Preface

> Until very recently, systematic thought about knowledge, beliefs, attitudes, desires, preferences and obligations has been concerned with only two kinds of agents: humans and gods. Until this century, the best mechanical analog for thought was clockwork … and the fantasy of creating something with knowledge could be achieved only by giving the mysterious quality of life to some dead or inert mass, risking the gods' wrath or vengeance.
>
> – Introduction to *Android Epistemology*
> Ford, Glymour, and Hayes, 1995

This book is about the nature of cognition, both natural and artificial. It has grown out of a program of research into intelligent functions like reasoning, problem solving and decision-making. These are well-established research topics, but our program is unusual in its focus on the *integration* of these and related cognitive processes. Many cognitive scientists seek a unified theory of their subject matter but, as in many other fields of scientific enquiry, the discipline tends to fragment into more and more specialist areas and unification eludes us. Our long-term aim is to develop intellectual and methodological tools that will foster a unified cognitive science.

We have been concerned to validate the tools and underlying concepts by building practical demonstrations of computational intelligences that can solve complex problems by virtue of the integration of a range of cognitive functions. The criteria for measuring success in this programme, therefore, are part theoretical and part practical. In this book we present our theoretical proposals formally and demonstrate their effectiveness with examples of practical applications.

Medicine provides the specific context for our discussions and practical demonstrations. As medical researchers, as well as cognitive scientists, our aim has been to create technologies which can help

nurses, doctors and other medical professionals make clinical decisions when managing the diagnosis and treatment of life-threatening conditions, such as cancer.

One might imagine that achieving such practical objectives would be a straightforward engineering problem. By *engineering* we mean solving problems using established techniques, without needing to develop new theoretical concepts. In fact, for reasons of necessity, curiosity and accident, the work has led us to address problems that we think of as scientific rather than engineering in nature. The practical problems that we have addressed are so hard that they seem to require new concepts and principles if they are to be solved.

Many of the problems we consider arise in other fields, such as aerospace, industrial control, financial and commercial management systems, and robotics. We believe that the work we describe is relevant to any field where decision-making is complex, involves significant uncertainty and has important consequences.

Some ways in which applied research can inform our theoretical understanding can be seen from the relationships between a number of topics discussed in this book and the more general areas of enquiry that they typify. The topics in the left column of the table below are widely studied by researchers who are interested in medical thinking. These are actually just special cases of the more general research areas shown on the right, which are all topics that are extensively studied by cognitive scientists (notably AI researchers but also psychologists, economists, logicians, statisticians, decision analysts, and others).

Medical Cognition	**Cognitive Science**
Medical expertise	Goal-directed problem solving Theories of knowledge
Clinical inference	Reasoning Logic
Diagnosis Treatment selection Prescribing	Decision making
Medical judgment	Reasoning under uncertainty Formal knowledge representation
Patient management	Planning and acting Scheduling

The first theme of this book is to show how cognitive agents can be designed to have some or all of the above types of capability, and to carry out a range of complex, knowledge-intensive tasks. We demon-

strate the design of such agents primarily in the world of medicine, but the techniques are applicable to many other domains. The returns for this substantial effort in cognitive science and engineering are both a practical method for the principled design of AI systems and some insights into natural intelligence as well.

In common with many other fields, medicine can do harm as well as good, and medical technologies can cause more harm as well as more good than human beings working alone. The second major theme of this book is how we can make intelligent systems safe, and, in particular, what we can learn from how human agents manage hazards that can be put to use in designing artificial ones. We shall also take the opportunity to express a deep concern about the safety of all AI technologies, and the need for professional responsibility among AI researchers and engineers. We must be confident that intelligent systems can be relied upon to do what they are intended to do, and only what they are intended to do.

This is particularly true in fields like the aerospace and nuclear industries, where even small technical faults can have catastrophic consequences. The rapid development of computer systems and their applications has been accompanied by a proliferation of innovative software techniques that have made important contributions in many fields. But software has also become notorious for its potential to produce undesirable behavior in unanticipated circumstances. In aerospace and other technical fields, therefore, the use of formal mathematical methods is increasingly important in the design and verification of software. Since medicine is manifestly a safety-critical field, formal methods have strongly influenced the material presented in this book.

