

Artificial Intelligence

A Modern Approach

Fourth Edition



**PEARSON SERIES
IN ARTIFICIAL INTELLIGENCE**

Stuart Russell and Peter Norvig, Editors

FORSYTH & PONCE

GRAHAM

JURAFSKY & MARTIN

NEAPOLITAN

RUSSELL & NORVIG

Computer Vision: A Modern Approach, 2nd ed.

ANSI Common Lisp

Speech and Language Processing, 2nd ed.

Learning Bayesian Networks

Artificial Intelligence: A Modern Approach, 4th ed.

Artificial Intelligence

A Modern Approach

Fourth Edition

Stuart J. Russell and Peter Norvig

Contributing writers:

Ming-Wei Chang

Jacob Devlin

Anca Dragan

David Forsyth

Ian Goodfellow

Jitendra M. Malik

Vikash Mansinghka

Judea Pearl

Michael Wooldridge



Copyright © 2021, 2010, 2003 by Pearson Education, Inc. or its affiliates, 221 River Street, Hoboken, NJ 07030. All Rights Reserved. Manufactured in the United States of America. This publication is protected by copyright, and permission should be obtained from the publisher prior to any prohibited reproduction, storage in a retrieval system, or transmission in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise. For information regarding permissions, request forms, and the appropriate contacts within the Pearson Education Global Rights and Permissions department, please visit www.pearsoned.com/permissions/.

Acknowledgments of third-party content appear on the appropriate page within the text.

Cover Images:

Alan Turing – Science History Images/Alamy Stock Photo
Statue of Aristotle – Panos Karas/Shutterstock
Ada Lovelace – Pictorial Press Ltd/Alamy Stock Photo
Autonomous cars – Andrey Suslov/Shutterstock
Atlas Robot – Boston Dynamics, Inc.
Berkeley Campanile and Golden Gate Bridge – Ben Chu/Shutterstock
Background ghosted nodes – Eugene Sergeev/Alamy Stock Photo
Chess board with chess figure – Titania/Shutterstock
Mars Rover – Stocktrek Images, Inc./Alamy Stock Photo
Kasparov – KATHY WILLENS/AP Images

PEARSON, ALWAYS LEARNING is an exclusive trademark owned by Pearson Education, Inc. or its affiliates in the U.S. and/or other countries.

Unless otherwise indicated herein, any third-party trademarks, logos, or icons that may appear in this work are the property of their respective owners, and any references to third-party trademarks, logos, icons, or other trade dress are for demonstrative or descriptive purposes only. Such references are not intended to imply any sponsorship, endorsement, authorization, or promotion of Pearson’s products by the owners of such marks, or any relationship between the owner and Pearson Education, Inc., or its affiliates, authors, licensees, or distributors.

Library of Congress Cataloging-in-Publication Data

Names: Russell, Stuart J. (Stuart Jonathan), author. | Norvig, Peter, author.
Title: Artificial intelligence : a modern approach / Stuart J. Russell and Peter Norvig.
Description: Fourth edition. | Hoboken : Pearson, [2021] | Series: Pearson series in artificial intelligence | Includes bibliographical references and index. | Summary: “Updated edition of popular textbook on Artificial Intelligence.”— Provided by publisher.
Identifiers: LCCN 2019047498 | ISBN 9780134610993 (hardcover)
Subjects: LCSH: Artificial intelligence.
Classification: LCC Q335 .R86 2021 | DDC 006.3–dc23
LC record available at <https://lcn.loc.gov/2019047498>

ScoutAutomatedPrintCode



ISBN-10: 0-13-461099-7
ISBN-13: 978-0-13-461099-3

For Loy, Gordon, Lucy, George, and Isaac — S.J.R.

For Kris, Isabella, and Juliet — P.N.

Preface

Artificial Intelligence (AI) is a big field, and this is a big book. We have tried to explore the full breadth of the field, which encompasses logic, probability, and continuous mathematics; perception, reasoning, learning, and action; fairness, trust, social good, and safety; and applications that range from microelectronic devices to robotic planetary explorers to online services with billions of users.

The subtitle of this book is “A Modern Approach.” That means we have chosen to tell the story from a current perspective. We synthesize what is now known into a common framework, recasting early work using the ideas and terminology that are prevalent today. We apologize to those whose subfields are, as a result, less recognizable.

