the feeling that life has to be more complex than a self-replicating, 10-base-pair-long RNA molecule or an elementary autopoietic system, these gut feelings require a scientific articulation. Is there a marker, a capacity or part that will allow us to reconstruct the whole system from it, a threshold beyond which we can agree that a system is alive?

Maturana and Varela presented powerful arguments in favor of an autopoietic organization as the manifestation of life. However, they did not suggest a marker and did not consider the conditions that would enable an autopoietic system to exhibit long-term persistence. A criterion that can be used to mark forms of life that can persist over time (we call such a criterion a "transition marker") was suggested by Gánti, and Maynard Smith and Szathmáry developed the idea.[32] They highlighted one characteristic of the system, heredity, and distinguished between limited and unlimited heredity. Systems that can have only very few hereditary variants are *limited heredity systems*, and they reside in the gray area between the nonliving and living stages. They are on the evolutionary route to fully fledged life if they evolve further and the number of their hereditary functional variations becomes great enough to be practically unlimited. Without open-ended heredity and the open-ended evolvability that comes with it (that also generates new niches of which the new variants are part), the lineage would soon go extinct. For Maynard Smith and Szathmáry, it is *the transition to unlimited heredity* that identifies sustainable living entities. According to this view, rather than a single and inevitably highly contentious line between life and nonlife, there is a gradual transition. A gray zone, rather than a transition point, marks the road to sustainable life.

A New (Living) Way of Being: Goals, Functions, and Functional Information

The transition from a chemical to a living system involved more than new and more sophisticated chemical structures, mechanisms, and dynamics. Life is not just wonderfully complicated chemistry but a drastically new way of being. With life, mere chemical processes and mechanisms became organized into systems to which a goal (self-maintenance) can be ascribed, and the parts and processes of such systems can be said to have functions.

Function is something that only parts or processes in goal-directed systems can have. Living beings, which reconstruct themselves and their parts, are paradigmatic goal-directed systems; so are systems designed by living creatures, such as human artifacts or termites' nests, and so are chemical systems on the verge of living, like the system described in figure 1.5. Biological function (also known as teleofunction) is defined as the role that a part, a process, or a mechanism plays within an encompassing system—a role that contributes to the goal-directed behavior of that system.[33] As we have already noted, the most basic goal-directed behavior of living organisms is self-maintenance (survival) and, in the long-term, reproduction.[34] *Functional information* is any difference that makes a systematic, causal difference to the goal-directed behavior of an encompassing system and in the case of simple living forms, to the system's self-sustaining dynamics.[35] Chemical processes that do not organize into self-maintaining entities do not have functional information since they are not parts of a goal-directed system. Function is not a new high-level chemical process or trait. In Aristotelian terms, it is a facet of the teleological cause "that for the sake of which" things exist. Functions and functional information are the very essence of living organisms and are irreducible to descriptions in terms of chemistry.[36]

Before the realization that matter is inherently active, before the recognition that life evolved, and before the early twentieth-century advances in the understanding of biochemical cycles, the dynamic goal-directed organization that is the hallmark of living organisms was seen as a deep mystery, even by biologically well-informed philosophers and naturalists. Kant could not envisage how a self-organizing living being (what he called "a product of nature") could be constructed in such a way that everything in it has a goal and yet is also, reciprocally, a (mechanism-based) means.[37] Now, however, as a result of the experimenting, theorizing, and philosophizing about the origin of life, the nature of living entities and their origins have lost their aura of unreachable mystery. The Kantian gap has been bridged by our better understanding of the dynamic nature of matter and our ideas

about how certain types of autopoietic dynamics can instantiate life (chemoton dynamics is one example); we can theoretically simulate autopoietic systems and figure out how new functions can arise through a process of natural selection. We have moved beyond the rather narrow notions of matter and mechanism that Kant had assumed. In fact, as Wittgenstein pointed out, the metaphysical problem of life has vanished: "The solution of the problem of life is seen in the vanishing of the problem."[38] The question we now ask is whether the approach to the evolutionary transition to life can serve as a model for understanding the transition to consciousness.

Back to Consciousness: The Qualia Gap

> How it is that any thing so remarkable as a state of consciousness comes about as the result of irritating nervous tissue, is just as unaccountable as the appearance of the Djin when Aladdin rubbed his lamp.
>
> —Huxley and Youmans 1868, p. 178

We believe that if scientists are able to understand what the transition to a conscious system entails, characterize this new way of being, provide a model describing the kind of biological dynamics that instantiates it, and define a transition marker, then subjective experiencing will become as well explained as the state of being alive and, as was the case with the latter, the mystery will slowly vanish. The case of the transition to life shows that the Kantian explanatory gap between mechanism-based and teleological descriptions of a living system can in principle be bridged, and there is almost universal agreement among scientists and philosophers that the problem is accessible, and its solution does not require Higher Intervention. It is not a coincidence that creationists are extremely worried about this kind of research.

Such a universally accepted dissolution of mystery has not yet happened with the problem of consciousness. Although the approaches of most philosophers and neuroscientists are firmly grounded in biology, some eminent philosophers still doubt the possibility of explaining consciousness using the traditional tools of biological investigation. The problem was well captured by the words of Huxley and Youmans quoted above, as well as by Thomas Nagel, who more than a century later fleshed it out in his article "What Is It Like to Be a Bat?" In this famous article, Nagel considers the subjective experiencing of species that sense the world in ways different from our own, such as that of a bat navigating using sonar; he points

levels can therefore be said to present Kantian explanatory gaps because in all three cases something completely new, a new way of being, emerged. In the case of the transition to life, it was the emergence of function and functional information; in the case of the transition to subjective experiencing, it was the emergence of first-person experiences and subjective needs; and in the case of the transition to rationality, it was the emergence of symbolic concepts and symbolic values like truth, justice, beauty, and freedom. In all three cases, a new, open-ended realm of possibilities opened up: with life it was open-ended evolution; with subjective experiencing, as we argue in later chapters, it was open-ended associative learning; and with rationality it was open-ended imagination and reasoning.

Yes, we concede that a biological account of subjective experiencing is a special and challenging problem, but it is not unique: there are three such problems. We agree with Nagel that the three teleological transitions—to life, to subjective experiencing, to human values—do pose special philosophical and evolutionary challenges. But we take a diametrically opposite position to his regarding what these teleological transitions mean. As we see it, all are explicable within a sophisticated evolutionary framework, which once understood seeps down (or up?) and reformulates the philosophical problems. So our suggestion is exactly the reverse of that of Nagel, who argues that philosophical obstacles imply the fundamental explanatory insufficiency of evolutionary theory. We argue that once we can account for the teleological evolutionary transitions, many of the problems that were deemed insoluble dissolve. We think that Nagel's problem is the well-recognized one of the failure of evolutionary imagination. As Darwin confessed, the evolution of the eye made him shudder, but as he explained, this was the problem of a failure of his imagination rather than a failure of his evolutionary theory.[45] In fact, we believe that having three problems rather than one is very helpful because two of the "hard" problems—life and rationality—are actually beginning to yield to evolutionary investigations. Recognizing that there is continuity between living, subjective experiencing, and rationalizing and that each one is the product of a transition to a new teleological system can be informative. We can look for analogies between the first, second, and third transitions and see, in very general terms, what we can learn from them and whether they throw light on the nature of subjective experiencing.

The first and most obvious thing to recognize is that the only system in which consciousness has ever been found is a living system, so a good starting point is to investigate the kind of living organization that instantiates the essential properties of subjective experiencing. (At this point we

are not interested in robots, although we shall discuss them later.) This is the approach taken by Evans Thompson, who combined an expanded autopoietic view of living with a phenomenological approach and emphasized the cognitive embodiment of biological systems, a view that he called "embodied dynamicism." This view redefines the mind-body problem as the *body-body* problem.[46] Feelings and thoughts cannot be attributed to a brain, however evolved: they can be attributed to an enbrained body, a living, active animal, which is a very different matter. A second facet of the continuity between living, subjectively experiencing, and rationalizing forms of life is the intriguing parallels between them (table 1.2). For example, just as life entails functions, so subjective experiencing entails qualia, and rationality entails symbolic concepts. Similarly, the teloi of minimal life are phylogenetic (survival and reproduction), those of consciousness are ontogenetic values (values that can be ascribed to newly learned complex stimuli and actions guiding open-ended learning), and those of rationality are symbolic values (symbolic categories ascribed to states and actions that guide human cultural behavior). Just as there is no life without the processes instantiating it having functions, so there is no consciousness without the instantiating processes having qualia, and no rationality without symbolic concepts. We come back to the table in the last chapter. Here we focus on the middle column, looking at subjective experiencing.

It is a remarkable fact that in spite of the self-evident and generally acknowledged usefulness of an evolutionary approach to all biological questions, including the origins of life and the origins of human reflective consciousness, attempts to investigate subjective experiencing using the traditional methods of evolutionary biology have been, until quite recently, surprisingly uncommon in both the biological and philosophical literature. Scientists studying subjective experiencing are committed to the theory of evolution, and some philosophers and several neurobiologists take it very seriously, but it was only during the first decade of the twenty-first century that detailed evolutionary scenarios began to be suggested.[47]

One of the reasons for the paucity of concrete, evolution-focused approaches seems to be the lack of agreement about how minimal subjective experiencing or minimal consciousness should be characterized. This makes it very difficult to decide which organisms have it (a problem known as "the distribution problem") and where and when in evolutionary history it first emerged. The problem is exacerbated by the nature of evolutionary changes, which usually are not sharp and clear. As we have already noted, when a new level of organization emerges, there is always a gray and fuzzy

Table 1.2
Suggested parallels between concepts used to describe living systems, experiencing animals, and rationalizing humans.