Most of the results reported here came out of two collaborative projects called PROMPT[1] (protocols for medical procedures and therapies) and RED[2] (rigorously engineered decisions). In the course of our work we came to an important conclusion: that we owe, to borrow a legal term, a duty of care to the patients and others who could potentially be harmed by our inventions.

We attempt to discharge this duty in several ways.

Our first obligation is to try to ensure that the designs of our systems are sound. We need to ask not only "do they work?" but also "do they work for good reasons?" Unfortunately, conventional software design is frequently ad hoc, and AI software design is little better and possibly worse. Therefore, the results of our efforts are not to be entirely trusted. Consequently, we place great emphasis on clear design principles, strong mathematical foundations for these princi-

ples and effective development tools that support and verify the integrity of the system. In this respect we are applying lessons from conventional safety engineering to construct a comprehensive and rigorous *safety case* which provides reassurance that the agent application can be confidently deployed.

We shall also present a somewhat unorthodox twist to the problem of safe design.

In a field like medicine, the levels of uncertainty (for example, about what can happen, and when) are so high that many hazards and emergencies can not be predicted by software designers who are working away from the realities and pressures of the clinical setting. Hazards will arise that were not anticipated during the design process and which can only be managed if, and when, they occur. A central proposal of this book is complementary to established methods of static software design and verification: the idea of *dynamic management* of hazards. In practical terms, agents should consider the potential consequences of any recommendations they make or actions they decide to take. For example, they need to be alert to possible interactions between the effects of their actions and unanticipated features of the clinical context, such as an unusual aspect of the patient's condition. Ideally they should anticipate all the hazards or dangers that could arise from what they do.

AI offers a distinctive approach to this problem; software agents should manage hazards and risks intelligently, using the same sorts of techniques they use for their normal reasoning and decision making. Specifically we shall develop the idea of a *guardian agent* that is specialized for managing hazards through its possession of a bag of tools and techniques which it can use if problems arise.

Whether the techniques we present in this book prove to be good, bad or indifferent, we believe that the AI community must take the safety issue seriously. We are creating a powerful technology, possibly more quickly than we think, that has unprecedented potential to create havoc as well as benefit. We urge the community to embark on a vigorous discussion of the issues and the creation of an explicit "safety culture" in the field.

Who Is the Book For?

This book has been written with several audiences in mind.

First of all the theoretical material is aimed at mainstream cogni-

tive scientists, particularly those interested in mathematical or psychological theories of reasoning, judgement, decision-making and planning.

Second, there are the practically inclined, notably those with an interest in medical informatics and healthcare professionals who are interested in technological issues; also, practical people who want to apply AI to other fields may see analogies that are applicable in their own domains.

Students (of medical decision making, computer science or cognitive science) may find the practical applications a complement to more academic course material.

Finally, we hope that researchers and technologists working on conventional approaches to software safety may find something of interest. We have drawn on many of their ideas in our discussions, and hope they will agree that AI can offer methods that complement their own contributions. We think that the idea of the *guardian agent* can be realized in lots of different ways for different purposes.

To all, we hope the book conveys a sense of excitement about the imminent possibilities of this field, as well as setting an important research challenge for us to ensure that AI technologies are safe and sound.

Acknowledgments

The European Union supported a number of research projects on the clinical decision support and patient management systems that are discussed in this book. Several of these applications are now in use. Apart from the practical benefits these projects provided opportunities to observe experienced doctors and others in a number of different medical domains, leading to general insights about how artificial agents might emulate some of the versatility and robustness of human expertise.

Most of the formal work in part three was developed during the RED project, which was supported by the UK's Engineering and Physical Sciences Research Council and the Department of Trade and Industry, project GR/H 12348. We gratefully acknowledge this funding, and the enthusiasm and efforts of many colleagues who have supported and contributed to the work. In particular we would like to thank Professor Bob Malcolm, the Director of the DTI's Software Safety Programme, who took an interest in our ideas from the begin-

ning. Dominic Semple was our ever-reliable and efficient project officer at EPSRC, and Alan Montgomery, David Elsdon and Tom Khabaza (Integral Solutions Ltd.) were our patient industrial collaborators. Rachel Burnett and Richard Susskind (Masons Solicitors), and Rob Martil and Isaac Khabaza (Lloyds Register) provided us with important insights from the world of law and the practical realities of industrial certification. Some of them are now working with *Infer*Med Ltd. a company that was created to exploit the research.

