From: Mindblindness: an essay on
autism and theory of mind.
Simon Baron-Cohen. 1995.
Cambridge, MA: MIT Press.

Learning, Development, and Conceptual Change

Lila Gleitman, Susan Carey, Elissa Newport, and Elizabeth Spelke, editors

John Macnamara, *Names for Things: A Study in Human Learning* (1982)

Susan Carey, *Conceptual Change in Childhood* (1985)

David Premack, *"Gavagai!" or the Future History of the Animal Language Controversy* (1986)

Daniel N. Osherson, *Systems That Learn: An Introduction to Learning Theory for Cognitive and Computer Scientists* (1986)

James L. Morgan, *From Simple Input to Complex Grammar* (1986)

Frank C. Keil, *Concepts, Kinds, and Cognitive Development* (1989)

Steven Pinker, *Learnability and Cognition: The Acquisition of Argument Structure* (1989)

Kurt VanLehn, *Mind Bugs: The Origins of Procedural Misconception* (1990)

Ellen M. Markman, *Categorization and Naming in Children: Problems of Induction* (1990)

Henry M. Wellman, *The Child's Theory of Mind* (1990)

Charles R. Gallistel, *The Organization of Learning* (1990)

Josef Perner, *Understanding the Representational Mind* (1991)

Eleanor J. Gibson, *An Odyssey in Learning and Perception* (1991)

Annette Karmiloff-Smith, *Beyond Modularity: A Developmental Perspective on Cognitive Science* (1992)

Simon Baron-Cohen, *Mindblindness: An Essay on Autism and Theory of Mind* (1995)

Mindblindness

An Essay on Autism and Theory of Mind

Simon Baron-Cohen

A Bradford Book
The MIT Press
Cambridge, Massachusetts
London, England

Tooby, J, & Cosmides, L. (1995). The language of the eyes as an evolved language of mind. Forward to: *Mindblindness: An essay on autism and theory of mind.* By Simon Baron-Cohen. Cambridge, MA: MIT Press.

Foreword

Just as common sense is the faculty that tells us that the world is flat, so too it tells us many other things that are equally unreliable. It tells us, for example, that color is out there in the world, an independent property of the objects we live among. But scientific investigations have led us, logical step by logical step, to escape our fanatically insistent, inelastic intuitions. As a result, we know now that color is not already out there, an inherent attribute of objects. We know this because we sometimes see physically identical objects or spectral arrays as having different colors—depending on background, circumstance, and context—and we routinely see physically different spectral arrays as having the same color. The machinery that causes these experiences allows us to identify something as the same object across situations despite the different wavelength composites that it reflects from circumstance to circumstance. Far from being a physical property of objects, color is a mental property—a useful invention that specialized circuitry computes in our minds and then "projects onto" our percepts of physically colorless objects. This invention allows us to identify and interact with objects and the world far more richly that we otherwise could. That objects seem to be colored is an invention of natural selection, which built into some species, including our own, the specialized neural circuitry responsible.

What is true for color is true for everything in our experienced worlds: the warmth of a smile, the meaning of a glance, the heft of a book, the force of a glare. Although it is a modern truism to say that we live in culturally constructed worlds, the thin surface of cultural construction is dwarfed by (and made possible by) the deep underlying strata of evolved species-typical cognitive construction. We inhabit mental worlds populated by the computational outputs of battalions of evolved, specialized neural automata. They segment words out of a continual auditory flow, they construct a world of local objects from edges and gradients in our two-dimensional retinal arrays, they infer the purpose of a hook from its shape, they recognize and make us feel the negative response of a conversational partner from the roll of her eyes, they identify cooperative intentions among individuals from their joint attention and common emotional responses, and so on.

Each of the neural automata responsible for these constructions is the carefully crafted product of thousands or millions of generations of natural selection, and each makes its own distinctive contribution to the cognitive model of the world that we individually experience as reality. Because these devices are present in all human minds, much of what they construct is the same for all people, from whatever culture; the representations produced by these universal mechanisms thereby constitute the foundation of our shared reality and our ability to communicate. Yet, because these evolved inference engines operate so automatically, we remain unaware of them and their ceaseless, silent, invisible operations. Oblivious to their existence, we mistake the representations they construct (the color of a leaf, the irony in a tone of voice, the approval of our friends, and so on) for the world itself—a world that reveals itself, unproblematically, through our senses.

Indeed, it is exactly because of their universal and automatic character that we have been blind to the existence of the machinery that constitutes most of the evolved architecture of the human mind—what might reasonably be called our cognitive instincts. Instinct blindness is sanity for the individual, but it has been crippling for scientific psychology. Scientists do not conduct research to find things whose existence they don't suspect. These mechanisms solve the many computational problems involved in constructing the world we deal with so automatically that the scientific community remained unaware for decades that these computational problems existed and were being solved as part of the ordinary functioning of the mind of every normal human being. As a consequence, most of psychology retained its empiricist orientation throughout the 20th century, resting on the assumption that a pre-packaged "world" acted though the senses and through general-purpose learning mechanisms to build our concepts, interpretative frameworks, and mental organization.