Living	Subjective experiencing	Symbolizing/rationalizing
Phylogenetic teloi: self-maintenance—survival and reproduction	Ontogenetic teloi: ascription of values to newly learned complex stimuli and actions	Symbolic teloi: symbolic values like freedom and justice
Function	Qualia	Symbolic concepts
Heredity (unlimited)	Memory/learning (unlimited)[a]	Transmission of adaptations involving symbolic representations (unlimited)[b]
Development	Recall	Social reconstruction[b]
Evolution (open-ended)[c]	Learning (unlimited); behavioral adaptation, open-ended[c]	History (open-ended)[c]

Notes:
[a] Our suggestion that unlimited memory and learning parallel unlimited heredity is a central theme in this book. We contend that only conscious living beings can learn in an unrestricted manner but *not* that all sentient beings (e.g., babies) have such learning ability.
[b] These notions can be seen as different facets of historical cultural change.
[c] The relationships between unlimited heredity; learning; symbolizing; and open-ended genetic, neural, and symbolic evolution are far from simple. We assume that at each level hereditary transmissible variations map onto functionally diverse, potentially novel phenotypes and lead to the construction of new selective environments. For a discussion of the relation between unlimited transmissible variations and open-ended evolution, see de Vladar, Santos, and Szathmáry 2017.

area where the classification of the system is uncertain; for most philosophers this presents a major problem, although biologists are much more tolerant of classificatory ambiguity—gray areas are inevitable, given evolutionary history. Nevertheless, once the transition has been identified, it is possible to recognize processes and properties in the pretransition stages, which, though not sufficient for the transition, are necessary for it to occur. Over evolutionary time, the necessary factors and processes accumulate, combine, and become sufficient. A new teleological system emerges.

Our strategy in this book is to employ the evolutionary, transition-oriented methodology that has proved so fruitful for the study of life to the study of the teleological system we call "consciousness." We therefore

present and discuss the "lists" of characteristics that neuroscientists and philosophers of mind have associated with consciousness and the very preliminary dynamic models that neurobiologists have suggested. On the basis of theoretical and empirical considerations, we suggest a transition marker for consciousness: unlimited associative learning (UAL). UAL refers to an animal's ability to ascribe motivational value to a compound stimulus or action pattern and to use it as the basis for future learning. We show that the features that enable UAL are based on computational mechanisms and neural structures generally believed to underlie the ability to form mental representations and presuppose the list of criteria and the dynamic organization that scholars of consciousness suggest. We provide evidence that the groups that exhibit UAL are the same as those having the capacities in our list, even when learning-independent criteria are used. Following the evolutionary origins of UAL enables us to identify its building blocks and attempt to reconstruct the system of which it is part—a system that, we argue, instantiates minimal consciousness. Finally, we consider how understanding the evolution of UAL enables us to work out how consciousness has changed during evolutionary history.

Dennett's Hierarchy and Phylogenetic Distributions: Locating the Experiencing (EX) Factor

> I want to propose a framework in which we can place the various design options for brains, to see where their powers come from. It is an outrageously oversimplified structure, but idealization is the price we should often pay for synoptic insight. I call it the Tower of Generate-and-Test; as each new floor of the Tower gets constructed it empowers the organisms at that level to find better and better moves, and find them more efficiently.
>
> —Dennett 1995, p. 373

Dennett provided a general, evolution-inspired framework for describing different levels of goal-directedness that we find useful for investigating the evolution of subjective experiencing. He used the term "intentionality"—the ability to represent or to stand for things, properties, and states of affairs—and called his approach the "intentional stance." His evolutionary-selectionist framework can be seen as an extension of the Aristotelian teleological approach, and we make use of it throughout the book. However, our standpoint differs from Dennett's because whereas Dennett believes it is convenient to talk about living and sensitive creatures as if they had a telos, we think that these selection-based systems are intrinsically teleological.[48]

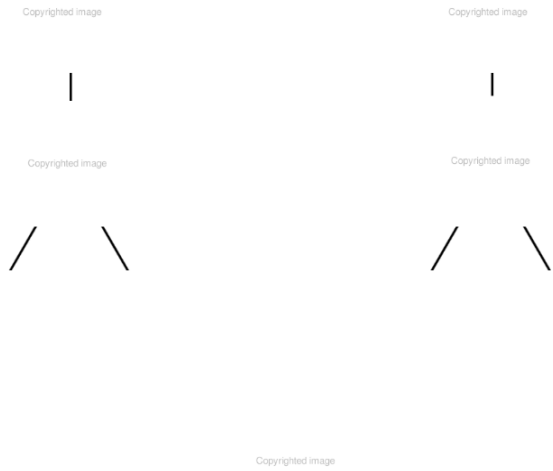


Figure 1.5
Darwinian organisms. These organisms, which multiply and transfer their characteristics to offspring, can "learn" through the mutation-selection mechanism. The lineage on the right is more adapted and has greater reproductive success than the lineage on the left.

Dennett calls the first type of goal-directed organisms, those that generate random variations and are tested by the external environment, "Darwinians organisms" (see figure 1.5). They are, of course, the products of the first great teleological transition—to life. They include organisms such as single-celled microorganisms, plants, fungi, and very simple animals like sponges, which Dennett thinks have limited flexibility (i.e., plasticity) and do not learn during their own lifetime. Although we think Dennett greatly underestimates their plasticity and we believe that all (or most) organisms, including the single-celled, exhibit some form of learning,[49] we agree with Dennett that Darwinian organisms do not experience (i.e., do not have phenomenal consciousness) because, as we claimed earlier, a central nervous system with a particular type of dynamics is necessary for consciousness.

Dennett's second type of organisms are the "Skinnerian organisms" (figure 1.6), which learn during their lifetimes (ontogenetically) through

Figure 1.8
Gregorian organisms. The women in this orchestra create, by using symbols (musical notation) and artifacts, a new world of experience.

significance. Humans are not intellectual geniuses with the emotions of a chimpanzee. In profound ways we experience the world differently. The transition to having abstract values (justice, freedom, beauty, and others) that regulate social and individual behavior—the transition to rationality—occurred, as far as we know, only in the hominin lineage.

Dennett's generate-and-test hierarchy can be seen as an evolutionary (selectionist) interpretation of Aristotle's teleological hierarchy. Using Aristotelian terms, Darwinian organisms can be said to have a vegetative soul, while Skinnerian and Popperian organisms have a sensitive soul (although the Popperian ones are endowed with quite a bit of imagination), and the Gregorian organisms have a rational soul.

Although Dennett's hierarchy is useful as a general framework, evolutionary reasoning requires grounding in terms of phylogenies and scenarios. For example, if one claims that vertebrates are subjectively experiencing animals, the questions an evolutionist will ask are: All vertebrates? Only vertebrates? (Maybe some chordate ancestor was also endowed with a little

bit of the EX factor? Maybe nonvertebrates, such as some arthropods or mollusks, are minimally conscious too?) Was it lost in some lineages? Did it evolve in parallel in some vertebrate and invertebrate lineages, and when did this happen? Can we work out the selection conditions (the scenarios) for each occurrence? What were the necessary (and jointly sufficient) molecular-neurological-cognitive-behavioral preconditions in the ancestral lineages, and how are these phylogenetically distributed? Every theory about consciousness suggests answers to these questions, and biologists have to produce the information needed. Even if one assumes that subjective experiencing is primitive, as Chalmers and Nagel think, or believes, like Lynn Margulis, that it is characteristic of all living organisms, the difference between a dog's and a tomato's levels of subjective experiencing is striking and requires a detailed and concrete biological explanation. Natural philosophers became fascinated by the evolutionary framing of this question as soon as theories of evolution emerged in the early nineteenth century. Their ideas paved the way, scaffolding and influencing subsequent research on the subject, and in spite of the many inevitable blind alleys they stumbled into, their work is full of intriguing and sometimes surprisingly relevant insights.

2 The Organization and Evolution of the Mind: From Lamarck to the Neuroscience of Consciousness

Beginning with the eighteenth-century associationists, we trace the origins and development of modern physiological and evolutionary ideas about mental life. The first evolutionary psychologist we discuss is Jean-Baptiste Lamarck, who at the beginning of the nineteenth century suggested that through adaptive self-organization mediated by use and disuse, apes were transformed into humans, and the nervous system changed from a system of communication in a humble polyp into the bicortical, sophisticated human brain. The second is Herbert Spencer, who argued that mental evolution proceeded from reflexes to habits and from habits to complex instincts, culminating in consciousness, whose different facets are feelings, reasoning, and willing. The third is Charles Darwin, who established the validity of the theory of descent with modifications and provided a powerful motor for adaptive evolution—natural selection. Although he carefully avoided the question of the origins of mentality, he used sexual selection to show that animals have mental traits, used the descent of man from the apes to show how mental faculties evolved in our lineage, and used the expression of emotion in man and the animals to show how emotions evolved through the processes of natural selection and use and disuse. The ideas of these thinkers deeply influenced Williams James, the father of modern consciousness studies, who applied the selection principle to the operation of the mind-brain, arguing that consciousness is adaptive scale tipping, which "loads the dice" of neural activities. For him, consciousness was a goal-directed, selection-based process that resolves perceptual ambiguities, guides actions, and enables inferences to be made. Although hugely influential, his view of psychology as the science of mind was soon supplanted by behaviorism, which dominated psychology for a large part of the twentieth century. Behaviorism redefined psychology as the study of controlling and manipulating observable behavior and, in its American version, banished mind and consciousness. The subsequent rise of the cognitive sciences paved the way for the return of James-inspired consciousness studies during the late twentieth century.