Also David Robertson, Tony Hunter, Ken Ford and several anonymous reviewers committed substantial time to reading and commenting on drafts of the manuscript. If the book has virtues it is in no small measure due to their efforts.

All those involved in the RED project remember Mike Clarke, one of the most multi-talented people we have ever known, and whose teaching skills and deep understanding of logic, mathematics and artificial intelligence gave us so much. He helped us to get the project started and gave us the confidence to confront the hazards of trying to say something new about safety. Sadly, Mike died of cancer during the early stages of our writing this book. We hope he would have approved of the results.

On a personal note JF wishes to thank Sir Walter Bodmer and Dr. Margaret Swain of the Imperial Cancer Research Fund, who provided long-term support for his laboratory, and Saki Hajnal, Ali Rahmanzadeh, Paul Taylor, and Richard Thomson, who have been close collaborators over many years. The work with Rick Cooper, on the development of the cogent simulation package and on empirical and computational studies of medical decision-making, has been one of the most productive and enjoyable anyone could wish for. The Vrij University of Amsterdam made it possible to visit Frank van Harmelen, Annette ten Teije, Jan Treur, Frances Brazier and their colleagues on several occasions and to benefit from their experience and criticism. Hugs and thanks too to Helen Johan, who kept things going at difficult moments, particularly during the production of the final drafts. Her gentle but firm editorial style has made this book a better read than it might otherwise have been.

Among many others who had a hand in this are Andrew Coulson, Zhan Cui, Paul Ferguson, David Frost, David Glasspool, Claude Gierl, Colin Gordon, Jun Huang, Nicky Johns, Andrew Newbigging, David Sutton, and Peter Yule, who provided much technical help as well as their amazingly diverse talents. The contribution of Peter Hammond of University College, London, particularly the material

in chapter 8 of the book, should be singled out for special thanks. The technical results could also not have been achieved without the clinical realism and imagination of many doctor colleagues, particularly Enrico Coiera, Robert Dunlop, Jon Emery, Andzrej Glowinski, Peter Johnson, Colin Lyons, Mike O'Neil, Jean-Louis Renaud-Salis, Nick Walker, Robert Walton, Peter Wilson and Jeremy Wyatt.

SD also wishes to acknowledge suggestions and constructive criticism of many colleagues over the past few years; the quality of the book has been greatly enhanced by these discussions. They include colleagues at Charles River Analytics, especially Paul Gonsalves, Greg Zacharias, and Dan Grecu, Sandeep Mulgund from Arthur D. Little, Alper Caglayan from Bowne International, and Peter Hammond, who contributed much to the logical developments. I am also highly indebted to each and every member of my family in India for their continuous encouragement and for patiently accepting my absence. The greatest sacrifice was certainly made by my wife Janique, who not only inspired me in the writing, but also provided the extra love and care for our wonderful children Sebastien and Kabita during my many absences from home.

Finally, we would both like to thank friends and colleagues at *Infer*Med Ltd. who helped to turn many of the ideas discussed in the book into the Arezzo® suite of products, and for permission to include screen shots which illustrate how some of the ideas can be used in practice. Also thanks to the staff of Charles River Analytics Inc. for their interest in this project, and their hospitality to JF during final production. Several lively discussions about whether our method of implementing intelligent agents could be used to design agents that could fly aircraft (safely) will be long remembered!

Notes

1. Funded by the European Union's Healthcare Informatics Programme (1995-98).

2. Funded by the UK's Department of Industry and Engineering and Physical Sciences Research Council. Participants in the RED project were the Imperial Cancer Research Fund, Queen Mary and Westfield College, London; Integral Solutions Ltd., Basingstoke; Lloyds Register, Croydon and Masons Solicitors, London.