THE SEARCH FOR A THEORY OF COGNITION

Early Mechanisms and New Ideas

Edited by
Stefano Franchi and Francesco Bianchini

Rodopi

Amsterdam - New York, NY 2011

Cover Image: www.morguefile.com

Cover Design: Studio Pollmann

The paper on which this book is printed meets the requirements of "ISO 9706:1994, Information and documentation - Paper for documents - Requirements for permanence".

ISBN: 978-90-420-3427-3
E-Book ISBN: 978-94-012-0715-7
© Editions Rodopi B.V., Amsterdam - New York, NY 2011
Printed in the Netherlands

CONTENTS

ACKNOWLEDGMENTS

The early roots of the collection go back to a volume we edited for the *Discipline filosofiche* series in philosophy published by Quodlibet in 2007. Stefano Besoli, editor of *Discipline filosofiche*, supported our initial research and provided useful advice. The present volume extends the interpretation of cognitive science and Artificial Intelligence we started to articulate in the earlier work. Roberto Cordeschi provided helpful advice in the early phase of the project. Raffaela Serrani was the first translator of the chapter by Claudio Pogliano. Douglas Hofstadter provided precious help in the revision of the manuscript. His critical comments on the introduction were particularly helpful. Elena Pallottini provided significant help with proof-reading. Manuela Marchesini's editing skills were instrumental in turning the manuscript into a publishable work. Many thanks go also to the publishers of the *Journal of Philosophy* and to The University of Chicago Press, publisher of *Critical Inquiry*, for the permission to reprint the essays by Solomon Feferman and Peter Galison. Stefano Franchi would also like to thank the authors of the free software package LyX and the TeX collection of software, which made the typesetting of the volume possible and almost enjoyable.

FOREWORD

Douglas Hofstadter

The origins of humanity's ongoing attempt to realize intelligence, understanding, creativity, and consciousness in a technological artifact are many, including the disciplines of mathematical logic, cognitive and perceptual psychology, computer science and the theory of computation, the study of the details of brain cells and brain architecture, the philosophy of mind, cybernetics, information theory, linguistics, robotics, physics, evolutionary biology, and numerous other disciplines and schools of thought. To further complicate matters, over the decades, many different countries have contributed crucial ideas that reflect, in various ways, their particular cultures. Among the most important countries in this brew have been the United States, Great Britain, and Russia, but France, Italy, and Japan have also played significant roles.

In this book, two insightful and original young philosophers from Italy have brought together scholars from several countries and who collectively represent all of the disciplines mentioned above, in an attempt to shed light on the fascinatingly complex, international, multicultural, pluridisciplinary, and labyrinthine story of artificial intelligence, focusing on, but not limited to, the decades in the middle of the twentieth century. Thus there are articles about the attempts to use biological ideas to model the brain, about the role played by the science of self-organizing systems, about the raging philosophical debates caused by such results as Gödel's incompleteness theorems and Turing's results on the unsolvability of the halting problem, about the very nature of intelligence and how it might be detected in an agent, about the nature of simulations and models, and about what limits there are to computation posed by the nature of mathematics and by the laws of physics. Moreover, there are case studies describing how Italy, France, and the Soviet Union incorporated, reacted to, and idiosyncratically added to these swiftly moving streams of thought coming largely from abroad.

As we look back today at the origins of the movement now called "artificial intelligence," we are inevitably influenced by events that have happened in the more recent past, such as the rise of such paradigms as connectionism (sometimes referred to as "parallel distributed processing"), nonrepresentationalism, dynamic systems theory, robotics, cognitive linguistics, emotive computing, and others. And if anything, the debates over many classic philosophical questions have only heated up as artificial intelligence has made various kinds of progress. Such questions include the existence or nonexistence of free will, the potential role of randomness in the functioning of mind and brain, the controversial nature of supposed "qualia" (that is, the idea that certain physical systems are intrinsically

capable of being a locus of experience, whereas others are intrinsically incapable of it), the difference between "primary intentionality" and "secondary intentionality" (that is, the question of the "aboutness," or lack thereof, of symbols that are manipulated by machines, which has given rise to a putatively strict dichotomy between "syntax" and "semantics"), the question of other minds (another way of looking at solipsism), the mind–body problem (the question of whether a mind is a direct result of physical processes or whether something about a mind transcends physics), the seeming paradox of rule-bound creativity, the realization of emotions in inanimate (or at least inorganic) objects, and of course the eternal riddle of consciousness (which, it must be added, is not a riddle at all, let alone an eternal riddle, for some thinkers, who argue that consciousness simply does not exist, or else that it is at best a kind of necessary illusion of any sufficiently complex perceptual system, which is caused by the drastic simplification that perceptual processes inevitably involve).

All of these burning questions are front and center in this stimulating international volume concerning a crucial period in the history of one of the most philosophically ambitious quests ever undertaken by our species—namely, the quest by *homo sapiens* to build a mechanical rival to itself. That is to say, artificial intelligence is our collective quest to build a rival to our very essence, which is our ability to think—and indeed to build a rival that may possibly overtake us and render us obsolete. This book therefore constitutes an important contribution to the literature seeking to understand the nature of this quixotic voyage that we humans have collectively embarked on.

— July 2010, Paris.

ON THE HISTORICAL DYNAMICS OF COGNITIVE SCIENCE: A VIEW FROM THE PERIPHERY

Stefano Franchi and Francesco Bianchini

In 1998, speaking from the privileged perspective of the East Coast researcher, Daniel Dennett presented a "logical geography of computational approaches," a conceptual map of cognitive science that demonstrated how, by simply locating theoretical proposals with respect to their spatial and geographical distance from MIT orthodoxy, we could shed a great conceptual clarity on the methods and goals of cognitive science. Concerns that are not too dissimilar lie also behind the present introduction. We investigate the same broad set of disciplines and approaches, with the difference that our focus has shifted from space to time: instead of a "logical geography," we aim at portraying a "historical dynamics." Like Dennett, we think useful to look at the discipline in terms of the opposition between a theoretical "center" and its "periphery." We set out to concentrate our attention on the periodical shifts that from time to time have reassigned the functional roles of "center" and "periphery" to different research programs, often thrusting research efforts that only yesterday were close to its core, out on the limb of the theoretical landscape.