A full account of the development of evolutionary psychology in the nineteenth and twentieth centuries would require several book-length treatises. This is not just because the topic was of interest to so many scholars—it is

also because those who had the most to say about it—Jean-Baptiste Lamarck, Herbert Spencer, Charles Darwin, and William James—were some of the most impressive and prolific natural philosophers of their time, and interesting lines of descent and dissent exist among them.[1] Lamarck laid down some of the theoretical evolutionary foundations on which Spencer developed his ideas, which then, together with Darwin's ideas on the evolution of human mentality and the expression of emotions, provided the scaffold for later ideas, such as those of William James. Although it is in the nature of scaffolds to be discarded after use, the intellectual structures erected by James and by his late twentieth-century followers form the basis of current theories of consciousness, including our own.

The common assumptions of all four thinkers sprang from physiological versions of the "associationist" theories developed in Britain and France during the late eighteenth century. These theories connected ideas to sensations and sensations to the physiology of the nervous system.[2] The notion that mental processes are made of (or as we will usually say, are constituted by) physiological processes in the nervous system facilitated the development of evolutionary ideas, and in the nineteenth century, it led to the inevitable conclusion that mental-physiological processes have evolved.

The Associationists

The view that mental activity is the result of the association of ideas is often traced back to Aristotle's reflections on recollection.[3] Like almost all of Aristotle's ideas, this proposal underwent many reformulations by theologians and philosophers, who broadened and qualified the basic claim that recollection occurs when one encounters features associated with the original experience.

In its eighteenth-century reincarnation, the explanation of recollection and other mental processes in terms of associations was based on several assumptions. The first, that complex thoughts and emotions are ultimately derived from sensations, was developed by John Locke. He believed that there are no innate ideas and that the mind is a blank slate, or, in his own words, "white paper," on which sensations are inscribed through the operation of innate, internal psychic principles.[4] The second assumption was that there are principles, like those enumerated by David Hume, that link sensation-derived ideas together: resemblance, spatial contiguity, cause and effect (which is for Hume contiguity in time), and contrast.[5] The third assumption was that this linkage is the result of the operation of some kind of force analogous to Newton's gravitational attraction. These associationist

assumptions were not, at first, linked to materialistic or physiological notions; the elementary sensations that were the building blocks of complex ideas were seen as belonging to the realm of the mental, which was assumed to be different from that of the physical and the physiological. It was only with David Hartley (1705–1757) in England, and more explicitly and boldly with Pierre Jean George Cabanis (1757–1808) in France, that mental processes began to be understood in terms of the physiological activities of nerves.

David Hartley, an English physician and philosopher, was the first to develop a physiological account of the association of ideas.[6] He started with the assumption that the body's "component particles" are subject to the same "subtle laws" that govern all other material entities. Like all eighteenth-century scholars, Hartley was deeply influenced by Newton's theory of gravitation and sought general, Newtonian laws of human nature. But although Hartley's (1749) *Observations on Man, His Frame, His Duty, and His Expectations* was highly regarded by late eighteenth-century and early nineteenth-century English philosophers, it was in France that associationism became the basis of a secular science of man and mentality. This science was promoted in revolutionary and postrevolutionary France by the "ideologues"—Parisian intellectuals who were so called because they declared an ambition to construct a *science of ideas* that would establish a physiological basis for social and cultural reforms. As Destutt de Tracy, who used the term "ideology," boldly stated, ideology is a part of zoology.[7] The evolutionary theories of Lamarck, especially his theory of mental evolution, shared and extended the aspirations and assumptions of the ideologues.

Pierre Jean Georges Cabanis, a physician, philosopher, and social reformer, was one of the ideologue leaders. His principal work *On the Relations between the Physical and Moral Aspects of Man* (*Rapports du physique et du moral de l'homme*) was published in 1802. As the title indicates, he wanted to establish a bridge between the physical features of humankind and its highest mental-social aspects, and he did so through the development of what we may call "physiological psychology."[8] As a doctor and a social reformer, Cabanis was well aware of and fully appreciated research showing that faculties such as memory and reflection are associated with states of the brain, on the one hand, and with the physical and social environment, on the other. His conclusion was that "the moral is only the physique from a certain point of view."[9] He suggested that brain disturbances can bring about madness and frenzy, while hysteria and languor can result from disorders in the genital center of sensibility. Consciously following in the steps of the ancient Hippocratic doctors, he suggested that age, sex,

Figure 2.1
Jean-Baptiste Pierre Antoine de Monet, Chevalier de Lamarck (1744–1828).
Source: Wellcome Library, London.

assertions, which he consistently repeated, match the penetrating characterization of Lamarck by the famous, temperamental nineteenth-century French literary critic Charles Augustin Sainte-Beuve (1804–1869), who as a very young man had attended elderly Lamarck's lectures in the Jardin des Plantes (box 2.1).

According to Lamarck, two basic principles accounted for the evolutionary changes that started with the tiny, fragile creatures produced by spontaneous generation and culminated (but did not end, since evolution is, at least in theory, open-ended) in humans. The first was the effect of the changing environmental conditions on physiological and hereditary processes. Lamarck assumed that living organization is very plastic. Indeed, to him, adaptive plasticity is inherent in the special type of self-organization that characterizes the living state. In young animals with malleable and soft tissues, adaptive plasticity is mediated through behavioral changes: changed environments lead to a need to cope with the change, and this leads to activities that alter the use of some organs, leading to new habits; when the conditions and habits persist for many generations, the new behavioral tendencies and their supporting organs become hereditarily stable. As a result, offspring develop more readily the characteristics that their ancestors had laboriously and imperfectly acquired, and the cumulative result of this process is the observed diversification and functional specialization of animals. The long neck of the giraffe is one of the many examples that Lamarck put forward to exemplify this process.[14] To this major principle, Lamarck added a supplementary process, hybridization. Once new races or species were formed, they sometimes hybridized, a process that further increased the diversity of living organisms and that can be seen in domestication.

The second basic principle behind evolutionary change explained the nature of adaptive plasticity. It was, Lamarck suggested, a result of the dynamics of the subtle fluids (heat, electricity) in the living entity that tended in time to increase complexity through activities involved in self-maintenance and growth. Self-maintenance was the result of a special kind of self-organization that occurred as the activities of the subtle fluids organized matter. This self-organization was the hallmark of life, and Lamarck elaborated on the process in the second part of *Philosophie zoologique*, which dealt with the origin of life. Becoming more complex was a physical inevitability, according to Lamarck, because the fluid movements were accelerated as the effects of their activities built up during the organism's growth and development. As we see it, Lamarck was expressing a very modern idea, which would today be described as a positive feedback reaction. He thought that the movements of fluids lead to the formation of better and deeper old and

Box 2.1
Charles Augustin Sainte-Beuve on Lamarck.

In his autobiographic novel, *Volupté*, Sainte-Beuve conveys his strong impressions of the lectures delivered by Lamarck, and of the man.

> I attended frequently, several times per decade, the lectures in natural history of Mr de Lamarck, in the Jardin des Plantes; his teaching, of whose hypothetical paradoxes, and conflicts with other more positive and more advanced systems I was unaware, had a powerful attraction for me because of the serious and fundamental questions which it always raised and the passionate and almost sorrowful tone that was mingled with it. Mr de Lamarck was by then the last representative of this great school of physicists and observers who had reigned since the times of Thalès and Democritus until Buffon: he was profoundly opposed to the tiny chemists, experimentalists and analysts, as he called them. His hate, his philosophical hostility to the Flood, to Biblical Creation and anything reminiscent of Christian theology, was profound. His conceptions had much simplicity, bareness, and much sadness. He constructed the world with the least number of elements, the least number of upheavals and with as great a duration as possible. According to him, things arose of themselves, one by one, continuously, enduring for sufficiently long time, without instantaneous transformation for overcoming catastrophes, without disasters or commotions, without centers, growth-nodes or organs deliberately arranged to help them reduplicate. A long blind patience, this was his kind of world. The actual form of the earth, according to him, depended only on the slow deterioration of pluvial waters, on the daily rotations and the successive displacement of the seas; he admitted no great bowel movements in this Cymbeline [goddess of nature], nor the renewal of her earthly face by some temporary heavenly body. So too, within the organic order, the mysterious power of life was rendered by him as small and as basic as possible, assumed to develop by itself, order itself, construct itself little by little through time; the obtuse need, the only [source of] habit in diverse circumstances, eventually gave rise to organs, opposing the constant, destructive power of nature; For Mr de Lamarck separated life from nature. Nature, in his opinion, was stone and cinder, granite of the tomb, death! Life intervened there as an artful and strange accident, an extended battle, with more or less success or equilibrium here and there, but always finally conquered; cold immobility reigned before as well as after its occurrence. I loved these questions of origin and finality, this view of a dismal nature, these sketches of obscure energy. My reason was suspended, pushed to its limits, enjoying its own confusion. I was, of course, far from accepting these much too simple hypotheses, this uniform series of continuity that went against my exuberant feeling of creation and vigorous youth, but the boldness of a man of genius made me think. (Sainte-Beuve 1834/1986; translation by Eva Jablonka)

Lamarck's world was indeed devoid of grand cataclysms and mysterious forces: the origin of life was a recurring, chemically predictable accident; living beings evolved very slowly; individuals persisted for a very short time; and there were no miracles, no afterlives, no external telos. But it was, for Lamarck, a beautiful and rich world, a feeling that, it seems, was not apparent to the young and impatient Sainte-Beuve.

new paths that facilitate further movements, that allow growth, that promote further activities, and so on. The challenges of new conditions usually (not always!) add to rather than diminish the existing order. Together, the cumulative effects of these two related factors—complexification and adaptation to contingent conditions—generated over very long periods of time the diversity and increase in organizational complexity observed in the living world. The result is a branching tree, which started from several events of spontaneous generation (probably three—one leading to plants and two leading to animals) that became firmly established and gave rise to all plants and all animals. This progressive and branching process, presented in the first part of *Philosophie zoologique*, accounts both for the diversity and for the increase in complexity in the living world and leads to a natural system of classification.