New to this edition

This edition reflects the changes in AI since the last edition in 2010:

- We focus more on machine learning rather than hand-crafted knowledge engineering, due to the increased availability of data, computing resources, and new algorithms.
- Deep learning, probabilistic programming, and multiagent systems receive expanded coverage, each with their own chapter.
- The coverage of natural language understanding, robotics, and computer vision has been revised to reflect the impact of deep learning.
- The robotics chapter now includes robots that interact with humans and the application of reinforcement learning to robotics.
- Previously we defined the goal of AI as creating systems that try to maximize expected utility, where the specific utility information—the objective—is supplied by the human designers of the system. Now we no longer assume that the objective is fixed and known by the AI system; instead, the system may be uncertain about the true objectives of the humans on whose behalf it operates. It must learn what to maximize and must function appropriately even while uncertain about the objective.
- We increase coverage of the impact of AI on society, including the vital issues of ethics, fairness, trust, and safety.
- We have moved the exercises from the end of each chapter to an online site. This allows us to continuously add to, update, and improve the exercises, to meet the needs of instructors and to reflect advances in the field and in AI-related software tools.
- Overall, about 25% of the material in the book is brand new. The remaining 75% has been largely rewritten to present a more unified picture of the field. 22% of the citations in this edition are to works published after 2010.

Overview of the book

The main unifying theme is the idea of an **intelligent agent**. We define AI as the study of agents that receive percepts from the environment and perform actions. Each such agent implements a function that maps percept sequences to actions, and we cover different ways to represent these functions, such as reactive agents, real-time planners, decision-theoretic

systems, and deep learning systems. We emphasize learning both as a construction method for competent systems and as a way of extending the reach of the designer into unknown environments. We treat robotics and vision not as independently defined problems, but as occurring in the service of achieving goals. We stress the importance of the task environment in determining the appropriate agent design.

Our primary aim is to convey the *ideas* that have emerged over the past seventy years of AI research and the past two millennia of related work. We have tried to avoid excessive formality in the presentation of these ideas, while retaining precision. We have included mathematical formulas and pseudocode algorithms to make the key ideas concrete; mathematical concepts and notation are described in Appendix A and our pseudocode is described in Appendix B.

This book is primarily intended for use in an undergraduate course or course sequence. The book has 28 chapters, each requiring about a week's worth of lectures, so working through the whole book requires a two-semester sequence. A one-semester course can use selected chapters to suit the interests of the instructor and students. The book can also be used in a graduate-level course (perhaps with the addition of some of the primary sources suggested in the bibliographical notes), or for self-study or as a reference.

Throughout the book, *important points* are marked with a triangle icon in the margin. Wherever a new **term** is defined, it is also noted in the margin. Subsequent significant uses of the **term** are in bold, but not in the margin. We have included a comprehensive index and an extensive bibliography.

The only prerequisite is familiarity with basic concepts of computer science (algorithms, data structures, complexity) at a sophomore level. Freshman calculus and linear algebra are useful for some of the topics.

Online resources

Online resources are available through pearsonhighered.com/cs-resources or at the book's Web site, aima.cs.berkeley.edu. There you will find:

- Exercises, programming projects, and research projects. These are no longer at the end of each chapter; they are online only. Within the book, we refer to an online exercise with a name like “Exercise 6.NARY.” Instructions on the Web site allow you to find exercises by name or by topic.
- Implementations of the algorithms in the book in Python, Java, and other programming languages (currently hosted at github.com/aimacode).
- A list of over 1400 schools that have used the book, many with links to online course materials and syllabi.
- Supplementary material and links for students and instructors.
- Instructions on how to report errors in the book, in the likely event that some exist.

Book cover

The cover depicts the final position from the decisive game 6 of the 1997 chess match in which the program Deep Blue defeated Garry Kasparov (playing Black), making this the first time a computer had beaten a world champion in a chess match. Kasparov is shown at the

top. To his right is a pivotal position from the second game of the historic Go match between former world champion Lee Sedol and DeepMind's ALPHAGO program. Move 37 by ALPHAGO violated centuries of Go orthodoxy and was immediately seen by human experts as an embarrassing mistake, but it turned out to be a winning move. At top left is an Atlas humanoid robot built by Boston Dynamics. A depiction of a self-driving car sensing its environment appears between Ada Lovelace, the world's first computer programmer, and Alan Turing, whose fundamental work defined artificial intelligence. At the bottom of the chess board are a Mars Exploration Rover robot and a statue of Aristotle, who pioneered the study of logic; his planning algorithm from *De Motu Animalium* appears behind the authors' names. Behind the chess board is a probabilistic programming model used by the UN Comprehensive Nuclear-Test-Ban Treaty Organization for detecting nuclear explosions from seismic signals.