From the very beginning of AI, a constant fluctuation among different focal points—the methodological, the ontological, and the epistemological—has characterized its historical development. Over time, the core interests of many researchers shifted towards the edge; what was once in the spotlight, slipped into a dark zone; the center became periphery. Conversely, returns to the center from the the periphery or even from the outer regions became the norm. Indeed, Artificial Intelligence has displayed a peculiarly non-straightforward history.

Is there a reason, we may ask, behind these rapid and frequent shifts? We believe that even a cursory examination of the temporal evolutions of the "science of the mind" may teach us some lessons about the forces at work behind its historical vicissitudes, and provide indications about the dynamics of the discipline that we may find useful for present and future reference. Cognitive science appears to be endowed with peculiar epistemological features that not only caused those oscillations, but also led to an unusual relationship between the current and the previous status of scientific research. Let us begin with a brief narrative.

Between the 1940s and the early 1960s, a large number of scientific disciplines began working on the mechanization of a range of cognitive processes that often included not just adaptive behavioral processes but also emotions and

volitions. Their approaches varied. Some researchers in the field (including Alan Turing himself) tried to extend to mental operations the conception of abstract machine that Turing had just proposed. Others came up with computational architectures alternative to Turing's. Efforts to translate the concepts of the emerging school of cognitive psychology and well established behaviorist psychology into computer science went hand in hand with the software and hardware simulation of physiological processes that scientists often, but not always, targeted at the neural level. In spite of their theoretical incompatibilities, researchers in disciplines as different as psychology, anthropology, mathematics, formal logic, economics, physics, physiology, and others—all with their distinctive languages, methodologies, and histories—were all eager to discuss the strengths and weaknesses of their respective approaches. Researchers often met in person in the numerous conferences, seminars, and summer schools that dotted those years. The best known example of these public gatherings, though far from being the only instance of its kind, were of course the Macy's conferences that took place between 1946 and 1953.

Practical and theoretical reasons made these meetings possible. On the one hand, the great technical and political successes of the Manhattan project and, later, the panicked reaction to the Soviet Union's Sputnik launch, fueled a surge in government funding for scientific research—in the United States in particular, but also in the UK and in other European countries; in spite of the broad sample of disciplines present at the meetings, the limited number of participants facilitated personal discussions; last but not least, the close political and economic ties existing among the scientists' nations of origin worked in the same direction.

The theoretical reasons, on the other hand, were equally diverse. First, all participants to the multifarious enterprise shared one fundamental goal: namely, to provide a scientific theory of the mind, the entity "more ghostly than a ghost," —in the words of Charles Sherrington—whose understanding remained beyond the reach of modern science. Warren McCulloch used Sherrington's clause in a passage that is worth quoting in its entirety. It shows how the seekers of a theory of cognition saw themselves: as heirs to a long-standing philosophical tradition and, at the same time, as radical reformers engaged in a sweeping enterprise that a set of common tools had just made possible:

> Sir Charles Sherrington [the 1932 Nobel laureate physiologist who studied the role of synapses and neural integration in the nervous system] … near the end of a life spent on studying the ways of the brain, was forced to conclude that "in this world, Mind goes more ghostly than a ghost." The reason for his failure was simply that *his physics was not adequate to the problem* that he had undertaken. That has so regularly been the shortcoming of scientists who would have approached this problem, that even Clark Maxwell[*sic*], who wanted nothing more than to know the relation between

> thoughts and the molecular motions of the brain, cut short his query with the memorable phrase, "but does not the way to it lie through the very den of the metaphysician, strewn with the bones of former explorers and abhorred by every man of science?" *Let us peacefully answer the first half of his question "Yes," the second half "No," and then proceed serenely*. (McCulloch, 1989[1948], p. 143, emphasis added)

With the serene confidence that is proper to the man of science, researchers felt in the position to take care of the problem that had afflicted the metaphysicians of the past. Not only does McCulloch, as many other mid-twentieth century scientists, identify the problem of his nineteenth century predecessors (metaphysics) but he also indicates that only the guidance of the natural sciences will unmistakably take us to the goal that had proved so elusive. Many researchers working in the natural and the social sciences shared, if perhaps not McCulloch's confidence, certainly his approach. Notwithstanding the frequent unease that the social scientists felt towards the methodological self-confidence of their colleagues in the natural sciences, the common allegiance to the methods of natural science provided a shared vocabulary that facilitated intellectual exchange among disciplines.The situation changed dramatically toward the end of the 1960s, when one particular approach gained the theoretical and institutional upper hand and consigned to the periphery methods once at the core of the enterprise. The winning approach has been variously labeled AI (*Artificial Intelligence* (McCarthy et al., 1956)), *cognitivism* (Neisser, 1967), GOFAI (*Good Old Fashioned* AI (Haugeland, 1985)), and even the *Boolean Dream* (Hofstadter, 1985).