The Physiological Evolution of the Mind

> In this fictitious entity [mind], which is not like anything else in nature, I see a mere invention for the purpose of resolving the difficulties that follow from inadequate knowledge of the laws of nature.
>
> —Lamarck 1809/1914, p. 286

Philosophie zoologique was part of a grand project: that of *Biologie*, the study of *all* living bodies—plants and animals—as well as their products—minerals. In the more modest project of *Philosophie zoologique*, the subject matter was animal diversity and functional biology explained through their physiology. The evolutionary origin and sophistication of animal life and mind were part of this scheme, and Lamarck devoted two-thirds of the book to the problems of the origin and evolution of life (part II) and mind (part III), dwelling with special pleasure, as he confessed, on the question of the origins and the evolution of mentality. His starting point, like that of Cabanis, was the assumption that the moral is an aspect of the physical.

Lamarck went further, however, than other ideologues, suggesting that only through an evolutionary analysis, starting from the simplest living beings, can one discover the basic principles underlying mental phenomena. Beginning with complex animals like man, on which Cabanis and many other philosophers focused, cannot lead very far, Lamarck argued, because the complexity of the organization of these already highly evolved animals makes their organization "the most difficult from which to infer the origin of so many phenomena."[15] It is only when we start with the simplest organisms that manifest mental life and follow the gradual evolution of mentality in animal lineages that we can discover what feelings and ideas actually

are and how human mentality came about. To do this we must study the organs and processes that constitute and generate feelings: we must study the anatomy and physiology of the nervous system.

Anatomically, the nervous system envisaged by Lamarck consists of three major parts: (1) a pulpy medullary mass, which is a kind of center of communication, with many extensions and threads; (2) protective sheaths enclosing the central mass and the threads (nerves) emanating from it; and (3) a subtle fluid that he said is fundamentally electric current, which is the factor active in transmitting sensory inputs, eliciting motor outputs, and tracing activity marks on the soft medullary mass. Although not all animals have a nervous system, claimed Lamarck, these three major elements are present in all of those that do. However, nervous system organization differs in different taxa, and there are many variations and gradations between the different systems. That is why the nervous system provides the best suite of characters for the classification of neural animals, just as in plants the best organs for classification are the reproductive parts, the flowers.

Lamarck recognized three major stages in the evolution of the nervous system, with endless gradations in-between. The simplest system was a nervous system with a medullary mass or masses (ganglia) with nerves, which could control and coordinate motor actions but not feelings; this system, he suggested, seems to characterize creatures such as sea urchins and possibly sea anemones. In the next major stage, seen in most invertebrates and in simple vertebrates, a more centralized sensory system evolved. In these animals, one ganglion (clusters of nervous tissue), the head ganglion, a true brain (according to Lamarck, in vertebrates it was the brain stem), was responsible for the integration of sensory inputs that produced feelings, and these feelings could then guide and coordinate behaviors by communicating with the more ancient motor ganglia. During the third stage, the two cortical hemispheres on top of the primary brain, which enable feelings to be combined and reorganized into thoughts, emerged.

Lamarck was well aware that the ability to have feelings was generally assumed to be an immaterial faculty, and he therefore emphasized again and again that he regarded it as the effect of an entirely physical, and rather simple, process. He claimed that the faculty of receiving sensations from the sense organs and the brain constitutes a feeling, a physical sensibility,[16] and that this is the hallmark of enbrained animals. Sensations, at their simplest, arrive directly from the sense organs via the nerves to the "sensory nucleus" in the brain, get distributed to the whole body, and come back again to be resent to the sensors from which they had originally arrived. Sensation is never a local process; it is always holistic. The mechanism of sensation

those integrated manifestations of the inner feeling evoked by the fundamental motivations that lead to action: anger, fear, joyous excitement, and so on. To these almost universal motivations/needs, he added "moral needs," such as moralistic anger, which are the products of thought and are characteristic of man.

Lamarck's "list" of characteristics of a state of consciousness or subjective experiencing are therefore: (1) a centralized nervous system with a sensory nucleus (where all sensations converge) and a rich array of afferent and efferent fibers through which electricity flows; (2) physiological integration processes that occur in the center of communication, relating the internal state of the whole organism to sensations arriving from the external world; this integration constitutes the inner feeling, which leads to interactions with the motor center(s) and the induction of actions; (3) value systems, or needs, the satisfying of which determines how the animal behaves. A point or zone of transition to consciousness was not spelled out by Lamarck, and the detailed scenario of evolution is unclear. It *is* clear, however, that Lamarck assigned some rudimentary ability for subjective experiencing to invertebrates possessing head ganglia and obvious sensory organs, although he thought that the "radiata" (e.g., sea urchins) do not experience, since they lack a head ganglion. According to Lamarck, once the cortical hemispheres started to evolve, higher intellectual functions, including reasoning and the ideas, emerged. This obviously happened in vertebrates and was a slow process, with man being its present culmination.

The mechanisms that drove the evolution of animal mentality, from sea urchins that merely coordinate their movements to humans who can reflect about evolutionary processes, were, Lamarck suggested, the already familiar processes underlying the transition from habits to instincts: in novel conditions, needs generate new positively or negatively reinforcing behaviors; new behaviors lead to new habits; and new habits become, in time, new instincts, supported by new anatomical structures. Instincts, Lamarck maintained, are neither automatic, passive reactions like those of plants or sea urchins nor "innate ideas," a notion he thought ridiculous; rather, they are compelling tendencies to behave in ways that satisfy the needs that in past conditions led to the persistent habits of the ancestors.

From feelings, instincts, and skills (which in invertebrates are just trains of instincts), Lamarck went on to derive higher mental faculties like thinking, imagining, and judging. Lamarck suggested that in order for thoughts to exist, there must be a supporting organ: the cortex of the brain or, as he called it, the "hypocephalon." This is necessary because it is in this highly complex and differentiated organ that the sensations and perceptions can

be engraved; it is on this special and soft neural material that they leave memory traces. The simplest thought was, for Lamarck, the recalling of sensations, when as a result of an "effort of the inner feeling," an effort that is an act of attention, the traces of past sensations are activated. Lamarck's emphasis on attention as a prerequisite for having ideas may appear to have modern overtones, and it is indeed one of his many intuitive leaps. However, it is important to note that attention was for him a result of neural fluid dynamics. It was the effect of the inner feeling transporting and directing the available part of the nervous fluid toward the cortex, making it ready to retrace and thus reawaken old ideas or create new traces, new ideas.[18] Complex ideas, he suggested, arise when different traces become coactivated and thus associated, and the animal combines and manipulates them.

Lamarck's assumption about the dynamics of the subtle fluids in the nervous system led him to the conclusion that thought, both simple and complex, is an action and process[19] and that some feelings (e.g., moral feelings) are the natural results of thought, although in animals without a cortex, feelings exist without thought. However, thought always depends on sensation because by definition it is only when sensations become engraved in the cortex, and the engraved pattern is reactivated, that thought arises. Moreover, for Lamarck this continuity between sensation and thought in encephalized animals meant that the mental evolution of these animals followed the same associationist principles as those of simpler cortexless animals. He suggested that the ideas produced in the cortex can become associated when the fluids pass through the cortical areas in which they are engraved at the same time or in succession. Moreover, complex mental habits can develop too. For example, in man, education creates the habit of thinking: fixing attention on many things, comparing, combining, and so on, become habits of thought. The evolution of man's mentality is a result of the accumulation of many such complex, open-ended mental habits, which were greatly reinforced and sophisticated by language. Will is an intellectual act, dependent on the ability to judge, and hence there is no free will, claimed Lamarck: the illusion that we have free will is the consequence of the great number of factors that impinge on our judgments, many of which we are hardly aware.

