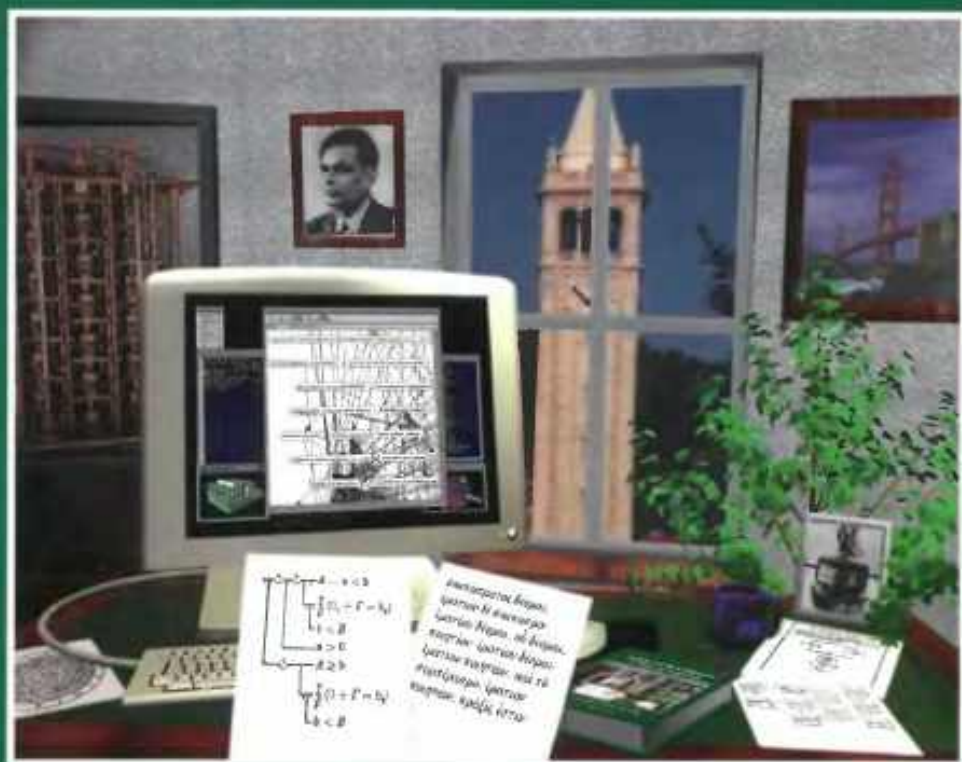


Книга об
интеллектуальных
агентах

Искусственный интеллект

Современный подход

ВТОРОЕ ИЗДАНИЕ



Стюарт Рассел • Питер Норvig

Artificial Intelligence

A Modern Approach

Second edition

Stuart J. Russel and Peter Norvig



Prentice Hall
Upper Saddle River, New Jersey 07458

Искусственный интеллект

Современный подход

Второе издание

Стюарт Рассел, Питер Норвиг

Москва • Санкт-Петербург • Киев
2006



ББК 32.973.26-018.2.75

P24

УДК 681.3.07

Издательский дом “Вильямс”

Зав. редакцией *С.Н. Тригуб*

Перевод с английского и редакция *К.А. Птицына*

По общим вопросам обращайтесь в Издательский дом “Вильямс” по адресу:

info@williamspublishing.com, <http://www.williamspublishing.com>

115419, Москва, а/я 783; 03150, Киев, а/я 152

Рассел, Стюарт, Норвиг, Питер.

P24 Искусственный интеллект: современный подход, 2-е изд.: Пер. с англ. — М. : Издательский дом “Вильямс”, 2006. — 1408 с. : ил. — Парал. тит. англ.

ISBN 5-8459-0887-6 (рус.)

В книге представлены все современные достижения и изложены идеи, которые были сформулированы в исследованиях, проводившихся в течение последних пятидесяти лет, а также собраны на протяжении двух тысячелетий в областях знаний, ставших стимулом к развитию искусственного интеллекта как науки проектирования рациональных агентов. Теоретическое описание иллюстрируется многочисленными алгоритмами, реализации которых в виде готовых программ на нескольких языках программирования находятся на сопровождающем книгу Web-узле.