In the last two decades, though, scientific psychology has finally begun to slip the bonds imposed by this seductive but misdirecting folk psychology. Cognitive scientists were awakened by a series of encounters with alien minds, whose starkly contrasting designs and surprising incapacities drew attention to previously overlooked natural human competences and to the computational problems they routinely solve. They encountered artificial mentalities in the computer lab that had obstinate difficulties in seeing, speaking, handling objects, understanding, or doing almost anything that humans do effortlessly. They encountered thousands of animal species each of which could solve a striking diversity of natural information-processing problems that other species could not. They encountered the developing minds of infants and children, which forced them to confront the intractable computational and philosophical problems that plague empiricist models of how children acquire knowledge. And they encountered neurologically impaired individuals who displayed unanticipated dissociations of cognitive deficits and abilities. These and a host of other factors alerted psychologists to the necessity for—and to the actuality of—a vast nonconscious realm of evolved, specialized, computational problem solvers that construct and interpret the world.

Instead of viewing the world as the force that organizes the mind, researchers now view the mind as imposing (on an infinitely rich and extensive world) its own pre-existing kinds of organization—kinds invented by natural selection during the species' evolutionary history to produce adaptive ends in the species' natural environment. On this view, our cognitive architecture resembles a confederation of hundreds or thousands of functionally dedicated computers (often called modules) designed to solve adaptive problems endemic to our hunter-gatherer ancestors. Each of these devices has its own agenda and imposes its own exotic organization on different fragments of the world. There are specialized systems for grammar induction, for face recognition, for dead reckoning, for construing objects, and for recognizing emotions from the face. There are mechanisms to detect animacy, eye direction, and cheating. There is a "theory of mind" module, and a multitude of other elegant machines.

These modules appear to be structured very differently from the general-purpose cognitive machinery—"attention," "short-term memory," "category induction," and so on—proposed in the previous generation of models of the mind. In order to solve its characteristic domain of problems, a module is designed to interpret the world in its own pre-existing terms and framework, operating primarily or solely with its own specialized "lexicon"—a set of procedures, formats, and representational primitives closely tailored to the demands of its targeted family of problem. These are the languages of the human mind: diagnostic facial-muscle configurations defined by an emotion-recognition system that maps the facial expressions of others onto models of their internal states; a language-acquisition device whose conceptual primitives include elements such as "noun phrase" and "verb phrase"; a rigid object mechanics that construes the world in terms of "solid objects," relative location, and mutual exclusivity within volume boundaries; social-exchange algorithms that define a social world of agents, benefits, requirements, contingency, and cheating; and—the focus of this book—a "theory of mind" module that speaks of agents, beliefs, and desires and links them to a language of the eyes. This language is generated by still other mechanisms that detect eye direction and feed the data into a variety of social inference modules.

The realization that the human mind is densely multimodular has propelled modern psychology into a new theoretical landscape that is strikingly different from the standard empiricist approaches of the past. In consequence, the outlines of the psychological science of the coming century are getting clearer.

In this new phase of the cognitive revolution, discovering and mapping the various functionally specialized modules of the human brain will be primary activities. Even more fundamentally, psychologists are starting to put considerable effort into making their theories and findings consistent with the rest of the natural sciences, including developmental biology, biochemistry, physics, genetics, ecology, and evolutionary biology: Psychology is finally becoming a genuine natural science.

The cognitive revolution solved many of the ontological problems that had prevented psychological concepts from being located with respect to the other sciences. (What manner of thing, after all, was a mental image or an inference or a goal, next to oxidation or mass or receptor sites?) As a result, the psychological architecture can now be mapped—simultaneously and complementarily—as a system of computational relationships and as a physical system that implements these relationships. As the operation of the genetic code is tracked through molecular biology and cell biology to developmental neurobiology, the processes that organize the developing nervous system are becoming increasingly intelligible. These developmental programs were "designed" by selection to build a physical structure that realizes certain functional informational relationships. Discovering what these relationships are is the province of still other fields, such as evolutionary biology and cognitive psychology.

One of the most significant trends in the naturalization of the psychological sciences is the application of data and conceptual tools forged in evolutionary biology, behavioral ecology, primatology, and human paleoanthropology. These fields have begun to contribute an increasingly detailed list of the native information-processing functions that the human brain was built to execute. Detailed theories of adaptive function can tell cognitive scientists what modules are likely to exist, what adaptive information-processing problems they must be capable of solving, and—since form follows function—what kind of design features they can therefore be expected to have. Evolutionary biology and related fields can supply this wealth of guidance because natural selection is the only known natural process that builds functional organization into the species-typical designs of organisms. Consequently, all reliably developing functional mechanisms in a species' psychological architecture must (1) be ascribed to the operation of natural selection, (2) be consistent with its principles, and indeed (3) be organized and specifically designed to solve the narrowly identifiable sets of biological information-processing problems defined by selection operating within the context of a species' ancestral mode of life. For humans, of course, this means the world of ancestral hunter-gatherers, foraging hominids, and even pre-hominid primates.

