Introduction

Keith Frankish and William M. Ramsey

Overview

Very generally, artificial intelligence (AI) is a cross-disciplinary approach to understanding, modeling, and replicating intelligence and cognitive processes by invoking various computational, mathematical, logical, mechanical, and even biological principles and devices. On the one hand, it is often abstract and theoretical as investigators try to develop theories that will enrich our understanding of natural cognition or help define the limits of computability or proof theory. On the other hand, it is often purely pragmatic as other investigators focus on the engineering of smart machines and applications. Historically, its practitioners have come from such disciplines as logic, mathematics, engineering, philosophy, psychology, linguistics, and, of course, computer science. It forms a critical branch of cognitive science since it is often devoted to developing models that explain various dimensions of human and animal cognition. Indeed, since its inception in the mid twentieth century, AI has been one of the most fruitful new areas of research into the nature of human mentality. Today, it is impossible to be a serious cognitive scientist or philosopher of mind without at least some familiarity with major developments in AI. At the same time, anyone who uses modern technology is probably enjoying features that, in one way or another, had their origin in AI research, and AI technology will undoubtedly play an increasingly large role in our lives in coming decades.

This volume of original essays aims to describe the state of the art in the field of AI and to highlight important theoretical and philosophical applications of current research. The book's focus is on theory rather than technical and applied issues, and the chapters should be useful not only to AI researchers and other cognitive scientists, but also to philosophers and people in the humanities. Each chapter is a specially commissioned survey article from a leading writer in the area – either a philosopher of AI or an AI researcher with strong theoretical interests. There is coverage of the foundations of the discipline, cognitive architectures, the various facets of AI research, and extensions of AI research, such as robotics and artificial life (see the chapter summary below for details about each contribution). The approach is thematic rather than historical, and although the chapters are primarily survey pieces, critical assessment is also included where appropriate. We have worked hard to make the material accessible to non-specialists, and readers are not expected to have any significant background in the subject areas covered.

The volume possesses a number of distinctive features. First, it provides concise and up-to-date coverage of a diverse and rapidly expanding field, written by leading researchers with important perspectives. The contributors include both senior figures who have helped to form the modern discipline and younger researchers doing work that will help shape its future. Second, it presents scientific work in a form that is accessible to a humanities audience and focuses on broad theoretical issues and applications rather than experimental work and technical details. Contributors present logical and mathematical principles in general terms, and the use of symbolic notation has been kept to a minimum. Third, the book includes coverage of important topics that have been relatively neglected by past overviews, including the ethics of artificial intelligence, artificial life, and machine consciousness. Fourth, the discussion is pitched at an intermediate level, suitable for both advanced students and scholars new to the area, and the book includes supporting material, such as a glossary and chapter-specific "Further reading" sections that enhance its value as a teaching text. A companion volume, *The Cambridge Handbook to Cognitive Science* (2012), serves to complement this handbook in both scope and aims.

Artificial intelligence and cross-disciplinarity

As its title indicates, this volume is intended as a guide to AI itself, rather than to the *philosophy* of AI. That is, the primary focus is on first-order research in AI, and on theoretical issues raised directly by that research, rather than on meta-level philosophical questions about the *science* of AI. Yet, as readers will note, several of the chapters are written by people who are usually characterized as philosophers of AI or of cognitive science, and the volume's co-editors both have their homes in philosophy departments and have done no significant first-order research in AI. Why, one might ask, do philosophers have such a large role in a volume about one of the sciences of the mind?

There are a number of reasons why philosophical involvement in a volume on AI is beneficial. First, in truth, the distinction between AI research and philosophical work on mentality is not a sharp one. As a number of the chapters make clear, there are no tidy and well-defined demarcations between, say, programming rationality on the one hand and philosophical work on logic or reasoning on the other. Foundational questions and challenges regarding the development of synthetic minds are unavoidably philosophical in nature, and AI researchers often need to reflect on the broader implications of their findings, speculate about abstract matters such as hidden assumptions and overarching themes, appeal to thought experiments, and invoke traditional philosophical concepts, such as knowledge, representation, and action. In other words, there is a lot of philosophical reasoning involved in being a cutting-edge investigator in AI. At the same time, philosophers of mind need to be well-versed in the theories, models, and major developments in AI, so that their own contributions are informed and useful.