Книга предназначена для использования в базовом университетском курсе или в последовательности курсов по специальности. Применима в качестве основного справочника для аспирантов, специализирующихся в области искусственного интеллекта, а также будет небезынтересна профессионалам, желающим выйти за пределы избранной ими специальности. Благодаря кристальной ясности и наглядности изложения вполне может быть отнесена к лучшим образцам научно-популярной литературы.

ББК 32.973.26-018.2.75

Все названия программных продуктов являются зарегистрированными торговыми марками соответствующих фирм.

Никакая часть настоящего издания ни в каких целях не может быть воспроизведена в какой бы то ни было форме и какими бы то ни было средствами, будь то электронные или механические, включая фотокопирование и запись на магнитный носитель, если на это нет письменного разрешения издательства Prentice Hall, Inc.

Authorized translation from the English language edition published by Prentice Hall, Copyright © 2003 by Pearson Education, Inc.

All rights reserved. No part of this book may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording or by any information storage retrieval system, without permission from the Publisher.

Russian language edition was published by Williams Publishing House according to the Agreement with R&I Enterprises International, Copyright © 2006

ISBN 5-8459-0887-6 (рус.)
ISBN 0-13-790395-2 (англ.)

© Издательский дом “Вильямс”, 2006
© by Pearson Education, Inc., 2003

ОГЛАВЛЕНИЕ

Предисловие	24
Об авторах	31
 Часть I. ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ	 33
Глава 1. Введение	34
Глава 2. Интеллектуальные агенты	75
 Часть II. РЕШЕНИЕ ПРОБЛЕМ	 109
Глава 3. Решение проблем посредством поиска	110
Глава 4. Информированный поиск и исследование пространства состояний	153
Глава 5. Задачи удовлетворения ограничений	209
Глава 6. Поиск в условиях противодействия	240
 Часть III. ЗНАНИЯ И РАССУЖДЕНИЯ	 281
Глава 7. Логические агенты	282
Глава 8. Логика первого порядка	341
Глава 9. Логический вывод в логике первого порядка	380
Глава 10. Представление знаний	440
 Часть IV. ПЛАНИРОВАНИЕ	 511
Глава 11. Основы планирования	512
Глава 12. Планирование и осуществление действий в реальном мире	564
 Часть V. НЕОПРЕДЕЛЕННЫЕ ЗНАНИЯ И РАССУЖДЕНИЯ В УСЛОВИЯХ НЕОПРЕДЕЛЕННОСТИ	 621
Глава 13. Неопределенность	622
Глава 14. Вероятностные рассуждения	660
Глава 15. Вероятностные рассуждения во времени	718
Глава 16. Принятие простых решений	778
Глава 17. Принятие сложных решений	815
 Часть VI. ОБУЧЕНИЕ	 863
Глава 18. Обучение на основе наблюдений	864
Глава 19. Применение знаний в обучении	902
Глава 20. Статистические методы обучения	945
Глава 21. Обучение с подкреплением	1010

Часть VII. ОБЩЕНИЕ, ВОСПРИЯТИЕ И ОСУЩЕСТВЛЕНИЕ ДЕЙСТВИЙ	1045
Глава 22. Общение	1046
Глава 23. Вероятностная обработка лингвистической информации	1102
Глава 24. Восприятие	1141
Глава 25. Робототехника	1188
Часть VIII. ЗАКЛЮЧЕНИЕ	1247
Глава 26. Философские основания	1248
Глава 27. Настоящее и будущее искусственного интеллекта	1277
Приложение А. Математические основы	1288
Приложение Б. Общие сведения о языках и алгоритмах, используемых в книге	1297
Литература	1302
Предметный указатель	1373