For our use here, we will adopt *cognitivism,* a term that perhaps embrace a broader concept, and that we take to refer to a variety of approaches that more often than not overlap with one another. Haugeland defines *cognitivism* as the "position that intelligent behavior can (only) be explained by appeal to internal 'cognitive processes,' that is, rational thought in general" (1981[1978], p. 243). In turn, we interpret these "cognitive processes," Haugeland says, as information processes that connect inputs to outputs by virtue of their ability to manipulate internal representations. "Cognitivism can be summed in a slogan," Haugeland concludes: "the mind is to be understood as an I[nformation] P[rocessing] S[ystem]." The performances of actual information processing systems must, he adds, grant empirical confirmation to the thesis: "[cognitive] simulations can function as concrete checks on whether particular proposed IPSS in fact have the abilities they are supposed to explain" (1981[1978], pp. 261, 278). Because of the tight link between a view of the mind as an information processing system and a reliance on cognitive simulations of those processes, cognitivism, formerly and formally a psychological theory, turns *ipso facto* into a hybrid theory that spans the divide between psychology and theory of computation. Daniel Dennett, who

later dubbed it "High Church Computationalism," reduced its theoretical content to three dogmas (1998, p. 217):

1. *Thinking is information processing*
2. *Information processing is computation (which is symbol manipulation)*
3. *The semantics of these symbols connects thinking to the external world*

Dennett's quasi-theological analysis has the merit of showing how, depending on which of the three dogmas we emphasize, the identification of *cognitivism* shifts from one domain to another. The traditional cognitivism that psychology (Neisser, 1967) develops, for instance, tends to put the emphasis on the first tenet; Fodor's Representational Theory of Mind (1987; sometimes called "representationism") emphasizes the third one; Herbert Simon and Allen Newell (defenders of the so-called "physical-symbol hypothesis," or "symbolicism" (1981[1976]) insist on the second one. In the USA, the philosophy of psychology has appropriated the term cognitivism; the philosophy of mind orbits toward "representationism" and the "RTM" (Representational Theory of Mind); the philosophy of Artificial Intelligence and of cognitive science prefers "computationalism" or "symbolicism." In Continental Europe (Varela, Thompson, and Rosch, 1991; Descombes, 2001; Dupuy, 2000), instead, scholars use *cognitivism* in a sense that is broader than their American counterparts', and that covers all of the approaches we just listed. Precisely because of this more extensive range of the term, of which we will say more shortly, we have opted—like the Europeans—for "cognitivism" and, alternatively, "the cognitivist paradigm."

Roughly at the same time, in the 1960s, scientists introduced another label that, like "cognitivism," was a convenient shorthand for the collection of disciplines working toward a scientifically rigorous and empirically testable theory of the mind: *cognitive science*. For the new research institute they founded at Harvard in 1960, George Miller and Jerome Bruner resorted to "cognitive studies." "Studies" slowly morphed into "science." By 1975, when Daniel Bobrow and Allan Collins used it in the title of their successful anthology (*Representation and Understanding: Studies in Cognitive Science*), the designation was popular enough that the editors did not feel compelled to explain it. Since George Miller was one of the main defenders of cognitivism, we may be tempted to simply define "cognitive science" as the kind of psychology that researchers of cognitivist persuasions practice. That would be a mistake. Even though for much of their history the two labels—*cognitive science* and *cognitivism*—have referred to the same research programs, to the same people, and to the same publishing ventures, we think that distinguishing them is crucial. *Cognitive science* (or, sometimes, "the cognitive sciences") refers to the discipline or disciplines that study cognition. Cognition comes from the Latin noun *cognitio* (from the verb *cognosco,* "to become acquainted with"= "to know"). As the etymology indicates, we should

understand "knowledge" and "thought" as the concrete, permanent results that lie at the end of a process—hence, an item of knowledge, a *cogitatum*, or a thought. We therefore could say that "cognitive science" is the discipline that seeks a theory of our thinking as knowledge-producing processes, and that keeps clear of other psychological processes like emotions, volitions, moods, drives, and so forth. But history, at least in this case, contradicts etymology. George Miller and Jerome Bruner introduced "cognitive science" to designate the discipline(s) that studied not just the smaller subset of rational processes*, but all* mental phenomena. Miller remarked that they had chosen "cognition" over the redundant "mental psychology" precisely in order to distinguish their approach from *behavioral* (non-mental) psychology:

> In reaching back for the word "cognition," I don't think anyone was intentionally excluding "volition" or "conation" or "emotion." … [I]n using the word cognition, we were setting off from behaviorism. We wanted something that was *mental*—but "mental psychology" seemed terribly redundant. "Commonsense psychology" would have suggested some sort of anthropological investigation, and "folk psychology" would have suggested Wundt's social psychology. What word do you use to label this set of views? We chose "cognition."[1]

In short, researchers introduced "cognitive science" for the study of *all* mental processes, to signal a broad perspective that included classic psychological investigations and methods and procedures coming from other fields such as computer science, philosophy, linguistics, and the neurosciences. The aim was to provide not just a theory of the input/output correlations proper to mental processes, as behaviorist psychology was wont to do, but a theory of the mental processes themselves. In the beginnings, *cognitive science* nurtured an explicit *negative* theoretical commitment against behaviorism but, unlike *cognitivism* vis-à-vis the natural sciences, no equally explicit *positive* commitment to any specific theoretical framework. As the disciplines matured, the whole situation changed. The *cognitivism* we have defined above, with the theoretical position it encompassed, became such a dominant force within *cognitive science*, that in the short span of ten years the two formally distinct terms fused together. Against the grain of its historical unfolding, *cognitive science* had turned into the investigation of the mind seen through *cognitivist* glasses. This view is still popular. A recent textbook, for instance, offers the following definition:

> Here is the central hypothesis of cognitive science: Thinking can best be understood in terms of representational structures in the mind and computational procedures that operate on those structures. Although there is much disagreement about the nature of the representations and computations that constitute thinking, the central hypothesis is general enough to encompass

> the current range of thinking in cognitive science, including connectionist theories. (Thagard, 2005, p. 10)