Second, philosophers of mind possess two attributes that provide them with a unique perspective on AI research itself. One is an understanding of the more general metaphysical, epistemological, and even ethical issues that arise in AI. These include questions about the nature of machine intelligence and consciousness, reductionism, levels of explanation, properties of symbols and representations, the moral status of non-humans, and so on. The other attribute is an appreciation of specific foundational issues in various areas of AI research. For centuries, philosophers have been thinking and writing about a wide array of phenomena that AI researchers are now exploring – phenomena such as intelligence, consciousness, rationality, mental representation, perceptual experience, and human action. These are *both* established areas of philosophical analysis *and* the target of increased scientific investigation. The philosophy of mind is itself being transformed by work in AI, and, at the same time, philosophers have a unique vantage from which they can elucidate empirical work, synthesizing ideas, making theoretical connections, and highlighting conceptual and methodological issues.

Despite this overlap, however, there has remained some distance between advanced AI researchers and those working in related disciplines, including philosophy. Consequently, a third and final reason to have input from relative outsiders is to help overcome this. The distance has been especially marked where AI research has been highly technical and focused upon specific applications and tasks. To some degree, AI researchers belong to a tighter, more homogeneous community than researchers in other areas of cognitive science, with a greater emphasis upon investigating precise theoretical problems or producing highly specific applications. This focus has led to considerable progress in many areas, but it has also made much of the work less accessible to those in other disciplines. Thus, one of our aims in compiling the present volume was to help bridge this gap by presenting even technical areas of research in an accessible manner, with a focus on general principles rather than specific details. In our role as editors, we view ourselves as representatives of readers who have a strong interest in AI but who do not currently work in the field and are not familiar with many developments. Consequently, we chose a mix of established figures and up-and-coming researchers, and then pressured them hard to explain the state of the art in a way that others could understand. The end result is a unique volume that allows those without formal training to gain a good overview of even highly complicated areas of AI research. We hope it will help to foster productive communication and collaboration between AI and other disciplines.

Summary of the volume

The volume is composed of fifteen chapters, collected into four sections: *Foundations*, *Architectures*, *Dimensions*, and *Extensions*. The focus narrows across the first three sections, which respectively survey foundational issues, general theories of mental architecture, and specific areas of research. The final section then extends the coverage to some related topics and research programs. Each section and each chapter stands alone and can be read individually (and in order to achieve this, we have allowed occasional overlap between chapter contents); but the chapters and sections are designed to complement each other, and the collection as a whole provides a systematic and comprehensive overview of the theoretical landscape of AI. Below we give a brief summary of the contents of each part of the volume.

Part I: Foundations

The essays in this section are devoted to explaining and discussing foundational issues in AI. In the first chapter, Stan Franklin introduces readers to the field, with a concise survey of the discipline's core themes, landmark moments, major accomplishments, main research areas, and recent trends. Chapter 2, by Konstantine Arkoudas and Selmer Bringsjord, looks in detail at the philosophical and conceptual roots of AI. For instance, the functionalist notion that the mind is like a computational program has traditionally served as a metaphysical basis for a great deal of work in AI, but it has also led to various powerful criticisms. Arkoudas and Bringsjord discuss these and other ways in which philosophy and AI intersect. The field of AI has also faced a number of philosophical challenges, both to the general project of creating artificial intelligence and to specific approaches within the field. In Chapter 3, William Robinson discusses the most important of these and the ways in which defenders have responded. For example, Robinson carefully explains Searle's well-known *Chinese Room* argument against the classical symbol-processing approach to AI and the responses that have been offered by proponents of that approach.

Part II: Architectures

This section of the volume looks at the three major architectural views (general theories of the representations and processes involved in intelligent thought and action) that have dominated work in AI. In Chapter 4, Margaret Boden provides an overview of the core features of classic computational architectures, often called GOFAI (for "Good Old-Fashioned AI"). She examines the strengths and weaknesses of the classical approach, and argues that the alleged failure of traditional AI has been grossly overstated, though she grants that it will probably need to be complemented by other approaches. Ron Sun presents a similar review of connectionism and neural networks in Chapter 5. After providing a historical perspective and an analysis of the distinctive features of connectionist models (such as distributed representations), Sun offers his own support for hybrid architectures, which combine both connectionist and traditional symbolic elements. In Chapter 6, Randall Beer gives an in-depth look at the dynamical and embedded approach to AI, or as he puts it, the Situated, Embodied, and Dynamical (SED) framework. As Beer notes, this approach focuses upon the dynamic interaction between intelligent systems and their environment, and leans heavily upon mathematical tools such as dynamical systems theory to capture that interaction through time. Beer provides an analysis of both the unique characteristics of the SED framework and its prospects for providing a radically new way of thinking about intelligent systems.