Lamarck's ideas about the evolution of mentality received very little attention, even from those nineteenth-century evolutionists who were interested in the same questions, such as Herbert Spencer, Charles Darwin, and William James. When they did refer to Lamarck, it was his general evolutionary theory, especially use and disuse, that they discussed, not his ideas about the evolution of the mind. Perhaps this is understandable because

his eighteenth-century neurophysiology was behind the times, and he left open some big and obvious questions, such as the origins of the first nervous system and the evolutionary origins of the cortical hemispheres. But Spencer, Darwin, and James were people who delighted in open questions. Did they read part III of Lamarck's *Philosophie zoologique*, or did they lose patience as they struggled through his outmoded neurophysiology? As we show in the next sections, there are some remarkable parallels between some of Lamarck's ideas about the nature and evolution of feelings and those developed later, especially the ideas put forward by Spencer and James, although both went, in many ways, far beyond Lamarck.

Herbert Spencer and the Evolutionary Principles of Psychology

Herbert Spencer (figure 2.3, *left*) was undoubtedly the most influential evolutionary philosopher in the nineteenth century—as important as Charles Darwin in the dissemination of evolutionary ideas and with far wider popular appeal. He was read widely by both professional academics in various disciplines and by lay people and had a huge influence on people's domestic life, on legislators, on education, on the establishment of the disciplines of psychology and sociology, and on the curricula of universities. Serious young lovers, like the young Ivan Pavlov and his fiancée in czarist Russia,[20] cemented their relationship by reading Spencer together (figure 2.3, *right*), factory workers discussed his work, and civil servants organized their lives in line with his philosophy.

The dramatic decline in Spencer's influence in the twentieth century—a decline to the point of near extinction among laypeople and contempt among academics—is an interesting and somewhat distressing story in the sociology of knowledge, a field that Spencer himself helped to establish. To a large measure, Spencer's decline can be traced to changes in the social and political milieu (including political-sociological thought), but it is also partly due to his adherence to unfashionable Lamarckian notions and his idiosyncratic "outsider" personality, which recoiled from any institutional associations.

We know quite a lot about Spencer's life and about his personality. He not only wrote an autobiography but was also associated with superb writers and observers of human nature, and his great fame ensured that people recorded their meetings with him and reported on his many eccentric behaviors. Spencer, completely focused on his work and on his own affairs, was a hypochondriac walking around with earmuffs, checking his pulse and retiring to rest his delicate nerves even when he invited guests to his

Figure 2.3
Spencer and his influence. *Left*, Herbert Spencer (1820–1903). *Right*, reading Spencer together as described by Chekhov's (1891/1916) *The Duel*, p. 7: "To begin with we had kisses, and calm evenings, and vows, and Spencer, and ideals and interests in common."

rented house (he never owned one). He refused most of the many honors that were offered to him, was obsessively concerned with issues of priority and with real and imaginary misunderstandings of his work, and always voiced his opinions, however unpopular.

The Law of Progressive Evolution

Evolution, according to Spencer, is a process of progressive complexification—a process inherent in any physical system, from atoms to societies. It is not extrinsically teleological: it is the result of obedience to physical principles, as inevitable as the fall of an apple to the ground because of the force of gravity. Spencer started from the reasonable assumption that a state of nature is, in the beginning, simple and relatively homogenous. The condition of homogeneity is, however, unstable because environmental

conditions, which are never perfectly symmetrical, impinge on living matter, and therefore the original homogenous state cannot last. Because physical entities are dynamic and inevitably interact, the state of the system changes, and since any single cause leads to more than one interaction and hence to more than one effect, it follows that an increase in heterogeneity is inevitable.[21] Since Spencer assumed that the heterogenous state is more stable than the homogenous one, he concluded there was a universal tendency toward increased heterogeneity and complexity—for progressive evolution in the broad sense. The arrow of time and the arrow of increasing complexity were coincident.

Spencer provided a general definition for this universal law of progressive change: "Evolution is definable as a change from an incoherent homogeneity to a coherent heterogeneity, accompanying the dissipation of motion and the integrations of matter."[22] How motion is dissipated and how integration occurs differ for different systems. In the bodies of living organisms, just as in animal and human societies, the dissipation and integration take the form of increased division of labor, exemplified by Karl Ernst von Baer's law of embryological development—the progression from the general and simple to the more complex and specialized. Individual organisms and groups of organisms can respond to challenges, such as diminished resources, with a more efficient division of labor.[23] Spencer's response to Thomas Robert Malthus's gloomy predictions about overpopulation was that when populations grow and resources run out, the group reorganizes, new specializations are formed, and the better use and production of resources that follow can support the increased group size. Moreover, he argued, complex animals tend to reproduce less than simpler ones because there is a trade-off between the fertility and complexity of organization, so the problem of overexploitation of resources further diminishes. Interestingly, toward the end of his early paper discussing this, Spencer (1852) suggested that differential survival of the better adapted was another result of the challenge of diminished resources, but he did not develop this suggestion. When Darwin's *On the Origin of Species* was published in 1859, Spencer recognized the power of this idea and regretted that he had overlooked its importance. From then on he regarded Darwinian natural selection as one of several factors driving evolution, usually assigning to it the negative role of weeding out the unfit. Lamarckian adaptation through use and disuse still remained central to his view of evolutionary change, especially with respect to complex ("higher") neural animals, which used their behavioral-neural plasticity to cope with the environment during their own lifetime, something that eventually led to the formation of new heritable habits.

subsequently recurs there is a certain tendency for the second to follow."[27] He illustrated this principle with the example of a shadow that becomes associated with touch. A touch normally elicits a reflex contraction response from an animal, and if a shadow is systematically associated with the touch, in time a link is formed in the nervous system, and the shadow alone will elicit a contraction response. In the second edition of *The Principles of Psychology*, Spencer introduced an additional type of associative learning, suggesting that successful (pleasurable) effects lead to the reinforcement of the actions that brought them about: "On the recurrence of the circumstances, these muscular movements that were followed by success are likely to be repeated; what was at first an accidental combination of motions will now be a combination having considerable probability."[28] This idea is the same as that suggested by Alexander Bain (1818–1903), one of the founders of modern psychology and a close associate of Mill, whom Spencer knew personally and whose books he read. Bain followed in Hartley's footsteps and explored the relations between neurological and psychological processes, stressing the role of early development and learning in the formation of complex behavior, although he did not introduce evolutionary principles. In his books *The Senses and the Intellect* (1855) and *The Emotions and the Will* (1858), Bain stressed the continuity between reflex actions, spontaneous activity, and learned behaviors. He suggested that beginning very early in development, shortly after birth, motor acts, which are at first spontaneous and chaotic, become organized: some actions become reinforced because they are pleasurable, while others are suppressed because they are unpleasant.[29]

Spencer did not attribute reinforcement learning to Bain, possibly because he regarded his own (critical) endorsement of Mill's associationism and his own discussion of learning by association in the first edition of *The Principles of Psychology* as covering this case. In the second edition, however, reinforcement learning, which in time became known as the Spencer-Bain principle (and in the twentieth century, as instrumental or operant conditioning), was given center stage. Spencer not only employed this principle to explain how complex behaviors and habits are formed during ontogeny but he also speculated about the neural basis of associative learning, something that Bain refused to do because he knew little about neurology. Spencer explained that as a result of pleasure accompanying particular motor acts (e.g., suckling), the neural routes of communication become active, facilitating the subsequent passage of neural discharges leading to the same action. If the association is very stable and persists for several generations, a new instinct will emerge. Moreover, he explained that the categories

of pleasure and pain are themselves products of evolution: beneficial and deleterious experiences have become associated with feelings that were categorized as pleasant or painful, respectively. This categorization, Spencer claimed, resulted from natural selection. He considered natural selection to be especially important during the early stages of mental evolution, although it always accompanies and reinforces evolutionary changes in the nervous system that are brought about through the inherited effects of use and disuse.

Spencer suggested that the unit of neural evolution was the "reflex action," a term describing nerve-muscle interaction, which had become well known following the neurophysiological studies of Marshall Hall in 1831. However, although he considered it to be a core building block of neural activity, Spencer did *not* consider the reflex as a unit of consciousness; it was a basic *neural* process from which other more complex processes evolved. Reflex action, which he described as a single contraction following a single sensory irritation, already presupposes a differentiation between afferent sensory and efferent motor nerves, a center of communication, and a contractible muscle. The sensory stimulus is carried to a ganglion (the communication center) through the afferent nerve and is reflected from it through an efferent nerve that elicits a muscle contraction. An instinct, which may be quite elaborate, is a compound reflex action: it is the result of several reflexes occurring in parallel and in succession following relatively simple stimulation. Complex instincts are, in turn, the platform on which the first vague and preliminary manifestations of consciousness appeared. The processes of composition and coordination that are involved in eliciting and executing highly complex instincts become less determined; conflict between different neural connections may occur, the system may "hesitate" as the most coherent representations become established, and as a result the elicitation and the response cease to be entirely automatic. These neural processes take time, and the ongoing activity of integrated sensory stimulations leads to Feelings, the basic units that constitute all forms of consciousness and include both felt sensations and perceptions: "It [the mind] consists largely, and in one sense entirely, of Feelings. Not only do Feelings constitute the inferior tracts of consciousness, but Feelings are in all cases the materials out of which, in the superior tracts of consciousness, Intellect is evolved by structural combination."[30]

Basic feelings were, however, just one facet of conscious activity. Spencer suggested that feelings, memory, reason, and will are different facets of the same psychical activity that constitutes the conscious (Feeling) state, which is characterized by being highly complex and hence inevitably nonautomatic and temporally "extended." For Spencer, as for Lamarck, "memory"

means not merely the reactivation of neural traces (neural internal relations) of past responses but active recollection, a mental search process involving conflicts and competition between different neural representations (the "inner relations" that correspond to the "outer relations"). Because these activities take time, the activity can be rendered conscious; Spencer firmly believed that for an inner neural state to become conscious, the neural activity has to have some minimal temporal persistence. The strongest links between neural representations are those that in the past occurred most frequently, and hence the "winning" relations are likely to be those best adapted to these circumstances. Spencer noted that "an [internal] action thus produced, is nothing else than a rational action,"[31] thus equating rationality with adaptation. Will, too, is a facet of the same process of neural reactivation, association, and searching among competing neural paths. Will involves the reactivation of an internal neural representation of a motor change: when the represented act is reactivated (recalled) and is in the process of becoming an actual motor act, the experience of willing emerges. Needless to say, with such a neurological notion of the will, the idea that there is free will was for Spencer a nonsensical, meaningless notion. There is a feeling of free will due, just as Lamarck had said, to the multiplicity of factors, few of which we are aware of, that lead to actions. Moreover, as Spencer emphasized, though we are free to do what we desire, we are not free to desire or not to desire. Free will is an illusion, a reification of a Feeling, which is then believed to have an independent existence.