We beg to disagree. Paul Thagard's definition may be historically correct for the period between the early 1960s and the late 1980s, but not earlier on. At any rate, even in the 1960s and 1970s, cognitivism's true golden age, notable differences were already present. For instance, Douglas Hofstadter, a well-known and respected researcher whose professional qualification—Distinguished Professor of Cognitive Science at Indiana University—designate him as a cognitive scientist, strongly disagrees with Thagard's identification of cognition with computation:

> In some domains, even in some relatively complex and technical ones, people have come up with programs that can [imitate chains of serial actions that come from verbal protocols of various experimental subjects]. But what about the simpler, noncognitive acts that in reality are the substrate for those cognitive acts? Whose program carries those out? At present, no one's. Why is this? It is because AI people have in general tended to cling to a notion that, in some sense, thoughts obey formal rules at the thought level, just as George Boole believed that "the laws of thought" amounted to formal rules for manipulating propositions. I believe that this Boolean dream is at the root of the slogan "*Cognition as computation*"—and I believe it will turn to be revealed for what it is: *an elegant chimera*. (Hofstadter, 1985, p. 654, emphasis added)

We are convinced that conflating *cognitive science* with *cognitivism*, or rather flattening the first term onto the second, is too narrow a move. It fails to provide an interpretive framework broad enough to include a remarkable wealth of theoretical positions that predated the onset of *cognitivism*, and that perdured during its hegemony—albeit in a more peripheral role—-until, in more recent years, they staged an impressive comeback. The central hypothesis behind the work that went into the preparation of this volume is that Thagard's identification of *cognitive science* with *cognitivism* is a historically contingent fact. It happened, but it did not need to. Convinced as we are that a historical and conceptual analysis of these issues may prove theoretically relevant, we commit this book to an examination of how it happened and at what theoretical costs.

Let us go back to the historical narrative. By the mid 1960s, the ascendancy of cognitivism within cognitive science led to a disciplinary and institutional consolidation around that unique approach and to the coincidental marginalization of alternative strategies. Since 1967, the increasingly dominant cognitivist paradigm served as the main inspiration behind new journals, research institutes, Ph.D. programs, and publishers' series, and the skills of the researchers tended to be more narrowly focused on the engineering disciplines.[2] The relatively few exceptions to this pattern were mathematicians, logicians, and those amongst the psychologists who, unlike engineers, were eager to translate their cognitivist

theories into computer simulations. Whereas only thirty years earlier an anthropologist (Gregory Bateson), a psychologist (Karl Lashley), a mathematician (Norbert Wiener), a neurophysiologist (Warren McCulloch), and a psychiatrist (W. Ross Ashby) could and had interacted face to face, the Artificial Intelligence of the 1970s strengthened its identity in international congresses (first and foremost the AAAI and the IJCAI series) where the sheer number of contributions was in inverse ratio to the breadth of their coverage.[3]

We could interpret this evolution as an unfolding that is typical of normal scientific progress.[4] When a new field of inquiry emerges, philosophers and, more generally, humanists belonging to different walks of intellectual life develop a multitude of varied and often mutually incompatible theoretical approaches. Then they test those approaches empirically. Little by little, they discard concepts and methods until the discipline—now past its initial "heroic" phase—enters a stage of normalization at the theoretical and institutional levels. An almost inevitable consequence of this second stage is the narrowing of interests and the standardization of the methods and techniques. Several examples corroborate this model. We might recall, for instance, the development of the science of linguistics: after its beginnings in the age of the Enlightenment, the Romantic philosophers and humanists laid the basis for the study of language as an autonomous field, even though their theories were still at the crossroads of many different disciplines, from philosophy and psychology, to philology and literary criticism. Only with the Neogrammarian revolution of the late 19th century, though, was linguistics to emerge as an autonomous science that supposedly obeyed laws as inflexible as those of the natural sciences. Later on, the Structuralist and Chomskian frameworks would refine and extend such an approach. A similar path we could also detect in the development of natural sciences such as biology and chemistry, two disciplines that underwent a long period of theoretical and methodological instability before reaching a normalization on the basis, respectively, of a biochemical reinterpretation of evolutionary theory and of a physical understanding of chemical properties such as valence.

This model construes scientific development as a linear, self-correcting, and quantitatively increasing process. The development of a science begins with the invention of novel yet necessarily imprecise concepts and methods, and it reaches maturity when a widely accepted methodological and conceptual framework—which the leading textbooks of the discipline typically canonize—comes to constitute its hard-core. While philosophical work can help scientific inquiry by clarifying difficult concepts and by removing theoretical blocks, historical analysis can never contribute to actual scientific progress. History studies past theories and concepts that were definitively refuted or, at best, were included as limit cases into newer and more powerful theories (the canonical example being the subsumption of Newtonian physics into Einsteinian relativity). This model of scientific progress downplays the scientific interest of any research approach

that is historically and conceptually peripheral to the mainstream. Even though an accurate historical reconstruction might revive the past, its results would be irrelevant to the actual path of present-day research.[5]

On the contrary, we find that the scientific study of the mind did not follow this pattern. We began assembling this collection with the assumption that a historical study of the research pathways that fell off the mainstream is indeed a necessary component of the scientific study of cognitive and sub-cognitive behavior. The motive behind our assumption is not our lack of confidence in the general validity of the standard model of scientific progress we described above. Instead, we are convinced that some precise and distinctive features proper to the scientific study of the mind force us to reject an assumption of historical and theoretical linear progress. As Roberto Cordeschi has shown (2002), the scientific field of "artificial studies" is non-traditional and non-cumulative. Why is that the case? Which peculiar features of a theory of cognition make the study of its theoretical and historical periphery relevant to contemporary research? Briefly stated, as far as we are concerned, two: the intrinsic self-reflexivity that pertains the study of cognition and its essential dependence on traditional philosophy. More precisely, these two aspects, which in fact summarize the dominant double feature of any philosophical reflection, boil down, like in the proverbial two-side coin, to a single one.