Acknowledgments

It takes a global village to make a book. Over 600 people read parts of the book and made suggestions for improvement. The complete list is at aima.cs.berkeley.edu/ack.html; we are grateful to all of them. We have space here to mention only a few especially important contributors. First the contributing writers:

- Judea Pearl (Section 13.5, Causal Networks);
- Vikash Mansinghka (Section 15.3, Programs as Probability Models);
- Michael Wooldridge (Chapter 18, Multiagent Decision Making);
- Ian Goodfellow (Chapter 21, Deep Learning);
- Jacob Devlin and Ming-Wei Chang (Chapter 24, Deep Learning for Natural Language);
- Jitendra Malik and David Forsyth (Chapter 25, Computer Vision);
- Anca Dragan (Chapter 26, Robotics).

Then some key roles:

- Cynthia Yeung and Malika Cantor (project management);
- Julie Sussman and Tom Galloway (copyediting and writing suggestions);
- Omari Stephens (illustrations);
- Tracy Johnson (editor);
- Erin Ault and Rose Kernan (cover and color conversion);
- Nalin Chhibber, Sam Goto, Raymond de Lacaze, Ravi Mohan, Ciaran O'Reilly, Amit Patel, Dragomir Radiv, and Samagra Sharma (online code development and mentoring);
- Google Summer of Code students (online code development).

Stuart would like to thank his wife, Loy Sheflott, for her endless patience and boundless wisdom. He hopes that Gordon, Lucy, George, and Isaac will soon be reading this book after they have forgiven him for working so long on it. RUGS (Russell's Unusual Group of Students) have been unusually helpful, as always.

Peter would like to thank his parents (Torsten and Gerda) for getting him started, and his wife (Kris), children (Bella and Juliet), colleagues, boss, and friends for encouraging and tolerating him through the long hours of writing and rewriting.

About the Authors

Stuart Russell was born in 1962 in Portsmouth, England. He received his B.A. with first-class honours in physics from Oxford University in 1982, and his Ph.D. in computer science from Stanford in 1986. He then joined the faculty of the University of California at Berkeley, where he is a professor and former chair of computer science, director of the Center for Human-Compatible AI, and holder of the Smith–Zadeh Chair in Engineering. In 1990, he received the Presidential Young Investigator Award of the National Science Foundation, and in 1995 he was cowinner of the Computers and Thought Award. He is a Fellow of the American Association for Artificial Intelligence, the Association for Computing Machinery, and the American Association for the Advancement of Science, an Honorary Fellow of Wadham College, Oxford, and an Andrew Carnegie Fellow. He held the Chaire Blaise Pascal in Paris from 2012 to 2014. He has published over 300 papers on a wide range of topics in artificial intelligence. His other books include *The Use of Knowledge in Analogy and Induction*, *Do the Right Thing: Studies in Limited Rationality* (with Eric Wefald), and *Human Compatible: Artificial Intelligence and the Problem of Control*.

Peter Norvig is currently a Director of Research at Google, Inc., and was previously the director responsible for the core Web search algorithms. He co-taught an online AI class that signed up 160,000 students, helping to kick off the current round of massive open online classes. He was head of the Computational Sciences Division at NASA Ames Research Center, overseeing research and development in artificial intelligence and robotics. He received a B.S. in applied mathematics from Brown University and a Ph.D. in computer science from Berkeley. He has been a professor at the University of Southern California and a faculty member at Berkeley and Stanford. He is a Fellow of the American Association for Artificial Intelligence, the Association for Computing Machinery, the American Academy of Arts and Sciences, and the California Academy of Science. His other books are *Paradigms of AI Programming: Case Studies in Common Lisp*, *VerbMobil: A Translation System for Face-to-Face Dialog*, and *Intelligent Help Systems for UNIX*.

The two authors shared the inaugural AAAI/EAAI Outstanding Educator award in 2016.

Contents

I Artificial Intelligence

1 Introduction	1
1.1 What Is AI?	1
1.2 The Foundations of Artificial Intelligence	5
1.3 The History of Artificial Intelligence	17
1.4 The State of the Art	27
1.5 Risks and Benefits of AI	31
Summary	34
Bibliographical and Historical Notes	35
2 Intelligent Agents	36
2.1 Agents and Environments	36
2.2 Good Behavior: The Concept of Rationality	39
2.3 The Nature of Environments	42
2.4 The Structure of Agents	47
Summary	60
Bibliographical and Historical Notes	60