Part III: Dimensions

In this section, authors provide a glimpse into cutting-edge research in different subfields of AI, corresponding to different dimensions of intelligence: learning, perception, reasoning, language, action, and consciousness.

The section begins with Chapter 7, in which David Danks provides an overview of the different forms of machine learning and the sorts of algorithms and methods that have proven successful. The chapter also provides an interesting analysis of some of the technical and philosophical challenges that confront researchers in this area. In Chapter 8, Markus Vincze, Sven Wachsmuth, and Gerhard Sagerer review work in artificial perception, particularly computer vision. They give a detailed account of recent developments in such critical areas as object recognition and categorization, tracking and visual servoing, and vision-based human–computer interaction, and they offer their assessment of the key challenges that lie ahead. In Chapter 9, Eyal Amir presents a survey of recent work in artificial reasoning and decision making – topics which are central to the AI subfield known as *Knowledge Representation and Reasoning* (KR&R). Amir surveys techniques for representing information and reasoning with it (including ones drawn from propositional logic and probabilistic theory), looks at the various formalisms that have guided work in automated decision making, and highlights some cross-cutting issues such as the emergence of applications combining logic-based and probabilistic methods.

Two other important subfields of AI have been Natural Language Processing and Computational Linguistics, both of which focus on artificial language processing. In Chapter 10, Yorick Wilks provides some of the background to these fields and explains how they relate to core areas of linguistics, including syntactic processing, semantics, and lexical disambiguation. He also explains some of the quantitative and statistical approaches that have recently been employed to develop AI systems with linguistic capacities. In Chapter 11, Eduardo Alonso looks at agency in AI and discusses the strategies used to develop AI systems that function as competent agents in different environments, both as a consequence of stored knowledge and as a result of learning. In particular, Alonso highlights the importance of an emerging agent-centered AI, in which software systems are designed to display behavior that is autonomous, adaptive, and social. The chapter looks at the challenges involved in developing artificial agents that can operate in multi-agent environments by using techniques such as negotiation and argumentation. In the final chapter in this section, Chapter 12, Matthias Scheutz provides an in-depth look at recent work on developing machines that experience emotions and even some form of consciousness. Reviewing the historical and philosophical dimensions of the issue (such as the possible functional role of emotions), Scheutz examines the central challenges to this line of research, noting that we have yet to develop clear criteria for deciding when a system does or does not possess subjective states.

Part IV: Extensions

This final section extends the volume's coverage to topics and research programs that are closely related to AI, including robotics, artificial life, and the ethical dimensions of AI. In Chapter 13, Phil Husbands looks at the recent history of building machines that engage in intelligent behavior, both for industrial applications and for research purposes. He examines efforts to mimic biological systems, especially insects, and surveys evolutionary approaches, which use processes of iterated replication and mutation to generate robotic controllers adapted to specific tasks and environments. Chapter 14, by Mark Bedau, presents a fascinating look at the rapidly growing field of artificial life, which seeks to synthesize life in a variety of different forms, including robotic systems, artificial cells made from biochemical building blocks, and software agents inhabiting virtual eco-systems. After examining the ways in which artificial life is linked to conventional AI, Bedau explores what it has taught us about living systems and highlights some new philosophical issues it raises. In the final chapter of the volume, Chapter 15, Nick Bostrom and Eliezer Yudkowsky explore a wide range of ethical issues associated with the creation of thinking and feeling machines. These include questions about our use of AI algorithms to govern financial or legal transactions, the creation of artificial persons with moral status and rights, and the development of machines that are smarter than humans. As Bostrom and Yudkowsky correctly note, whereas these issues once belonged to the domain of science fiction, they are increasingly becoming real moral dilemmas we must face.