СОДЕРЖАНИЕ

Предисловие	24
Краткий обзор книги	25
Отличия от первого издания	26
Как использовать эту книгу	27
Использование Web-узла	28
Благодарности	28
Об обложке	30
Об авторах	31
Часть I. Искусственный интеллект	33
Глава 1. Введение	34
1.1. Общее определение искусственного интеллекта	34
Проверка того, способен ли компьютер действовать подобно человеку: подход, основанный на использовании теста Тьюринга	36
Как мыслить по-человечески: подход, основанный на когнитивном моделировании	37
Как мыслить рационально: подход, основанный на использовании “законов мышления”	38
Как мыслить рационально: подход, основанный на использовании рационального агента	39
1.2. Предыстория искусственного интеллекта	40
Философия (период с 428 года до н.э. по настоящее время)	40
Математика (период примерно с 800 года по настоящее время)	43
Экономика (период с 1776 года по настоящее время)	45
Неврология (период с 1861 года по настоящее время)	46
Психология (период с 1879 года по настоящее время)	49
Вычислительная техника (период с 1940 года по настоящее время)	51
Теория управления и кибернетика (период с 1948 года по настоящее время)	52
Лингвистика (период с 1957 года по настоящее время)	53
1.3. История искусственного интеллекта	54
Появление предпосылок искусственного интеллекта (период с 1943 года по 1955 год)	54
Рождение искусственного интеллекта (1956 год)	55
Ранний энтузиазм, большие ожидания (период с 1952 года по 1969 год)	56
Столкновение с реальностью (период с 1966 года по 1973 год)	60
Системы, основанные на знаниях: могут ли они стать ключом к успеху (период с 1969 года по 1979 год)	62

Превращение искусственного интеллекта в индустрию (период с 1980 года по настоящее время)	65
Возвращение к нейронным сетям (период с 1986 года по настоящее время)	65
Превращение искусственного интеллекта в науку (период с 1987 года по настоящее время)	66
Появление подхода, основанного на использовании интеллектуальных агентов (период с 1995 года по настоящее время)	68
1.4. Современное состояние разработок	69
1.5. Резюме	71
Библиографические и исторические заметки	72
Упражнения	73
 ГЛАВА 2. ИНТЕЛЛЕКТУАЛЬНЫЕ АГЕНТЫ	75
2.1. Агенты и варианты среды	75
2.2. Качественное поведение: концепция рациональности	78
Показатели производительности	78
Рациональность	79
Всезнание, обучение и автономность	80
2.3. Определение характера среды	82
Определение проблемной среды	83
Свойства проблемной среды	86
2.4. Структура агентов	90
Программы агентов	91
Простые рефлексные агенты	93
Рефлексные агенты, основанные на модели	96
Агенты, основанные на цели	97
Агенты, основанные на полезности	99
Обучающиеся агенты	100
2.5. Резюме	103
Библиографические и исторические заметки	104
Упражнения	106
 ЧАСТЬ II. РЕШЕНИЕ ПРОБЛЕМ	109
 ГЛАВА 3. РЕШЕНИЕ ПРОБЛЕМ ПОСРЕДСТВОМ ПОИСКА	110
3.1. Агенты, решающие задачи	110
Хорошо структурированные задачи и решения	113
Формулировка задачи	115
3.2. Примеры задач	116
Упрощенные задачи	116
Реальные задачи	120
3.3. Поиск решений	122
Измерение производительности решения задачи	126
3.4. Стратегии неинформированного поиска	127
Поиск в ширину	127

Поиск в глубину	130
Поиск с ограничением глубины	131
Поиск в глубину с итеративным углублением	133
Двунаправленный поиск	135
Сравнение стратегий неинформированного поиска	136
3.5. Предотвращение формирования повторяющихся состояний	136
3.6. Поиск с частичной информацией	139
Проблемы отсутствия датчиков	140
Проблемы непредвиденных ситуаций	142
3.7. Резюме	144
Библиографические и исторические заметки	145
Упражнения	147
 ГЛАВА 4. ИНФОРМИРОВАННЫЙ ПОИСК И ИССЛЕДОВАНИЕ ПРОСТРАНСТВА СОСТОЯНИЙ	 153
4.1. Стратегии информированного (эвристического) поиска	154
Жадный поиск по первому наилучшему совпадению	155
Поиск A*: минимизация суммарной оценки стоимости решения	157
Эвристический поиск с ограничением объема памяти	163
Обучение лучшим способам поиска	166
4.2. Эвристические функции	167
Зависимость производительности поиска от точности эвристической функции	168
Составление допустимых эвристических функций	170
Изучение эвристических функций на основе опыта	173
4.3. Алгоритмы локального поиска и задачи оптимизации	174
Поиск с восхождением к вершине	175
Поиск с эмуляцией отжига	180
Локальный лучевой поиск	181
Генетические алгоритмы	182
4.4. Локальный поиск в непрерывных пространствах	187
4.5. Поисковые агенты, действующие в оперативном режиме, и неизвестные варианты среды	189
Задачи поиска в оперативном режиме	190
Агенты, выполняющие поиск в оперативном режиме	193
Локальный поиск в оперативном режиме	194
Обучение в ходе поиска в оперативном режиме	197
4.6. Резюме	198
Библиографические и исторические заметки	199
Упражнения	204
 ГЛАВА 5. ЗАДАЧИ УДОВЛЕТВОРЕНИЯ ОГРАНИЧЕНИЙ	 209
5.1. Задачи удовлетворения ограничений	210
5.2. Применение поиска с возвратами для решения задач CSP	214
Упорядочение переменных и значений	217
Распространение информации с помощью ограничений	219