Cognitive science was the discipline that would provide a rigorous and complete theory of the most distinctive aspects of human nature grounded in the natural sciences. However, since its main targets were language and thought (as well as emotional and adaptive behaviors), the discipline would never be able to leave behind its "heroic" phase and acquire full autonomy from traditional philosophical inquiries, because its central concerns were exactly the same as philosophy's. The new science of the mind could only aim to replace the old speculations with new and scientifically more adequate theories. Content-wise, however, the new science and the old philosophy would be equivalent, thereby preventing scientific autonomy according to the traditional model.

Let us expand this argument, by taking care of a possible objection first. We could say that a similar attitude has often been present in most scientific disciplines. After all, the history of science has witnessed many forms of reductionism like biologism, psychologism, sociologism, physicalism, just to name a few. In each case, researchers made the effort to explain a variety of human phenomena—at the individual and the social levels—as the surface manifestation of an underlying reality whose complete theory only the discipline operating the scientific reduction, be it biology, psychology, sociology, and so on, could provide.[6] Some of the research projects working toward the mechanization of the mind had reductionist leanings. We could mention, for instance, some interpretations of the early cybernetics in the style of Wiener (Richter, 1976), as well as later versions of Ashby's theory of the brain (J. G. Taylor, 1962). In the

first case, we were confronted with a rather strict form of physicalist reduction, while in the second case biology came to the forefront. Nonetheless, the most common attitude in the early years—an attitude that AI would adopt—was deeply anti-reductionist in the following sense: researchers did not reduce the multiplicity of phenomena that they could directly or indirectly classify as distinctly human to any single aspect—be it psychological, physical, or biological. Instead, they interpreted them as different manifestations of an essentially human nature whose complete theory the future discipline would eventually provide. How broad is this theoretical goal? Perhaps more than we usually acknowledge. We will try to characterize the scope of cognitive science with the help of a terminology that Kant made popular first, when discussing the architectonics and method of philosophical research. Our reference to Kant is not gratuitous, we think, since at least one of the key figures in the history of cognitive science, Herbert Simon, has recognized the close affinities between his project and Kant's philosophical system.[7]

Kant divides the scope of philosophical reflection into three main categories. In order of increasing generality, they are: *special metaphysics*, *general metaphysics*, and *propaedeutic metaphysics* or *critique*. Special metaphysics is the analysis and clarification of the specific objects and concepts that a particular discipline uses. In physics, for instance, it would clarify the concepts of force, motion, mass, and so on. General metaphysics is the analysis and clarification of objects in general (of what there *is*). We nowadays call it "ontology": an analysis of the general conditions under which human beings can think of objects and their properties and make successful references to them in their communicative acts. Finally, the *critique* is the preparatory and most general philosophical inquiry into the capabilities of human reason, independently of any particular object to which humans may apply them (Kant, 1998, pp. 696ff.=A841/B869ff.). Nowadays, when biologists discuss the ontological priority of the gene over the organism (Dawkins, 1976; Eldredge, 1995), or when physicists argues over strings and quarks as the ultimate constituents of physical reality (Greene, 2000; Penrose, 2005), we witness instances of scientists that engage in special metaphysics. General metaphysics and critique, on the other hand, tend to be the specialized province of philosophers. Labels might have changed since Kant's time, but the theoretical tasks are not so different. Even though general metaphysics has become ontology, it struggles with some of the same problems, including the realism/anti-realism controversy, the debate over the nature of causality, the definition of objects and parts, and so on. Today hardly anyone, except hard-core Kantians, still calls the general inquiry into the capacity of human reason a *critique*. Nonetheless, it has become one of the prevalent modes of philosophical discussion, especially—but not only—in the so-called Analytic tradition that flourishes in the USA and other English-speaking countries. Much contemporary work in the philosophy of mind, in epistemology, in the philosophy of language and in the philosophy of logic is a

reflection—very similar to Kant's own—upon the structural features and overall limits of rationality and its application to the real world.

Here is also where the peculiar character of cognitive science lies, however: the scope of its investigation was (and still is) sufficiently broad to cover, in addition to empirical research, not only its special metaphysics, but also general metaphysics and critique. Examples abound. We cited above McCulloch's declaration about a science of the mind that would finally solve the puzzles of metaphysics. Likewise, in the works of Herbert Simon and Allen Newell (1990) we may find similarly all-comprehensive attitudes toward the methods and goal of cognitive science.[8] A similar perspective endures in more recent works in sub-fields as diverse as cognitive neuroscience (Bennett and Hacker, 2003), evolutionary robotics (Di Paolo, Rohde, and Jaegher, 2011), and the biology of cognition (Weber and Varela, 2002). Patricia Kitcher's critique of cognitive science (1992) springs from the same claim: like Freud's, Kitcher argues, we may interpret cognitive science as a broad and interdisciplinary research program whose efforts aspire to reach an overall theory of how the mind works and how it produces theories. Phrased in traditional philosophical terms, we claim that cognitive science is intrinsically self-reflective because it produces a theory of its conditions of possibility (a theory of itself, in other words). A geological theory is not a rock, nor is a physical theory a billiard ball: their results and predictions do not apply to themselves. A physical theory may provide an account of how a billiard ball came into being, or at least an account of how it achieved its current position and momentum. But that theory does not have to provide an account of how the physical theory *itself* came into being. On the contrary, a theory of cognition *is* cognition, and as such it must include an account of how it was produced. This is why we cannot draw a clean-cut separation between the theory's objects on the one hand, and the theory's methods and concepts on the other: they are cut from the same cloth. And this is why we cannot apply the simple—and arguably simplistic—model of inexorable, linear, and cumulative scientific progress to cognitive science. In principle, every new theory of cognition necessarily rethinks, to a greater or lesser degree, what a theory is, what its proper objects and its proper methods are, an so on. With the exception of a few momentous paradigm shifts, such as the transformation of the teleologically-oriented closed world of Aristotelian physics into the infinite universe of Galilean dynamics that Alexandre Koyré (1957) documented, a new scientific theory does not *usually* change the definition of its objects. The new theoretical apparatus is external to its domain, for it provides new explanations for the same phenomena.[9] Cognitive science does not follow the pattern of the other scientific disciplines.