Spencer's evolutionary psychology is rich and complex, and in this chapter we are merely pointing to some elements that are of particular relevance to his views about the evolution of Feeling. However, like Lamarck, Spencer did not provide a particular scenario for the evolutionary emergence of Feeling. What he did do was provide a detailed functional account of how, in general terms, subjective feelings and related conscious faculties progressively evolved from reflexes, their combinations and coordination, the coordination of coordinations, and so on. The account of how nerves and ganglia, the tissues involved in reflex actions and instinctive acts, themselves originated and evolved was presented with ample speculative detail in the 1870 second edition of *The Principles of Psychology*. We can infer from this account that he believed that some invertebrates that have complex sense organs and manifest complex behavior (e.g., bees, cephalopods) probably have some elementary ability to subjectively experience; that all vertebrates manifest basic consciousness; and that some vertebrates (those with a cortex or its equivalent) also have the ability to form representations of representations—that is, abstract ideas.

Spencer, like Darwin, was interested in the expression of emotions, in the "language of emotions," as he called it, and this interest seems to have stemmed from his love of physical beauty and of music, which were for him the supreme manifestation of the expression of complex, and in this case utility-free, emotions. Like everything else he ever considered, he analyzed the expression of emotions, including the social emotions, from an evolutionary point of view. Some of these analyses were published in articles he wrote in the 1850s and 1860s, but a summary with further extensions formed a chapter in the second edition of *The Principles of Psychology*, which appeared shortly before Darwin's *The Expression of Emotions* was published. In that chapter, Spencer suggested that the expression of emotions—facial expression, vocalizations, gestures—result either from "diffuse neural discharge," such as that seen when overwhelming joy or overwhelming anger overtakes one (when the neural discharges flow along the least resistant neural paths), or from "restricted discharge." Restricted discharge, which is far more specific, can be either undirected—the result of evolutionarily established instincts and their underlying neural structures—or directed—the result of deliberate voluntary activities. As we discuss in the next section, Darwin had very similar ideas. According to Spencer, the evolution of the diffuse and unguided restricted discharges followed the use and disuse principle. He explained, for example, that expressions of fear are reactions that first appeared, for good adaptive reasons, in combat situations, and because of their functional adequacy they became innate. Nevertheless, although he thought that the expressions of many emotions are underlain by innate dispositions, he believed that humans possess a considerable ability to control some of them. Human emotional expressions are a complex mix of voluntary and involuntary discharges and motors actions.

Spencer's general evolutionary theory, including his evolutionary psychology, is largely forgotten today; when mentioned, with the exception of a few historians, it is treated with derision. Ernst Mayr, the most dominant evolutionary biologist in the second half of the twentieth century, devoted only three paragraphs to Spencer in his massive and hugely influential 1982 book about the history of biological thought because, he claimed, Spencer's "positive contributions [to evolutionary theory] were nil."[32] We beg to differ.

Charles Darwin and the Mental Continuity between Man and Animals

In sharp contrast to Spencer, Charles Darwin (figure 2.4) is idolized by biologists, philosophers, and historians, and everything that he wrote is treated with an almost religious reverence, an attitude that we suspect would have

Figure 2.4
Charles Darwin (1809–1882), with an unmistakable expression of tenderness and love, and his son William Erasmus Darwin in 1842. Reproduced by permission of Cambridge University Library.

deeply embarrassed and disturbed him. There is, however, little doubt that Darwin's ideas are crucial to any evolutionary account we may offer, so although Darwin himself dwelled neither on the evolutionary origins of the ability to experience nor on the evolution of the nervous system, he did contribute to our understanding of the evolution of mentality. First, he developed and firmly established the idea of descent with modification. This idea was not original to him: as already indicated, both Lamarck and

spread in the hominid lineage—and with them anything that enhanced cooperation, such as emotions of shame and guilt, which in turn further reinforced human cooperative social life. Sexual selection, too, helps to explain the evolution of enhanced mental faculties. For example, Darwin explained that human males, who compete for females and are chosen by them, evolved to be ever more strong and intelligent; because of male choice, females evolved to become more beautiful.

Darwin left no doubt that both sexual and group selection are found in other animals. Selection among family groups, for example, is used to explain the evolution of social insects' caste organization and altruistic behavior, and the whole second part of *The Descent of Man* is devoted to sexual selection in animals, starting with insects and ending with humans. Sexual selection, especially mate choice, explains the evolution of certain secondary sexual characteristics and differences among members of different groups ("races") that cannot be explained by natural selection for survival. Sexual selection also suggests that the "lower" animals possess mental powers: intelligence is useful when competing for mates, and mate choice implies a mental preference. Darwin wrote in the first chapter of *The Descent of Man* that "with respect to animals very low in the scale, I shall have to give some additional facts under Sexual Selection, shewing that their mental powers are higher than might have been expected."[36] When he found evidence for sexual selection in a taxon, it meant, to Darwin, that its members are endowed with will, desire, and choice. Darwin found evidence for sexual selection (and the associated mental powers) not only in all vertebrates but also in several groups of insects, notably the Homoptera, Orthoptera, and Coleoptera. He found no evidence for it in the protists, coelenterates, echinoderms, and annelids, an absence that confirmed his belief in the "lower mental powers" and the "imperfect senses" of these animals. However, Darwin did not regard the absence of sexual selection as definite evidence for a lack of mental powers, for he found no evidence for sexual selection in cephalopods, whose high intelligence he recognized.

In *The Expression of the Emotions in Man and Animals*, Darwin discussed a somewhat less controversial and far more empirically accessible topic—the similarity in the way emotions are expressed in animals and humans. The emotions that are expressed, unlike the emotions one feels, are in the public, observable ("objective") domain; they are not inferences from behavior, they are behaviors. They can be systematically studied and compared by (1) observing infants and blind people, who cannot imitate from others the expression of the emotions they feel, thus indicating which expressions

of emotions are innate; (2) observing the expression of emotions in animals, especially the higher apes, which can demonstrate the similarity and evolutionary continuity between apes and man in this respect; (3) observing the insane, whose voluntary control of their emotions is reduced and who therefore express emotions (relatively) unmodified by social norms; (4) scrutinizing works of art, which, because of the talent of artists, can highlight wide ranges and aspects of expression; and (5) comparatively studying the expression of emotions in widely different cultural groups using photographs and questionnaires. The use of photographs is a novel methodology that Darwin introduced for comparing the way emotions are expressed and, with modern modifications (videos), is still in use today.

The principles that, according to Darwin,[37] underlie the expression of emotions all stem from the associationist ideas that a response to a stimulus that is spatially or temporally contiguous with, or contrasts to a stimulus that elicited the original expression, becomes in time, after much repetition, innate. Darwin suggested three principles: "The principle of Serviceable Associated Habits.—Certain complex actions are of direct or indirect service under certain states of the mind, in order to relieve or gratify certain sensations, desires, &c.; and whenever the same state of mind is induced, however feebly, there is a tendency through the force of habit and association for the same movements to be performed, though they may not then be of the least use."[38] Expressions of aggression (anger, hate, spite), which are not only grounded in the physiology of the animal but, like the baring of the canines, were (and may still be) useful in situations of conflict, are good examples of this principle at work. The similarity to Spencer's "undirected restricted discharge" is obvious. The second principle is "The principle of Antithesis.—Certain states of the mind lead to certain habitual actions, which are of service, as under our first principle. Now when a directly opposite state of mind is induced, there is a strong and involuntary tendency to the performance of movements of a directly opposite nature, though these are of no use; and such movements are in some cases highly expressive."[39] Examples are the expressions of submission in dogs, or affection in cats. Here Darwin used the associationist law of contrast, extrapolating from the association of contrasting ideas to the association of contrasting expressions of emotions. He suggested that expressions that are very different from those in the first category and are perceived by observers (mates, members of the same social group) as manifestations of *opposite* states of mind will elicit contrasting actions. Although the communicative function of the expression of emotion is not stressed by Darwin, he suggested that

the principle of antithesis is best explained by assuming that, although at first it was the by-product of directly serviceable expressions, it later evolved for communication.