II Problem-solving

3 Solving Problems by Searching	63
3.1 Problem-Solving Agents	63
3.2 Example Problems	66
3.3 Search Algorithms	71
3.4 Uninformed Search Strategies	76
3.5 Informed (Heuristic) Search Strategies	84
3.6 Heuristic Functions	97
Summary	104
Bibliographical and Historical Notes	106
4 Search in Complex Environments	110
4.1 Local Search and Optimization Problems	110
4.2 Local Search in Continuous Spaces	119
4.3 Search with Nondeterministic Actions	122
4.4 Search in Partially Observable Environments	126
4.5 Online Search Agents and Unknown Environments	134
Summary	141
Bibliographical and Historical Notes	142
5 Adversarial Search and Games	146
5.1 Game Theory	146
5.2 Optimal Decisions in Games	148

5.3	Heuristic Alpha–Beta Tree Search	156
5.4	Monte Carlo Tree Search	161
5.5	Stochastic Games	164
5.6	Partially Observable Games	168
5.7	Limitations of Game Search Algorithms	173
	Summary	174
	Bibliographical and Historical Notes	175
6	Constraint Satisfaction Problems	180
6.1	Defining Constraint Satisfaction Problems	180
6.2	Constraint Propagation: Inference in CSPs	185
6.3	Backtracking Search for CSPs	191
6.4	Local Search for CSPs	197
6.5	The Structure of Problems	199
	Summary	203
	Bibliographical and Historical Notes	204
 III Knowledge, reasoning, and planning		
7	Logical Agents	208
7.1	Knowledge-Based Agents	209
7.2	The Wumpus World	210
7.3	Logic	214
7.4	Propositional Logic: A Very Simple Logic	217
7.5	Propositional Theorem Proving	222
7.6	Effective Propositional Model Checking	232
7.7	Agents Based on Propositional Logic	237
	Summary	246
	Bibliographical and Historical Notes	247
8	First-Order Logic	251
8.1	Representation Revisited	251
8.2	Syntax and Semantics of First-Order Logic	256
8.3	Using First-Order Logic	265
8.4	Knowledge Engineering in First-Order Logic	271
	Summary	277
	Bibliographical and Historical Notes	278
9	Inference in First-Order Logic	280
9.1	Propositional vs. First-Order Inference	280
9.2	Unification and First-Order Inference	282
9.3	Forward Chaining	286
9.4	Backward Chaining	293
9.5	Resolution	298
	Summary	309
	Bibliographical and Historical Notes	310

10 Knowledge Representation	314
10.1 Ontological Engineering	314
10.2 Categories and Objects	317
10.3 Events	322
10.4 Mental Objects and Modal Logic	326
10.5 Reasoning Systems for Categories	329
10.6 Reasoning with Default Information	333
Summary	337
Bibliographical and Historical Notes	338
11 Automated Planning	344
11.1 Definition of Classical Planning	344
11.2 Algorithms for Classical Planning	348
11.3 Heuristics for Planning	353
11.4 Hierarchical Planning	356
11.5 Planning and Acting in Nondeterministic Domains	365
11.6 Time, Schedules, and Resources	374
11.7 Analysis of Planning Approaches	378
Summary	379
Bibliographical and Historical Notes	380
IV Uncertain knowledge and reasoning	
12 Quantifying Uncertainty	385
12.1 Acting under Uncertainty	385
12.2 Basic Probability Notation	388
12.3 Inference Using Full Joint Distributions	395
12.4 Independence	397
12.5 Bayes' Rule and Its Use	399
12.6 Naive Bayes Models	402
12.7 The Wumpus World Revisited	404
Summary	407
Bibliographical and Historical Notes	408
13 Probabilistic Reasoning	412
13.1 Representing Knowledge in an Uncertain Domain	412
13.2 The Semantics of Bayesian Networks	414
13.3 Exact Inference in Bayesian Networks	427
13.4 Approximate Inference for Bayesian Networks	435
13.5 Causal Networks	449
Summary	453
Bibliographical and Historical Notes	454
14 Probabilistic Reasoning over Time	461
14.1 Time and Uncertainty	461
14.2 Inference in Temporal Models	465

14.3	Hidden Markov Models	473
14.4	Kalman Filters	479
14.5	Dynamic Bayesian Networks	485
	Summary	496
	Bibliographical and Historical Notes	497
15	Probabilistic Programming	500
15.1	Relational Probability Models	501
15.2	Open-Universe Probability Models	507
15.3	Keeping Track of a Complex World	514
15.4	Programs as Probability Models	519
	Summary	523
	Bibliographical and Historical Notes	524
16	Making Simple Decisions	528
16.1	Combining Beliefs and Desires under Uncertainty	528
16.2	The Basis of Utility Theory	529
16.3	Utility Functions	532
16.4	Multiattribute Utility Functions	540
16.5	Decision Networks	544
16.6	The Value of Information	547
16.7	Unknown Preferences	553
	Summary	557
	Bibliographical and Historical Notes	557
17	Making Complex Decisions	562
17.1	Sequential Decision Problems	562
17.2	Algorithms for MDPs	572
17.3	Bandit Problems	581
17.4	Partially Observable MDPs	588
17.5	Algorithms for Solving POMDPs	590
	Summary	595
	Bibliographical and Historical Notes	596
18	Multiagent Decision Making	599
18.1	Properties of Multiagent Environments	599
18.2	Non-Cooperative Game Theory	605
18.3	Cooperative Game Theory	626
18.4	Making Collective Decisions	632
	Summary	645
	Bibliographical and Historical Notes	646
V	Machine Learning	
19	Learning from Examples	651
19.1	Forms of Learning	651