Kant's terminology lets us grasp a peculiarity in the historical evolution of cognitive science that has important repercussions on any historical analysis of the discipline. Because AI must produce—alongside its specific, testable theories concerning particular aspects of human activities—the theory of those theories

(namely, its special metaphysics), and the theory about the general possibility of a theory (the general metaphysics as well as the critique of "Reason"), AI is intrinsically unable to become an autonomous discipline by following the same path of physics or biology. In Kuhnian terms, we might even say that cognitive science is almost always operating in a paradigm-shifting mode: its historical development is less linear, less cumulative, and not at all monotonic. This is why history becomes an important tool for actual research, and not only for an antiquarian reconstruction of the past. If the development is not linear and not cumulative, then we may find actual scientific progress in the reactualization and rediscovery of past lines of inquiry.[10]

The simultaneous presence of intertwined epistemological levels reveals several ontological questions about the objects that the discipline studies. In order to sanction the explicative power of the scientific theories that the discipline produces, we must properly address those ontological questions. Two different and opposite needs have always inhabited research in AI—and rightly so, for the structural reasons we just sketched. Narrowly focused, hyper-specialized testable projects always, unsurprisingly, go hand in hand with vast, often too vast, generalizations. In the golden years of classic AI—roughly, between the mid-1960s and the late 1970s—these two components entered into a virtuous circle. The positive, although limited results that researchers obtained at the particular level (early semantic systems, expert systems, heuristic programs, and so forth) provided an empirical verification of the general symbolic approach, which, in turn, refined and renewed the general concepts ("heuristic navigation," "semantic representation," etcetera) to be tested in the field. On the one hand we could interpret the circle as vicious, the symptomatic behavior of a discipline that fails to become a science (Matteuzzi, 2005), or on the other as the virtuous mixture of a series of related disciplines that would produce a meta-science of the mind (Gardner, 1985).

Everything changed, however, when the cognitivist paradigm entered a prolonged period of crisis. The advent, in the 1980s, of the so-called "winter of AI" (Winston, 1987; Crevier, 1993) broke the virtuous circle between narrow empirical projects and wide theoretical inquiries. The only element left from the original unity was interdisciplinarity, although often the naturalistic reductionism of neuroscience was responsible for it. The history of AI's crisis has already been told, sometimes in considerable detail (see, for instance, Boden, 2006). We would like to emphasize some of its epistemological consequences. Clearly, breaking the tight link between special inquiries and general investigations causes a problem that we can solve by developing the discipline in two different and incompatible directions. First, we can sever the link once and for all by focusing attention on narrow pieces of research that researchers will no longer see as contributions to a general science of the mind. They will reinterpret their work as solutions to specific technical problems that arise within computer science or

within software engineering. Second, replacing the discarded cognitivism with a different approach could reestablish the link (in Kant's language) between general and special metaphysics that used to guarantee the existence of a virtuous circle in the first place. In the two decades that separate us from the "winter of AI," both options have been repeatedly tried. Stuart Russell and Peter Norvig's successful textbook (1995) canonized the first alternative. It reinterprets Artificial Intelligence as a technical discipline that is firmly located within the field of computer science and that is oriented solely toward the solution of specific technical problems bound to produce usable artifacts—as does every other branch of engineering. The second option, on the contrary, has generated a continuous search for an adequate theoretical framework that could fill the epistemological role that cognitivism once played. We might mention, roughly in chronological order, some candidates for the position: the connectionism that Rumelhart and McClelland relaunched in the early 1980s; the focus on sub-cognitive abilities that Hofstadter championed; the *nouvelle AI* that Brooks's behavior-based roboticsinspired; the dynamicist approach to cognition, evolutionary robotics, all the way up to artificial life and research paradigms that theoretical neuroscience brought forth.[11] As Margaret Boden shows in her historical monograph, we have no reason to believe that the search for cognitivism's heir will be over any time soon. Her comprehensive review ends with a chapter that lists all the candidates that, by possibly producing a theoretically unified field comparable with the one cognitive science had once enjoyed, might revitalize the study of cognitive and sub-cognitive phenomena.

The present work aims to contribute to the historical and theoretical reconfiguration that cognitive science is currently going through. We have chosen the concept of "periphery" as a privileged standpoint from which to look at the current situation. Now that the theoretical center of gravity of Artificial Intelligence has for all practical purposes, dissolved, historians and philosophers of cognitive science must direct their interest toward the formerly marginal approaches that were at an (often considerable) theoretical, geographical, and cultural distance from the center. We say "must" because we are convinced that a careful study of the periphery is crucial for several different purposes. Anyone wishing, like Margaret Boden, to reestablish a new unifying core for the science of the mind should designate the periphery as a privileged object of study. A historical and theoretical study of it is a prerequisite for all discussions about the possibility and necessity of a new center that might replace the current fragmentation. We view such an analysis as preliminary to a fruitful practical and theoretical approach. Finally, a dispassionate look at the periphery is crucial for all philosophical discussions about the possibility of a project that, like AI, turned out to be an updated and naturalized version of Kant's system.