Darwin's third principle, "The principle of actions due to the constitution of the Nervous System, independently from the first of the Will, and independently to a certain extent of Habit" was influenced by Spencer's principle of "diffuse neural discharge." Examples include overwhelming emotions, such as horror, that lead to trembling, secretions from glands, defecating, and more, which never had any functional, selectable significance. Of course, Darwin, like Spencer, was fully aware that observed expressions are often the result of a combination of these different principles.

Darwin suggested that the processes that led to the evolution of the various human expressions of emotion (figure 2.5 illustrates some of them) were first and foremost the inherited effects of use and disuse—and to some extent also natural selection. *The Expression of the Emotions in Man and Animals* is the most "Lamarckian" book Darwin ever wrote. Paul Ekman, in his afterword to a 1998 reprint of the third edition, suggested that this is one of the reasons for the book's decline in popularity in the first two-thirds of the twentieth century. The other reasons are, according to Ekman, Darwin's reliance on anecdotal evidence; his relatively scant attention to the communicative function of the expressions; the rise of behaviorism, which focused on learned behavior and ignored mental states; and the dominance of cultural relativism in the first half of the twentieth century, which rendered universalistic claims about the expression of emotions or any other human attributes suspect.[40]

In spite of Darwin's refusal to speculate about the origins of animal mentality, we want to stress four points that are relevant to our topic. First, Darwin believed that many invertebrates are endowed with feelings and mental states, some exhibiting will, aesthetic preferences, and desires. Thus, for him the evolution of consciousness has deep phylogenetic roots. Second, he believed that the expression of emotions and the feelings that constitute emotions are related. "He who gives way to violent gestures will increase his rage; he who does not control the signs of fear will experience fear in a greater degree," he wrote on the last page of *The Expression of the Emotions in Man and Animals*, thus paving the way for James's theory of emotions. Third, natural selection, as one of the processes driving mental evolution, played a role. Fourth, although himself unwilling to tackle the subject, Darwin relegated the study of the evolution of mentality to his young successor and protégée George Romanes, who wrote three books about it.[41] Following

Figure 2.5
Expressions of emotions.

Darwin, all psychologists, philosophers, and biologists who considered mental evolution and the evolutionary origins of mentality explained it in terms of natural selection. William James's view of consciousness, which is the basis of twenty-first-century ideas on the topic, incorporates Darwin's theory of evolution by natural selection and challenges Spencer's account of the evolution of consciousness.

The Psychological Investigations of William James

Modern psychology sprang from the introspective psychophysics that developed in Germany in the second half of the nineteenth century, from the associationist philosophical psychology of John Stuart Mill and Alexander Bain, and from the evolutionary psychology of Spencer and Darwin. It reached a magnificent and idiosyncratic peak with the work of William James (figure 2.6), who integrated the three strands, but the study of mentality fell into sharp decline for sixty years with the reign of behaviorism in North America. It slowly came back to life and entered through the back door with the advent of the cognitive revolution and became more assured following the convergence of several paths of inquiry in the early 1990s. Here we focus on James's contribution to the understanding of consciousness—on *The Principles of Psychology*, which is a challenge and alternative to Spencer's identically titled book. James's book is not only full of remarkable introspective insights—it also provides, despite its many omissions, the foundations for a twenty-first-century science of consciousness.

A charmer—imaginative, artistic, rebellious, generous, self-centered, and often exasperating—William James was loved and admired by many people, and several excellent biographies have been devoted to him. His first biographer and former student, Ralph Barton Perry, pointed to his defiant spirit: "A natural poacher, with the poacher's characteristic dislike of the gamekeeper."[42] He was described by his sister Alice as "a creature who speaks in another language as H [Henry James, their brother] says from the rest of mankind and who could lend life and charm to a treadmill."[43] His brother Henry, mourning his death, wrote to H. G. Wells of his loss: "He did surely shed light on man, and gave, of his own great spirit and beautiful genius, with splendid generosity."[44]

The Principles of Psychology: Consciousness as a Selecting Agency

James's motivation in writing *The Principles of Psychology* was very different from that which guided Lamarck's and Darwin's writings on psychology. *The Principles of Psychology* is *not* a book about mental evolution: although

beautifully expressed by Thomas Huxley and quoted in *The Principles of Psychology*:

> The consciousness of brutes would appear to be related to the mechanism of their body simply as a collateral product of its working, and to be as completely without any power of modifying that working as the steam-whistle which accompanies the work of a locomotive engine is without influence on its machinery. … It seems to me that in men, as in brutes, there is no proof that any state of consciousness is the cause of change in the motion of the matter of the organism. If these positions are well based, it follows that our mental conditions are simply the symbols in consciousness of the changes which take place automatically in the organism; and that, to take an extreme illustration, the feeling we call volition is not the cause of a voluntary act, but the symbol of that state of the brain which is the immediate cause of that act. We are conscious automata. (James 1890, 1:131)

Though it sounds convincing, Huxley's reasoning, James argued, is philosophically problematic. First, the argument from continuity can work both ways: since we humans are aware of feeling, of focusing attention, of willing and of thinking, and these mental states, James believed, guide our actions, we can extrapolate and attribute this causal efficacy to the brutes' mental states. Second, the assumption that feelings and thoughts do not have causal power, but muscles and nerves do, is philosophically weak: ever since the days of Hume and Kant, James argued, we have been aware that we do not have a very good idea of what causality—physical or psychic—actually is, so making dogmatic and confident statements about material causality while denying causal power to feelings and thoughts is rather hasty: "As in the night all cats are gray, so in the darkness of metaphysical criticism all causes are obscure."[46]

More importantly, according to James, there are positive reasons for believing that consciousness has causal power. Higher animals have more complex brains than lower animals and seem to be more intelligent and more conscious. However, brain complexity breeds problems. James argued that an increase in brain complexity makes it unstable and prone to mistakes, and this vulnerability points to the function of consciousness: it is consciousness that "loads the dice" and guides and increases the efficiency of the unstable brain:

> Loading its dice would mean bringing a more or less constant pressure to bear in favor of those of its performances which make for the most permanent interests of the brain's owner; it would mean a constant inhibition of the tendencies to stray aside. (James 1890, 1:140)

James advanced a series of arguments to support this view. First, the distribution of consciousness and the fact that higher animals are more conscious than lower ones reinforces the idea that consciousness evolved gradually through natural selection because of the guidance it provides to animals with increasingly large and unstable brains. No doubt Huxley would have answered that the neurophysiological machinery had progressively evolved to be more complex and, inevitably, so had its epiphenomenal, mental "shadow." A second and more convincing observation supporting the causal effects of consciousness is that it disappears with habit. Habit, for James, was a result of the brain's plasticity, which was for him, as it was for Lamarck before him, a result of the flexibility of the material of the brain, a physical property of the matter of brain tissues that allows currents to carve and shape them. The fact that habit simplifies the movements necessary to accomplish an action and diminishes the attention needed for it supports the view that when neural reactions become, through habit, simple and reflex-like, consciousness disappears—it is no longer needed to guide action. A positive argument for the causal power of consciousness is that we are aware of being conscious when we are learning something new or are faced with a difficult decision. Consciousness is most conspicuous when we dither and deliberate (a very Jamesian state of mind), which is a major characteristic of the conscious effort to decide and act upon the decision. The goal-directed nature of consciousness is particularly evident when there is an obstacle or impediment in an animal's (including human's) way—for example, when a human becomes handicapped and finds alternative ways to achieve her goals. Our felt needs, especially our conscious goals, direct our actions. Finally, the fact that the mental state of pleasure usually comes with activities that improve the chances of surviving and leaving offspring (fitness-increasing activities), while the mental state of pain accompanies states and activities that decrease fitness, suggests that these mental states are not mere epiphenomena but have a clear adaptive function. All these suggest that we are not passive mirrors, merely reflecting the relations that exist in the environment. We select among them and make them relevant. The agency of organisms, especially of humans, is central to their everyday existence as well as to their evolution.

How does this "loading the dice" actually happen? James suggested a general principle: selection, in the case of consciousness, is the result of differential stabilization among the endless variable activities in the nervous system; the telos of consciousness is to satisfy, through elicited and stabilized feelings and thoughts, the goals of the animal. Just as reproductive

success is both the outcome and the telos of Darwinian organisms, so the manifestation of human consciousness is both the outcome and the telos of feeling and thinking. The ultimate telos is still survival, but there is now a mediating telos, a felt desire to survive:

> Every actually existing consciousness seems to itself at any rate to be a *fighter for ends*, of which many, but for its presence, would not be ends at all. Its powers of cognition are mainly subservient to these ends, discerning which facts further them and which do not. (James 1890, 1:141)

James did not tell us when and how consciousness emerged during evolutionary history—at what point the brain becomes too complex to work without the "guidance" of such selection and what the neural activity that enables it is. Nor did he tell us whether the notorious polyp has, by extrapolation from humans, a measure of consciousness. But the idea of selection in the brain as a necessary aspect of what we may call "mental life" is dominant in *The Principles of Psychology* and is, as we discuss in later chapters, a recurring theme among late twentieth- and early twenty-first-century philosophers like Dennett, neurobiologists like Changeux and Edelman, and theoretical biologists like Szathmáry and Fernando. There is something compelling about the idea, for, like natural selection, neural selection occurs without an external designer and can lead to complex adaptations. However, James, unlike modern scholars, never grounded the analogy in *intrinsic* neural selection, in a choice without a chooser; the selector for him was selective attention, which he regarded as a mental selecting activity that expressed itself most clearly when attention and action-selection required an effort. Consciousness is therefore not a stuff, material or transcendental, nor is it a neural mechanism or a property of neurophysiological activity, although it seems to be constituted by and composed of neural activity. It is a very special kind of activity in a whole animal (rather than just the brain of the animal), something the enbrained body does. James would have been firmly opposed to brains in vats and to conscious computer software.