Coverage and scope

We shall add some remarks on the coverage and the scope of the volume. Our first point concerns the relation between the chapters. This handbook can be read straight through by those wanting a broad overview of the field of AI, but we also wanted it to be of use to those seeking to know only about specific topics or particular areas of AI research. Consequently, the chapters are written as stand-alone pieces that can be read individually for a thorough treatment of a given topic. Readers with a particular interest in, say, machine learning can go straight to David Danks' chapter and find an extensive examination of the topic that includes discussion of several closely related issues and does not assume familiarity with the material in earlier chapters. While we regard this as an important virtue of the volume, it also means that there is occasional overlap between different chapters. For example, any chapter on the philosophical foundations of AI, such as Arkoudas and Bringsjord's, would be incomplete without discussion of John Searle's famous Chinese Room argument, since that argument played a significant role in shaping the way some people think about the limits of AI. But, of course, discussion of Searle's argument also belongs in Robinson's chapter on the philosophical challenges to AI and in Boden's chapter on classical ("Good Old-Fashioned") AI as well. The same can be said for a variety of other arguments, ideas, theories, and developments that are critical for understanding the different aspects of AI described in these chapters.

While this overlap is unavoidable, it should be noted that each chapter has a unique focus, emphasizing different aspects of shared topics. Thus, Robinson places Searle's argument in the context of philosophical and conceptual challenges to AI, assessing the scope of the argument (via a useful distinction between flexibility and understanding) and explaining the major replies to the argument and Searle's responses to them. Boden, on the other hand, discusses the argument's impact on classical AI's supposed commitment to Strong AI – the view that running a suitable computer program is sufficient for mentality – and she goes on to discuss the extent to which classical AI is actually committed to that doctrine. Similar points could be made with regard to other topics that surface in different places, such as connectionism, dynamical systems theory, the Turing Test, and the frame problem. Thus, although readers may come across particular themes or topics in more than one chapter, we believe that each encounter will be informative and helpful.

Our second point concerns the scope of the volume. There are some topics relevant to AI research that are not covered in depth here. In particular, there is no systematic coverage of the technical, nuts-and-bolts aspects of AI research, such as programming languages and techniques, information search strategies, different styles of knowledge representation (such as scripts and frames), and so on – though there is some discussion of these matters in the chapters on the subfields of AI research. These "tools" of AI are of course essential topics for those planning to work in the field of AI. However, they are covered in detail in standard textbooks (see the "Further reading" section at the end of this Introduction), and, given space constraints, we decided to focus the volume primarily on what AI researchers are doing and thinking, rather than on how to actually do AI research.

Another area in which the coverage could have been greatly extended is that of applications and implications of AI research. Developments in AI are so far-reaching and have become so interwoven in our lives that our section on the extensions of AI could have been much, much larger. Indeed, whole volumes could be (and have been) devoted to topics such as AI and creativity, AI and the internet, AI on the battlefield, AI in film and literature, and so on. These issues are touched on in many places throughout the volume. However, this is not, strictly speaking, an "AI-And-Something-Else" volume, and we limited the Extensions section to two major research areas that are closely intertwined with AI – robotics and artificial life – and discussion of the major ethical issues arising directly from AI research. These well illustrate the important and fascinating ways in which the field of AI has grown, influenced other areas of inquiry, and generated new topics of enquiry, new problems, and new challenges.

The focus of the volume, then, is on the middle ground – the central conceptual and theoretical issues in AI and the major research programs – excluding technical details of AI system building on the one hand and the wider social, cultural, and economic implications of AI research on the other. We feel this makes for a manageable and accessible volume, which can easily be complemented by other works with a narrower or wider focus. We offer some suggestions for complementary works in the "Further reading" section at the end of this Introduction. Finally, as already mentioned, this volume has a companion, *The Cambridge Handbook of Cognitive Science*, which has a similar structure and scope and was designed to complement this one. That book looks at many of the same topics (the nature of mentality, cognitive architecture, the structure and function of various mental faculties) from the perspective of researchers on human cognition, as well as discussing allied research programs such as evolutionary psychology. AI and cognitive science have influenced each other strongly throughout their history, and the discussions in each volume will give further depth and perspective to those in the other.