19.2	Supervised Learning	653
19.3	Learning Decision Trees	657
19.4	Model Selection and Optimization	665
19.5	The Theory of Learning	672
19.6	Linear Regression and Classification	676
19.7	Nonparametric Models	686
19.8	Ensemble Learning	696
19.9	Developing Machine Learning Systems	704
	Summary	714
	Bibliographical and Historical Notes	715
20	Learning Probabilistic Models	721
20.1	Statistical Learning	721
20.2	Learning with Complete Data	724
20.3	Learning with Hidden Variables: The EM Algorithm	737
	Summary	746
	Bibliographical and Historical Notes	747
21	Deep Learning	750
21.1	Simple Feedforward Networks	751
21.2	Computation Graphs for Deep Learning	756
21.3	Convolutional Networks	760
21.4	Learning Algorithms	765
21.5	Generalization	768
21.6	Recurrent Neural Networks	772
21.7	Unsupervised Learning and Transfer Learning	775
21.8	Applications	782
	Summary	784
	Bibliographical and Historical Notes	785
22	Reinforcement Learning	789
22.1	Learning from Rewards	789
22.2	Passive Reinforcement Learning	791
22.3	Active Reinforcement Learning	797
22.4	Generalization in Reinforcement Learning	803
22.5	Policy Search	810
22.6	Apprenticeship and Inverse Reinforcement Learning	812
22.7	Applications of Reinforcement Learning	815
	Summary	818
	Bibliographical and Historical Notes	819
VI Communicating, perceiving, and acting		
23	Natural Language Processing	823
23.1	Language Models	823
23.2	Grammar	833

23.3	Parsing	835
23.4	Augmented Grammars	841
23.5	Complications of Real Natural Language	845
23.6	Natural Language Tasks	849
	Summary	850
	Bibliographical and Historical Notes	851
24	Deep Learning for Natural Language Processing	856
24.1	Word Embeddings	856
24.2	Recurrent Neural Networks for NLP	860
24.3	Sequence-to-Sequence Models	864
24.4	The Transformer Architecture	868
24.5	Pretraining and Transfer Learning	871
24.6	State of the art	875
	Summary	878
	Bibliographical and Historical Notes	878
25	Computer Vision	881
25.1	Introduction	881
25.2	Image Formation	882
25.3	Simple Image Features	888
25.4	Classifying Images	895
25.5	Detecting Objects	899
25.6	The 3D World	901
25.7	Using Computer Vision	906
	Summary	919
	Bibliographical and Historical Notes	920
26	Robotics	925
26.1	Robots	925
26.2	Robot Hardware	926
26.3	What kind of problem is robotics solving?	930
26.4	Robotic Perception	931
26.5	Planning and Control	938
26.6	Planning Uncertain Movements	956
26.7	Reinforcement Learning in Robotics	958
26.8	Humans and Robots	961
26.9	Alternative Robotic Frameworks	968
26.10	Application Domains	971
	Summary	974
	Bibliographical and Historical Notes	975
VII Conclusions		
27	Philosophy, Ethics, and Safety of AI	981
27.1	The Limits of AI	981

27.2	Can Machines Really Think?	984
27.3	The Ethics of AI	986
	Summary	1005
	Bibliographical and Historical Notes	1006
28	The Future of AI	1012
28.1	AI Components	1012
28.2	AI Architectures	1018
A	Mathematical Background	1023
A.1	Complexity Analysis and $O()$ Notation	1023
A.2	Vectors, Matrices, and Linear Algebra	1025
A.3	Probability Distributions	1027
	Bibliographical and Historical Notes	1029
B	Notes on Languages and Algorithms	1030
B.1	Defining Languages with Backus–Naur Form (BNF)	1030
B.2	Describing Algorithms with Pseudocode	1031
B.3	Online Supplemental Material	1032
	Bibliography	1033
	Index	1069