Интеллектуальный поиск с возвратами: поиск в обратном направлении	224
5.3. Применение локального поиска для решения задач удовлетворения ограничений	226
5.4. Структура задач	228
5.5. Резюме	233
Библиографические и исторические заметки	233
Упражнения	236
 ГЛАВА 6. ПОИСК В УСЛОВИЯХ ПРОТИВОДЕЙСТВИЯ	 240
6.1. Игры	240
6.2. Принятие оптимальных решений в играх	242
Оптимальные стратегии	242
Минимаксный алгоритм	245
Оптимальные решения в играх с несколькими игроками	246
6.3. Альфа-бета-отсечение	247
6.4. Неидеальные решения, принимаемые в реальном времени	252
Функции оценки	252
Прекращение поиска	254
6.5. Игры, которые включают элемент случайности	257
Оценка позиции в играх с узлами жеребьевки	260
Сложность оценки ожидаемых минимаксных значений	261
Карточные игры	262
6.6. Современные игровые программы	264
6.7. Обсуждение изложенных сведений	268
6.8. Резюме	270
Библиографические и исторические заметки	271
Упражнения	276
 ЧАСТЬ III. ЗНАНИЯ И РАССУЖДЕНИЯ	 281
 ГЛАВА 7. ЛОГИЧЕСКИЕ АГЕНТЫ	 282
7.1. Агенты, основанные на знаниях	284
7.2. Мир вампуса	286
7.3. Логика	290
7.4. Пропозициональная логика: очень простая логика	294
Синтаксис	295
Семантика	296
Простая база знаний	299
Логический вывод	300
Эквивалентность, допустимость и выполнимость	301
7.5. Шаблоны формирования рассуждений в пропозициональной логике	303
Резолюция	306
Прямой и обратный логический вывод	311
7.6. Эффективный пропозициональный логический вывод	316
Полный алгоритм поиска с возвратами	316

Алгоритмы локального поиска	318
Трудные задачи определения выполнимости	320
7.7. Агенты, основанные на пропозициональной логике	322
Поиск ям и вампусов с помощью логического вывода	322
Слежение за местонахождением и ориентацией	324
Агенты на основе логических схем	325
Сопоставление двух описанных типов агентов	330
7.8. Резюме	332
Библиографические и исторические заметки	333
Упражнения	337
 ГЛАВА 8. ЛОГИКА ПЕРВОГО ПОРЯДКА	 341
8.1. Дополнительные сведения о представлении	341
8.2. Синтаксис и семантика логики первого порядка	347
Модели для логики первого порядка	347
Символы и интерпретации	349
Термы	351
Атомарные высказывания	351
Сложные высказывания	352
Кванторы	352
Равенство	357
8.3. Использование логики первого порядка	357
Утверждения и запросы в логике первого порядка	358
Проблемная область родства	358
Числа, множества и списки	361
Мир вампуса	363
8.4. Инженерия знаний с применением логики первого порядка	366
Процесс инженерии знаний	367
Проблемная область электронных схем	369
8.5. Резюме	374
Библиографические и исторические заметки	374
Упражнения	376
 ГЛАВА 9. ЛОГИЧЕСКИЙ ВЫВОД В ЛОГИКЕ ПЕРВОГО ПОРЯДКА	 380
9.1. Сравнение методов логического вывода в пропозициональной логике и логике первого порядка	381
Правила логического вывода для кванторов	381
Приведение к пропозициональному логическому выводу	382
9.2. Унификация и поднятие	384
Правило вывода в логике первого порядка	384
Унификация	386
Хранение и выборка	388
9.3. Прямой логический вывод	390
Определенные выражения в логике первого порядка	390
Простой алгоритм прямого логического вывода	392
Эффективный прямой логический вывод	394