Like his psychology, James's philosophy stresses the creative and unanticipated element in evolution, which is only judged a posteriori. In his course on the philosophy of evolution, James contrasted his position to that of Spencer.[47] While Spencer focused on physical laws and the "adjustments of inner relations to outer relations" through use and disuse, James highlighted the natural selection of chance variations, of agency, and of choice. In the last chapter of *The Principles of Psychology*, there is a strong endorsement of the neo-Darwinian view of evolution by the natural selection of chance variations, a view developed by August Weismann, who denied, on

theoretical and empirical grounds, the inheritance of acquired characters and the role of use and disuse in evolution.[48] Although for James the undirected origin of human-specific logical and aesthetic faculties was of greater concern, undirected variation also held an important place in his general views about the nature of external reality and evolution and was closely related to his notion of consciousness as a creative "voting" element.

Chance and Creativity

In order to distinguish and evaluate the role of directed (environmentally induced) and undirected (stochastic) variation in the construction of mentality, James distinguished between internal and external constructive factors. He admitted that relations in space and time can be learned and that learned behavior can become a habit, so in this sense inner relations may be said to adjust to outer relations. However, it was equally clear to him that certain aptitudes can have either external *or* purely internal (and unlearned) causes: a boy can become a musician because he practices a lot (external influence) or because of a lucky accident in the ovum that made him musically talented with little need for exercise (internal influence). Therefore, both internal (innate) and external (learned) factors can affect brain activity.

But even if we grant external relations some stamping power, how does this happen? The blue of the sky that we see is not a copy of a blue sky up there—there is no blue up there, as we well know. Similarly, pain and pleasure are internal mental categories and have no existence outside us. What we experience as external factors and the relations among them depends on the inner structure of our brain, which at some time during evolutionary history, as a result of a lucky accident, enabled us to make better discriminations than our ancestors and was therefore established through natural selection. The world is not given, nor is it orderly. As John Stuart Mill, whom James cited, put it: "The order of nature, as perceived at a first glance, presents at every instant a chaos followed by another chaos. We must decompose each chaos into single facts. We must learn to see in the chaotic antecedent a multitude of distinct antecedents, in the chaotic consequent a multitude of distinct consequents."[49] James was undoubtedly also influenced by his friend Chauncey Wright's conception of the world. Wright described the order of the world with a weather metaphor, a "cosmic weather" view, as his friends called it. The weather is a physical system we cannot predict, although it is subject to deterministic physical laws. The complexity and unpredictability of the ever-changing initial atmospheric conditions—and (as we would say today) nonlinear interactions—preclude any universal and predictable long-term regularities. Except for

the incessant change brought about by heat and gravitation, everything is transient, directionless.[50]

Given the ceaseless activity and instability of the brain, given the chaos of the world, it is not clear how a baby or a puppy learns to see objects as stable features of the world. James was aware, of course, that Spencer and Mill might agree that once an ability to impose some order on the plenitude of existence (to discriminate between colors, or feel pleasure and pain) becomes established by the natural selection of blind variations, then habit and learning could build on this and lead to the evolution of new and complex faculties. But this, he argued, could work only for simple sense impressions: "The only cohesions which experience in the literal sense of the word produces in our mind are … the proximate laws of nature, and habitudes of concrete things, that heat melts ice, that salt preserves meat, that fish die out of water, and the like."[51] Once we go further—to the human intellectual and emotional faculties; to the ability to compare and abstract; to logical laws, metaphysical generalization, and aesthetic judgment—there is no way that we can deduce these features of our understanding from our sensory impressions. On the contrary, what science and logic do is extract order out of sensory chaos. As every scientist knows, it requires great ingenuity and hard work to create the artificial conditions in which a regularity or a natural law becomes apparent.

Of course, James is right: abstract human categories and discovered laws of nature are not the result of a simple, cumulative combination of sensory associations, a position he attributes, unfairly, to Spencer and to John Stuart Mill.[52] We need something else, and for James this something else was a mode of creative consciousness that can be reduced neither to the external world nor to the brain as a responsive organ. At the phylogenetic level, he suggested that creativity can be best explained biologically as the consequence of random variation and natural selection. But it is not at all clear how random variation in the ability for abstraction can become established in a population. What is the initial function of the selected variation? Our human mental faculties do not simply and directly serve survival and reproduction; James (1878) rebelled against this vulgar and simplistic idea in his anti-Spencer paper of 1878. What led to human-specific faculties must be a by-product of a variation that was useful for other, more mundane reasons. James did not, however, tell us the by-product of what these faculties might be.[53]

It is not just in the case of the evolution of human reflective self-consciousness that James preferred the selection of chance variations to the inheritance of learned associations. James did not subscribe to the

called the "specious present,"[56] is the basis of our intuition of time and is based on a view of consciousness-as-activity. Although this activity depends on the brain, the brain is not sufficient. It is the animal, not the brain, that feels. The sense of self, most fundamentally of the bodily self, is basic to the animal's feeling as an agent, and this feeling of self becomes broader and more intricate as mental development proceeds from birth to maturity.

Instincts, Emotions, and the Will

Instincts are the foundations on which James, like Spencer and Lamarck before him, built his theory of consciousness. However, the nature of instinct and its relation to learning, reasoning, and complex emotions was, for James, very different from that of his predecessors. In *The Principles of Psychology*, he started his chapter on instinct with a definition that makes it clear that instinct is *any bias* for neurally mediated action, an inevitable accompaniment of the bodily construction of neural animals:

> Instinct is usually defined as the faculty of acting in such a way as to produce certain ends, without foresight of the ends, and without previous education in the performance. That instincts, as thus defined, exist on an enormous scale in the animal kingdom needs no proof. *They are the functional correlatives of structure. With the presence of a certain organ goes, one may say, almost always a native aptitude for its use.* (James 1890, 2:383; emphasis added)

Looked at in this way, instincts are everywhere and, as James stressed in this chapter, humans are particularly well endowed with instincts.[57] Following Spencer, he suggested that conflict and competition among different instincts (e.g., the baby's instincts to both approach and recoil from an unknown person) may lead to choice and voluntary action. However, instincts for James are not rigid and fixed. They can almost always be modified by learning, which leads to the formation of habits that sometimes result in the inhibition of the initiating instinct. Thus, the rat's instinct to approach and eat food when hungry can be inhibited if the food is associated with a trap. The learned association, if very traumatic or repeated, will make the rat avoid food in certain conditions. Instincts can be modified by learning in another way: they can become associated with a very narrow range of contexts. For example, a rabbit always deposits its feces in the same corner as that in which it first deposited them. Furthermore, many instincts (for example, the baby's suckling instinct) are transient. James built on the concept of "imperfect instincts," which had been developed by Douglas Spalding (1841–1877), the brilliant Scottish biologist who was the first to examine instincts experimentally, and George Romanes (1848–1894), who laid the

foundations of comparative psychology. Both are often cited in James's chapter on instinct. Spalding emphasized the condition-dependent nature of many instincts: sixty years before Lorenz studied and made famous the behavior of young chicks that persistently follow the first moving object they see after hatching (a behavior called "filial imprinting"), Spalding had experimented on this following behavior. He showed that following depended on the chicks' stage of development when they saw their first moving object, thus emphasizing the importance of maturation for triggering certain instinctive behaviors. Romanes, who performed cross-species fostering experiments in birds, also regarded instincts as "imperfect" and educable. James saw this modifiability of instincts as a basic property: far from being a sign of "imperfection," the modifiability of instincts is their very raison d'être because it enables the development of habits through learning, thus fulfilling the instincts' role and making it redundant:

> *Most instincts are implanted for the sake of giving rise to habits, and that, this purpose once accomplished, the instincts themselves, as such, have no raison d'être in the psychical economy, and consequently fade away*. (James 1890, 2:403)

The role of instincts as the scaffolds on which habits are built is related to James's view of the development of motivations and desires. On its first appearance, an instinct is blind and is elicited by a very simple stimulus, but when experienced for the second time, it comes with past-based expectations that lead to desires and create motivation.

James's view of instincts, especially the way in which initial instincts construct mental life through learning, is intimately related to his theory of emotions (a term he often used interchangeably with feelings). This famous Jamesian theory, which was developed independently by the physiologist Carl Lange,[58] suggests that emotions (e.g., fear) are not the causes of bodily changes; rather, emotions are made up of (or constituted by) bodily (physiological) changes and follow or accompany them. This theory is based on two assumptions: first, that the whole organism is the sounding board for whatever excites the nervous system[59] and second, that these bodily changes are felt as they occur. From these assumptions James came to his theory, which he contrasted with conventional wisdom, which regarded emotions as the causes of actions and bodily changes:

> My theory, on the contrary, is that the bodily changes follow directly the perception of the exciting fact, and that our feeling of the same changes as they occur IS the emotion. Common-sense says, we lose our fortune, are sorry and weep; we meet a bear, are frightened and run; we are insulted by a rival, are angry and strike. The hypothesis here to be defended says that this order of sequence is