Simon Baron-Cohen's trailblazing research gives us a preview of what psychological science will look like in the new century. In this conversational, understated volume, he attacks some of the most fundamental questions about how human beings mentally construct their commonly inhabited social world. He explores how a universal, evolved language of the eyes, which is mutually intelligible to all members of our species, can bring two separate minds into an aligned interpretation of their interaction. What we take for granted—the achievement of coordinated models of our mutual social interactions—he shows to be a triumph of automated modules and evolutionary cognitive engineering. Baron-Cohen lays out a series of elegant hypotheses outlining the design features and interrelationships of the modules responsible for these daily triumphs: an eye-direction detector, an intentionality detector, a shared-attention module, and so on. In showing how his proposals account for many dimensions of human social and mental life, he goes far beyond his own penetrating cognitive experiments and neuroscience research. In building his account, he weaves together a seamless tapestry from cognitive science, developmental psychology, primatology, philosophy, cognitive neuroscience, evolutionary biology, anthropology, neurology, behavioral ecology, and literature to create the first natural-science account of the mental machinery that implements the language of the eyes. It is exactly this focus on integrating—within a framework that simultaneously reconciles cognitive, evolutionary, and neural

levels of explanation—research from so many disciplines that we suspect will be the most salient characteristic of 21st-century psychology.

If we have eye-direction detectors and companion modules that define and speak the language of the eyes, what do they talk to? Normal humans everywhere not only "paint" their world with color, they also "paint" beliefs, intentions, feelings, hopes, desires, and pretenses onto agents in their social world. They do this despite the fact that no human has ever seen a thought, a belief, or an intention. A growing community of cognitive scientists has concluded that humans everywhere interpret the behavior of others in these mentalistic terms because we all come equipped with a "theory of mind" module (ToMM) that is compelled to interpret others this way, with mentalistic terms as its native language. We are "mindreaders" by nature, building interpretations of the mental events of others and feeling our constructions as sharply as the physical objects we touch. Humans evolved this ability because, as members of an intensively social, cooperative, and competitive species, our ancestors' lives depended on how well they could infer what was on one another's minds. Precisely because such an interpretive system does model the world in terms of unobservable entities (thoughts, intentions, beliefs, and desires), it needs to be coupled to confederate modules that can construct a bridge from the observable to the unobservable. Unobservable entities are invisible to association-learning mechanisms, but they are "visible," over the long run, to natural selection. As chance created alternative cognitive designs, this process "selected" those that implemented the best "betting" system. Over innumerable generations, the evolutionary process selected for modules interpenetrating our perceptual systems that could successful isolate, out of the welter of observable phenomena, exactly those outward and visible signs in behavior that reliably signaled inward and invisible mental states. These modules were built to expect, hook onto, and exploit patterns in the observable world that they already know how to recognize, and to use these targeted cues to fill in the blanks in the ToMM's pre-existing models of other people's mental states. By linking observable cues (such as direction of gaze) to representations of unobservable mental states (such as wants and beliefs), they create what one can think of as the "psychophysics" of the social world.

Yet even well-designed machinery can break down. When the machinery is fundamental to the operation of our minds, the results can be tragic—and deeply illuminating for the cognitive scientist. Breakdowns of specific modules result in subtractions from the impaired individual's model of and experience of the world. A color-blind individual loses one dimension of the visual world. A blind individual loses the entire visual world. But someone whose ToMM is impaired is blind to the existence of other minds, while still living in the same physical, spatial, visual, and many-hued world as unimpaired people do. For beings who evolved to live woven into the minds of mothers, fathers, friends, and companions, being blind to the existence of others' minds is a catastrophic loss. Simon Baron-Cohen and his colleagues were the first to propose that an individual with autism was one whose ToMM had been damaged. They persuasively explained how this hypothesis accounted for the bizarre constellation of symptoms autistics manifest. By considering what companion mechanisms the ToMM needed to function, Baron-Cohen and his colleagues could detect and experimentally track its computational links to what he has termed the eye-direction detector (EDD), the shared-attention mechanism (SAM), and the intentionality detector (ID). As the capstone of this research program, he and his colleagues used these new cognitive models to develop a method for detecting autism far earlier than anyone believed possible and successfully tested it on a base population of 16,000 children.

This sequence of discoveries is one of the key achievements of modern cognitive science. It deserves the careful attention of everyone studying social cognition and development, and because it encapsulates so many of the themes of psychology's metamorphosis it will become recognized as a milestone in the naturalization of the psychological sciences.

John Tooby
Leda Cosmides

Preface

This is a complicated book to write, as I have in mind readers from quite different backgrounds. First, I am writing for my colleagues in the biological and cognitive sciences, whom I hope will find the theory I advance here of sufficient interest that they will respond to the ideas and take them further than I have managed to. Second, I am writing for students in psychology (and related disciplines), for whom I want to make the topic exciting enough that they decide to stay in the field and make their own contributions. Finally, and not least, I am writing for the general reader who has no background in psychology but who wants to keep in touch with where science is going. Keeping in mind all three types of readers on each and every page requires a fair degree of acrobatics, I find. At times I have despaired that this juggling exercise cannot be done. I apologize if I occasionally lapse in this endeavor.