9.4. Обратный логический вывод	399
Алгоритм обратного логического вывода	399
Логическое программирование	401
Эффективная реализация логических программ	403
Избыточный логический вывод и бесконечные циклы	406
Логическое программирование в ограничениях	408
9.5. Резолюция	409
Конъюнктивная нормальная форма для логики первого порядка	410
Правило логического вывода с помощью резолюции	412
Примеры доказательств	413
Полнота резолюции	416
Учет отношения равенства	420
Стратегии резолюции	421
Средства автоматического доказательства теорем	423
9.6. Резюме	428
Библиографические и исторические заметки	429
Упражнения	435
 ГЛАВА 10. ПРЕДСТАВЛЕНИЕ ЗНАНИЙ	 440
10.1. Онтологическая инженерия	440
10.2. Категории и объекты	443
Физическая композиция	445
Меры	448
Вещества и объекты	449
10.3. Действия, ситуации и события	451
Онтология ситуационного исчисления	451
Описание действий в ситуационном исчислении	453
Решение проблемы представительного окружения	455
Решение проблемы выводимого окружения	457
Исчисление времени и событий	459
Обобщенные события	460
Процессы	462
Интервалы	464
Флюентные высказывания и объекты	465
10.4. Мыслительные события и мыслимые объекты	466
Формальная теория убеждений	467
Знания и убеждения	469
Знания, время и действия	470
10.5. Мир покупок в Internet	471
Сравнение коммерческих предложений	476
10.6. Системы формирования рассуждений о категориях	477
Семантические сети	478
Описательные логики	482
10.7. Формирование рассуждений с использованием информации,	
заданной по умолчанию	483
Открытые и закрытые миры	484

Отрицание как недостижение цели и устойчивая семантика модели	486
Логика косвенного описания и логика умолчания	488
10.8. Системы поддержки истинности	491
10.9. Резюме	494
Библиографические и исторические заметки	495
Упражнения	503

Часть IV. Планирование	511
------------------------	-----

Глава 11. Основы планирования	512
11.1. Задача планирования	513
Язык задач планирования	514
Выразительность и расширения языка	516
Пример: воздушный грузовой транспорт	518
Пример: задача с запасным колесом	519
Пример: мир блоков	520
11.2. Планирование с помощью поиска в пространстве состояний	521
Прямой поиск в пространстве состояний	522
Обратный поиск в пространстве состояний	523
Эвристики для поиска в пространстве состояний	525
11.3. Планирование с частичным упорядочением	527
Пример планирования с частичным упорядочением	532
Планирование с частичным упорядочением и несвязанными переменными	534
Эвристики для планирования с частичным упорядочением	535
11.4. Графы планирования	536
Применение графов планирования для получения эвристической оценки	539
Алгоритм Graphplan	541
Завершение работы алгоритма Graphplan	544
11.5. Планирование с помощью пропозициональной логики	545
Описание задач планирования в пропозициональной логике	546
Сложности, связанные с использованием пропозициональных кодировок	549
11.6. Анализ различных подходов к планированию	551
11.7. Резюме	553
Библиографические и исторические заметки	554
Упражнения	558

Глава 12. Планирование и осуществление действий в реальном мире	564
---	-----

12.1. Время, расписания и ресурсы	564
Составление расписаний с ресурсными ограничениями	567
12.2. Планирование иерархической сети задач	570
Представление декомпозиций действий	572
Модификация планировщика для его использования в сочетании с декомпозициями	574

Обсуждение вопроса	577
12.3. Планирование и осуществление действий в недетерминированных проблемных областях	580
12.4. Условное планирование	584
Условное планирование в полностью наблюдаемых вариантах среды	584
Условное планирование в частично наблюдаемых вариантах среды	589
12.5. Контроль выполнения и перепланирование	594
12.6. Непрерывное планирование	600
12.7. Мультиагентное планирование	605
Кооперация: совместные цели и планы	605
Многопольное планирование	606
Механизмы координации	608
Конкуренция	610
12.8. Резюме	611
Библиографические и исторические заметки	612
Упражнения	616