The present volume does not exhaust all possible peripheral approaches, but it offers a collection of studies that examine four different ways to gain a

perspective on the theoretical center of AI research. Part I deals with the approach that we might consider as the most important periphery of classic AI research. So much so, in fact, that it constituted a second center or even an altogether autonomous suburb. We refer to cybernetics. Firmly established as the driving force behind the new science of the mind in the 1940s and 1950s, cybernetics became the main rival of classic AI research until after a well-known institutional fight it was ultimately marginalized.[12] Stefano Franchi, Peter Galison, and Peter Asaro are the authors of the three essays in part I.

Franchi focuses his attention on a cybernetic artifact, Ashby's Homeostat. He closely follows its development from the biological theories of the early 20th century to Ashby's full-blown discussion in *Design for a Brain* (1960). Then he examines its disappearance in the following decades, until he bears witness to the unexpected forays of the Homeostat into some of the most recent research projects in cognitive science (evolutionary robotics in particular). The Homeostat is reborn. This rebirth promises a radical reorientation of the field that would reintroduce the philosophical subject—endowed with needs, freedom, and self-generated goals—within the scope of cognitive science. Franchi concludes by suggesting that such a perspective, albeit thrilling, may be also, at least for now, perhaps too hasty—unless psychoanalytic conceptions of the self will come to supplement homeostasis-rich theories. In his seminal essay on the historical evolution of cybernetics' theoretical framework, *Galison* traces the development of Norbert Wiener's thought from its formation during the war years to the end of Wiener's life. Galison shows the crucial role that the metaphor of the "enemy," which allowed cybernetics to develop ideas and artifacts focused on reaching a target that might be (non-linearly) moving across time and space, played. *Asaro* discusses the concept of "scientific model" in cybernetics, in cognitive science, and in the neurosciences, and shows how the different referents—the mind, the brain, the physical world—force the concept to play radically different theoretical roles.

Part II focuses on the geographical periphery of AI research. It examines the translations into countries that had different cultures and traditions—Italy, France, and the Soviet Union—of the theories and techniques developed in the United States, AI's home ground. Claudio Pogliano, Patrice Maniglier, and Slava Gerovitch are the authors of the essays.

Pogliano discusses the diffusion of cybernetics and proto-AI ideas in Italy in the 1950s and 1960s. Thanks mostly to the efforts of US-educated intellectuals and businesspeople, Italian culture was quick to absorb the new ideas and the computing devices that were coming from the United States. Yet, while welcoming the new theories, Italy's solid humanistic and metaphysical tradition brought into question some of the most radical epistemological and philosophical claims that those ideas accompanied. The result was an innovative integration between classic humanistic values and new technological artifacts that aimed at providing the foundations for a new "mechanical civilization." *Maniglier* examines the

development of the main tenets of cognitive science from the point of view of French Structuralism, a theoretical framework that was being developed at the same time as Artificial Intelligence. He claims that the theories that were part of Artificial Intelligence's background, such as von Neumann and Morgenstern's game theory and Shannon's theory of information, partly inspired also Structuralism. Maniglier shows that AI and Structuralism offer complementary solutions that we could integrate with a positive outcome, if we were to combine them: the static view of cognitive processes that classic AI developed may fruitfully complement the dynamic view of evolution of cultures and semiotic systems that classic Structuralism articulated. *Gerovitch* examines the diffusion of cybernetic and Artificial Intelligence ideas in the Soviet Union. In spite of the cultural and political rivalry that separated the two countries, researchers in the USSR were eager to adopt some of the new insights that were coming from the USA. Their different cultural and political background, however, led scientists of the Soviet era to interpretations that were often new and different from their sources'. In particular, Gerovitch shows how Soviet researchers tended to sideline the category of "rational choice" that was crucial to North American cybernetics and to AI (especially in heuristic programming), in favor of the category of (not necessarily rational) "creativity."

Part III focuses on AI's periphery understood in the cultural and historical meaning of the term. It includes essays that locate some of the central concepts of AI, such as representation and computability, within a broader background, be it philosophical (Descartes, Aristotle, Leibniz) or technical (programming theory and practice). Francesco Bianchini, Maurizio Matteuzzi, and Christopher Kelty are the contributors to this part.

Bianchini examines the debate around the experimental validation of computer simulations of intelligent behavior. Under which conditions can we consider a machine intelligent? Since Turing's original test, researchers have developed several criteria that have often generated lively debates about their adequacy. Bianchini offers a new perspective. He recommends that we locate the evolution of Turing's ideas in the historical development of the epistemology of AI. By showing that we might interpret contemporary criteria such as "behavior," "formalism," and "realism" as technical and pragmatical reformulations of philosophical conceptions of the mind that philosophers such as especially Descartes and Leibniz developed during the modern era, Bianchini offers some reflections toward a hopefully fruitful new awareness and reemployment of those concepts. *Matteuzzi* discusses the concept of Turing-computability from the point of view of mathematical constructivism and with the help of Leibniz's conception of computation. He suggests that even within the domain of constructivist approaches to mathematics (as Bernard Bolzano and Karl Weierstrass defined them), the use of random choice produces computations that bypass the limitations of Turing-computability. Ever since its inception, AI has identified computation with Turing's formalization of

it, while the notion that mainstream computer science uses, Matteuzzi suggests, has been more flexible and more aware of its limitations. The close relationship between Turing-computability and the simulations of intelligent behavior that Artificial Intelligence attempts raises the possibility that a broader conception of computability may substantially renew the theoretical framework we use to model cognitive behavior. *Kelty* reconstructs the history of regular expressions and shows that this powerful programming technique has played a notable practical role in the development of AI software, in particular in the sub-fields of knowledge representation and problem solving. Kelty shows that we can consider the transition from formal theory to algorithm construction as a magnifying glass revealing the process of identification of thought and logic that was characteristic of the years immediately preceding the birth of AI. Such an identification still exerts a strong influence even on the most recent theories of cognition that Artificial Intelligence has developed.