General themes

Although the chapters were written individually, the reader will notice a number of common themes appearing throughout the volume. For example, in the past a good deal of AI research involved the promotion of specific architectures or theories, such as classic computationalism or connectionism. Often, simple toy systems were developed, in part to demonstrate how a given type of computational framework could in principle perform a particular task. However, with rapid developments in technology and computing power, it has become possible to build far more sophisticated AI systems that can do real work, and, as a result, pragmatic engineering concerns have become more central. Many researchers today are more concerned with employing whatever works best for the particular job in hand, rather than promoting a particular global architecture. Consequently, an increasing amount of AI work involves hybrid systems, as investigators employ different sorts of architectures for different sub-processes or tasks. This mirrors recent trends in cognitive science, such as "dual-process" theories of cognition, which claim that different types of computational architectures underlie different cognitive capacities. Much of the work discussed in this volume suggests that past theoretical divisions and battles over competing architectures are fading while a growing number of researchers are adopting a more inclusive and ecumenical outlook.

Another trend is a shift in the theoretical, conceptual, and philosophical issues that occupy investigators. Traditional questions about whether a computer could feel pain or about the nature of computational symbols are still important, but we also see new philosophical concerns becoming central. For example, the development of dynamic models in AI has encouraged people to reconsider what counts as a computational or intelligent system, the boundaries of such systems, and how intelligent behavior ought to be explained. At the same time, many investigators are rethinking the need for representations or internal models that in the past were assumed to be essential for guiding the behavior of AI systems. More and more investigators are accepting the importance of situatedness and embodiment and exploring the degree to which a system's interactions with a real-world environment are crucial. Thus, many of the chapters reveal that traditional assumptions are no longer taken for granted, and that radically new ideas and principles are taking center stage.

At the same time, developments in popular culture and technology are having a greater impact on how researchers regard the field, with investigators commonly making reference to the internet, to the explosion of special applications for smart devices, and to the influence of computer gaming. Similarly, many of the chapters reflect a sophisticated awareness of important findings in biology, animal cognition, neuroscience, and the psychology of perception. Thus, one thing that has not changed is the degree to which the field of AI continues to evolve as it assimilates the newest developments in technology and the most recent discoveries in the life sciences.

We have enjoyed putting this volume together and hope it will serve as an entry-point to the fascinating and hugely important work being done in AI. Work in this field already affects our lives in countless ways, and future developments are likely to transform our world radically. On the small scale, AI applications will increasingly pervade our lives, reshaping and enhancing our interactions with our environment and each other, while on a larger scale, theoretical advances in artificially modeling intelligence will not only give us a deeper understanding of our own minds, but may also confront us with new minds with vastly different capacities from ours. We hope this volume will help to make this exciting but often highly technical field more accessible, and that readers will be inspired to learn more about AI, to apply its insights in other fields, and to reflect upon its implications for humanity.

Further reading

As explained earlier, this volume's focus is on the middle ground of AI research – the main conceptual and theoretical issues and the major research programs. As such, the volume is designed to be self-standing. However, some readers may wish to complement it with other reading that has either a narrower focus, on technical matters, or a wider one, on the social and cultural implications of AI research. In this section we offer some brief suggestions for such reading.

Technical matters

One of the most popular undergraduate textbooks of AI concepts and techniques and among the best single-volume introductions to the nuts and bolts of AI is:

Russell, S. and Norvig, P. (2010). *Artificial Intelligence: A Modern Approach* (3rd edn.). Upper Saddle River, NJ: Prentice Hall.

Two other recommended volumes, on more specialized topics, are:

Norvig, P. (1992). *Paradigms of Artificial Intelligence Programming: Case Studies in Common Lisp*. San Francisco, CA: Morgan Kaufmann. A little dated, but still a classic textbook on AI programming, which teaches the reader how to build and debug real AI programs and illustrates important techniques and concepts.

Whitten, I. H., Frank, E., and Hall, M. A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques* (3rd edn.). Burlington, MA: Morgan Kaufmann. A standard textbook on the statistical techniques of machine learning and data mining, which are now central to much AI research. Written at an accessible introductory level.

Wider perspectives

There is a growing literature on the social, economic, and cultural effects of AI research, from which the following is a selection.