Часть V. НЕОПРЕДЕЛЕННЫЕ ЗНАНИЯ И РАССУЖДЕНИЯ В УСЛОВИЯХ НЕОПРЕДЕЛЕННОСТИ 621

ГЛАВА 13. НЕОПРЕДЕЛЕННОСТЬ	622
13.1. Действия в условиях неопределенности	622
Учет наличия неопределенных знаний	623
Неопределенность и рациональные решения	626
Проект агента, действующего в соответствии с теорией решений	627
13.2. Основная вероятностная система обозначений	628
Высказывания	628
Атомарные события	629
Априорная вероятность	630
Условная вероятность	632
13.3. Аксиомы вероятностей	635
Использование аксиом вероятностей	635
Теоретическое обоснование аксиом вероятностей	636
13.4. Логический вывод с использованием полных совместных распределений	638
13.5. Независимость	642
13.6. Правило Байеса и его использование	644
Применение правила Байеса: простой случай	644
Использование правила Байеса: комбинирование свидетельств	646
13.7. Еще одно возвращение в мир вампуса	648
13.8. Резюме	652
Библиографические и исторические заметки	653
Упражнения	656

ГЛАВА 14. ВЕРОЯТНОСТНЫЕ РАССУЖДЕНИЯ	660
14.1. Представление знаний в неопределенной проблемной области	660

14.2. Семантика байесовских сетей	664
Представление полного совместного распределения	664
Отношения условной независимости в байесовских сетях	669
14.3. Эффективное представление распределений условных вероятностей	669
14.4. Точный вероятностный вывод в байесовских сетях	675
Вероятностный вывод с помощью перебора	676
Алгоритм устранения переменной	678
Сложность точного вероятностного вывода	681
Алгоритмы кластеризации	682
14.5. Приближенный вероятностный вывод в байесовских сетях	683
Методы непосредственной выборки	684
Вероятностный вывод по методу моделирования цепи Маркова	690
14.6. Распространение вероятностных методов на представления в логике первого порядка	694
14.7. Другие подходы к формированию рассуждений в условиях неопределенности	699
Методы на основе правил для формирования рассуждений в условиях неопределенности	700
Представление незнания: теория Демпстера–Шефера	703
Представление неосведомленности: нечеткие множества и нечеткая логика	704
14.8. Резюме	706
Библиографические и исторические заметки	707
Упражнения	712
 ГЛАВА 15. ВЕРОЯТНОСТНЫЕ РАССУЖДЕНИЯ ВО ВРЕМЕНИ	 718
15.1. Время и неопределенность	719
Состояния и наблюдения	719
Стационарные процессы и марковское предположение	720
15.2. Вероятностный вывод во временных моделях	724
Фильтрация и предсказание	725
Сглаживание	728
Поиск наиболее вероятной последовательности	731
15.3. Скрытые марковские модели	734
Упрощенные матричные алгоритмы	734
15.4. Фильтры Калмана	737
Обновление гауссовых распределений	738
Простой одномерный пример	739
Общий случай	743
Области применения калмановской фильтрации	744
15.5. Динамические байесовские сети	746
Процедура создания сетей DBN	747
Точный вероятностный вывод в сетях DBN	752
Приближенный вероятностный вывод в сетях DBN	754
15.6. Распознавание речи	758
Звуки речи	760

Слова	763
Предложения	765
Разработка устройства распознавания речи	769
15.7. Резюме	770
Библиографические и исторические заметки	771
Упражнения	774
 ГЛАВА 16. ПРИНЯТИЕ ПРОСТЫХ РЕШЕНИЙ	 778
16.1. Совместный учет убеждений и желаний в условиях неопределенности	778
16.2. Основы теории полезности	780
Ограничения, налагаемые на рациональные предпочтения	781
В начале была Полезность	783
16.3. Функции полезности	784
Полезность денег	784
Шкалы полезности и оценка полезности	788
16.4. Многоатрибутные функции полезности	790
Доминирование	790
Структура предпочтений и многоатрибутная полезность	793
16.5. Сети принятия решений	795
Способы представления задачи принятия решений с помощью сети принятия решений	795
Вычисления с помощью сетей принятия решений	798
16.6. Стоимость информации	798
Простой пример	799
Общая формула	800
Свойства показателей стоимости информации	802
Реализация агента, действующего на основе сбора информации	802
16.7. Экспертные системы, основанные на использовании теории принятия решений	803
16.8. Резюме	807
Библиографические и исторические заметки	808
Упражнения	810
 ГЛАВА 17. ПРИНЯТИЕ СЛОЖНЫХ РЕШЕНИЙ	 815
17.1. Задачи последовательного принятия решений	816
Пример	816
Оптимальность в задачах последовательного принятия решений	819
17.2. Итерация по значениям	822
Полезности состояний	823
Алгоритм итерации по значениям	824
Сходимость итерации по значениям	826
17.3. Итерация по стратегиям	829
17.4. Марковские процессы принятия решений в частично наблюдаемых вариантах среды	831
17.5. Агенты, действующие на основе теории решений	836
17.6. Принятие решений при наличии нескольких агентов: теория игр	839