Part IV focuses on the limitation of classic forms of computations and possible alternatives to them. Solomon Feferman, Rossella Lupacchini, and Giorgio Sandri are the authors of the essays.

Feferman discusses the issue of the intrinsic limitations of the mind on the basis of the debate around the publication of Ernst Nagel and James Newman's *Gödel's Proof*—the first popular book about Gödel's work on incompleteness. A close examination of the correspondence between Gödel and Nagel and Newman allows Feferman to show how the debate on the limitation of human thought originated and developed over the decades and how it influenced the philosophy of mind, AI, and cognitive science. Feferman suggests that a renewed appreciation of that debate can lead us to a new understanding of the formal/informal dichotomy of mathematical thought and of the closely associated notion of computability. *Lupacchini* examines a more peripheral notion of computation. She discusses the possibility that we might use concepts derived from quantum theory to broaden the notion of effective computability. As John von Neumann himself realized in the 1920s, the intrinsic non-determinism of quantum mechanics makes it quite difficult for researchers to provide a satisfactory mathematical axiomatization. Since it appears to require non-deterministic choices, which are quite difficult to formalize within classic computability theory, the simulation of intelligent behavior faces a similar difficulty. We could resolve the difficulty, Lupacchini suggests, through the development of quantum computers that would rely upon the lack of determinism that is characteristic of quantum theory. *Sandri* examines the relationship between, on the one hand, the classic theory of computability that Alan Turing formulated in the years immediately preceding AI's birth and, on the other, the objections that Gödel raised a few years later. Sandri argues that we can trace the difference between Turing's and Gödel's notion of computability back to a different conception of mental states. He stresses how cybernetics' emphasis

on information and real-time processing had a profound effect on the notion of computability.

NOTES

1. Baars, 1986, p. 210. Later in the interview, Miller gives an even more salient example of the breadth of cognitive psychology as he intends it, by declaring: "I read Freud's *The Interpretation of Dreams* and chapter 7 of that book [the theoretical chapter about the unconscious, repression, etc.] is to me one of *the* great essays in cognitive psychology" (*ibid.*, p. 214, Miller's emphasis).

2. For a snapshot of the publishing industry that sprouted around the newly minted "Cognitive Science," from the early 1970s (when the label entered common use) onward, see Boden, 2006, pp. 354–365. Boden does not mention that several among the researchers themselves pioneered some publishing ventures. For instance, Nils Nilsson had a leading role in the establishment of Morgan Kauffmann, one of the main publishers of technical literature in the field (including the proceedings of its major conferences, AAAI and IJCAI). 1967 is also the year that saw the publication of Marvin Minsky and Seymour Papert's *Perceptrons*, the final and (provisionally) winning shot in the debate with cybernetics.

3. This claim is hard to prove in strict numeric terms, but a quick look at the table of contents of the early *vs.* the later conferences may offer a useful indication. The 1952 Macy's conference had ten contributions coming from eight different disciplines spanning a spectrum that goes from logic to anthropology, and from computer science to anatomy (von Foerster, Mead, and Teuber, 1953). Thirty years later, the *8th International Joint Conference on Artificial Intelligence* (Bundy, 1983) had over 240 contributions divided among technical sub-fields such as "Natural Language [Processing]" (30), "Expert Systems" (24), "Knowledge Representation" (20), "Logic Programming" (10), "Robotics" (10), "Vision" (15), and so on. "Cognitive Modeling" (with 10 contributions, less than 5% of the total) is the only subdivision to reach out of the strict confines of computer science.

4. This is how Patrick Winston (1987) reads it.

5. For a detailed analysis of the reconstructive vs. interpretive views of history our discussion relies upon, see Gadamer, 1996, pp. 164–169.

6. A small sample of such attempts would include at least Comte's and Durkheim's sociological theories of science, along with their more recent incarnations in the so-called "strong program" of the sociology of science (Bloor, 1976; Latour and Woolgar, 1986; Latour, 1987); the biological reductions that evolutionary psychology attempted (Tooby and Cosmides, 1992); the psychologism in logic and mathematics that the German school at the turn of the 19th century carried out (see Kusch, 1995); Quine and his followers, who

continued and revived psychologism's program by naturalizing epistemology (Kornblith, 1985; Kornblith, 2002); Jean Piaget's genetic epistemology (1950; 1971); and recent variations on the same theme provided by the emergence of forms of reduction that are rooted in the neurosciences—such as, for instance, the "Neurophilosophy" that Patricia Churchland, among others, practices (1986; 2002).

7. Even though Carnap's interpretation (Carnap was Simon's teacher at The University of Chicago) of Kantism was "epistemologized"; see Proust, 1987; Franchi, 2006. Kant adopted his terminology from Christian Wolff and from his follower, Baumgarten, whose textbook on metaphysics Kant routinely used when lecturing.

8. In a personal communication to one of the Editors, Simon once remarked: "Franchi is right in considering Artificial Intelligence as an effort to produce a metaphysics [even though I prefer the term 'epistemology'] with non-philosophical means. The reason why AI can do so, and has already started to do so, is that we must rephrase epistemology's questions as empirical questions (how mind works and interacts with its environment)." See Franchi, 2007, for a detailed discussion.

9. See Matteuzzi, 1981, for a sophisticated discussion of the complex relationship between theories and their objects.

10. Gould, 2002 offers a vivid example of a practicing scientist who integrates historical research with scientific inquiry.

11. See Rumelhart and McClelland, 1986a; Hofstadter, 1985; Hofstadter and Fluid Analogies Research Group, 1995; Brooks, 1999; Beer, 1990; van Gelder and Port, 1995; Harvey et al., 2005; Nolfi and Floreano, 2000.

12. See Dupuy, 1994; Dupuy, 2000; Heims, 1991; Franchi and Güzeldere, 2005a; Minsky and Papert, 1967.