17.7. Проектирование механизма	851
17.8. Резюме	855
Библиографические и исторические заметки	856
Упражнения	859

Часть VI. ОБУЧЕНИЕ	863
---------------------------	------------

Глава 18. ОБУЧЕНИЕ НА ОСНОВЕ НАБЛЮДЕНИЙ	864
--	------------

18.1. Формы обучения	864
18.2. Индуктивное обучение	867
18.3. Формирование деревьев решений на основе обучения	870
Деревья решений, рассматриваемые как производительные элементы	870
Выразительность деревьев решений	872
Индуктивный вывод деревьев решений на основе примеров	873
Выбор проверок атрибутов	877
Оценка производительности обучающего алгоритма	879
Шум и чрезмерно тщательная подгонка	880
Расширение области применения деревьев решений	883
18.4. Обучение ансамбля	884
18.5. Принципы функционирования алгоритмов обучения: теория	
вычислительного обучения	889
Оценка количества необходимых примеров	890
Обучение списков решений	892
Обсуждение полученных результатов	894
18.6. Резюме	895
Библиографические и исторические заметки	896
Упражнения	899

Глава 19. ПРИМЕНЕНИЕ ЗНАНИЙ В ОБУЧЕНИИ	902
---	------------

19.1. Логическая формулировка задачи обучения	902
Примеры и гипотезы	903
Поиск текущей наилучшей гипотезы	905
Поиск на основе оценки наименьшего вклада	908
19.2. Применение знаний в обучении	913
Некоторые простые примеры	914
Некоторые общие схемы	915
19.3. Обучение на основе объяснения	917
Извлечение общих правил из примеров	919
Повышение эффективности правила	921
19.4. Обучение с использованием информации о релевантности	923
Определение пространства гипотез	923
Обучение и использование информации о релевантности	924
19.5. Индуктивное логическое программирование	927
Практический пример	928
Нисходящие методы индуктивного обучения	931

Индуктивное обучение с помощью обратной дедукции	934
Совершение открытий с помощью индуктивного логического программирования	937
19.6. Резюме	939
Библиографические и исторические заметки	940
Упражнения	943
 ГЛАВА 20. СТАТИСТИЧЕСКИЕ МЕТОДЫ ОБУЧЕНИЯ	 945
20.1. Статистическое обучение	946
20.2. Обучение с помощью полных данных	950
Обучение параметрам с помощью метода максимального правдоподобия: дискретные модели	950
Наивные байесовские модели	953
Обучение параметрам с максимальным правдоподобием: непрерывные модели	954
Обучение байесовским параметрам	956
Определение путем обучения структур байесовских сетей	959
20.3. Обучение с помощью скрытых переменных: алгоритм ЕМ	961
Неконтролируемая кластеризация: определение в процессе обучения смешанных гауссовых распределений	962
Обучение байесовских сетей со скрытыми переменными	966
Обучение скрытых марковских моделей	969
Общая форма алгоритма ЕМ	970
Определение с помощью обучения структур байесовских сетей со скрытыми переменными	971
20.4. Обучение на основе экземпляра	972
Модели ближайшего соседа	973
Ядерные модели	975
20.5. Нейронные сети	976
Элементы в нейронных сетях	977
Структуры сетей	979
Однослойные нейронные сети с прямым распространением (персептроны)	980
Многослойные нейронные сети с прямым распространением	985
Определение в процессе обучения структур нейронных сетей	990
20.6. Ядерные машины	991
20.7. Практический пример: распознавание рукописных цифр	995
20.8. Резюме	998
Библиографические и исторические заметки	1000
Упражнения	1005
 ГЛАВА 21. ОБУЧЕНИЕ С ПОДКРЕПЛЕНИЕМ	 1010
21.1. Введение	1010
21.2. Пассивное обучение с подкреплением	1012
Непосредственная оценка полезности	1014
Адаптивное динамическое программирование	1015

Обучение с учетом временной разницы	1016
21.3. Активное обучение с подкреплением	1020
Исследование среды	1021
Определение функции “действие—стоимость” с помощью обучения	1025
21.4. Обобщение в обучении с подкреплением	1027
Приложения методов обучения к ведению игр	1031
Применение к управлению роботами	1032
21.5. Поиск стратегии	1033
21.6. Резюме	1037
Библиографические и исторические заметки	1039
Упражнения	1042

Часть VII. ОБЩЕНИЕ, ВОСПРИЯТИЕ И ОСУЩЕСТВЛЕНИЕ ДЕЙСТВИЙ 1045

ГЛАВА 22. ОБЩЕНИЕ	1046
22.1. Общение как действие	1047
Основные понятия языка	1048
Составные этапы общения	1050
22.2. Формальная грамматика для подмножества английского языка	1054
Словарь языка \mathcal{E}_0	1054
Грамматика языка \mathcal{E}_0	1055
22.3. Синтаксический анализ (синтаксический разбор)	1056
Эффективный синтаксический анализ	1058
22.4. Расширенные грамматики	1065
Субкатегоризация глагола	1068
Порождающая мощь расширенных грамматик	1071
22.5. Семантическая интерпретация	1071
Семантика небольшой части английского языка	1072
Время события и времена глаголов	1074
Введение кванторов	1075
Прагматическая интерпретация	1078
Применение грамматик DCG для производства языковых конструкций	1079
22.6. Неоднозначность и устранение неоднозначности	1080
Устранение неоднозначности	1083
22.7. Понимание речи	1085
Разрешение ссылок	1085
Структура связной речи	1087
22.8. Индуктивный вывод грамматики	1089
22.9. Резюме	1092
Библиографические и исторические заметки	1093
Упражнения	1097

ГЛАВА 23. ВЕРОЯТНОСТНАЯ ОБРАБОТКА ЛИНГВИСТИЧЕСКОЙ ИНФОРМАЦИИ 1102

23.1. Вероятностные языковые модели	1103
Вероятностные контекстно-свободные грамматики	1106

Определение с помощью обучения вероятностей для грамматики PCFG	1108
Определение с помощью обучения структуры правил для грамматики PCFG	1110
23.2. Информационный поиск	1110
Сравнительный анализ систем информационного поиска	1114
Совершенствование информационного поиска	1115
Способы представления результирующих наборов	1117
Создание систем информационного поиска	1119
23.3. Извлечение информации	1121
23.4. Машинный перевод	1124
Системы машинного перевода	1127
Статистический машинный перевод	1127
Определение с помощью обучения вероятностей для машинного перевода	1132
23.5. Резюме	1134
Библиографические и исторические заметки	1135
Упражнения	1138
 ГЛАВА 24. ВОСПРИЯТИЕ	 1141
24.1. Введение	1141
24.2. Формирование изображения	1143
Получение изображения без линз — камера-обскура	1144
Системы линз	1145
Свет: фотометрия формирования изображения	1146
Цвет — спектрофотометрия формирования изображения	1147
24.3. Операции, выполняемые на первом этапе обработки изображения	1148
Обнаружение краев	1150
Сегментация изображения	1153
24.4. Извлечение трехмерной информации	1154
Движение	1156
Бинокулярные стереоданные	1158
Градиенты текстуры	1161
Затенение	1162
Контуры	1164
24.5. Распознавание объектов	1168
Распознавание с учетом яркости	1171
Распознавание с учетом характеристик	1172
Оценка позы	1175
24.6. Использование системы машинного зрения для манипулирования и передвижения	1177
24.7. Резюме	1180
Библиографические и исторические заметки	1181
Упражнения	1184
 ГЛАВА 25. РОБОТОТЕХНИКА	 1188
25.1. Введение	1188

25.2. Аппаратное обеспечение роботов	1190
Датчики	1190
Исполнительные механизмы	1192
25.3. Восприятие, осуществляемое роботами	1195
Локализация	1197
Составление карты	1203
Другие типы восприятия	1206
25.4. Планирование движений	1207
Пространство конфигураций	1207
Методы декомпозиции ячеек	1210
Методы скелетирования	1214
25.5. Планирование движений в условиях неопределенности	1215
Надежные методы	1217
25.6. Осуществление движений	1220
Динамика и управление	1220
Управление на основе поля потенциалов	1223
Реактивное управление	1225
25.7. Архитектуры робототехнического программного обеспечения	1227
Обобщающая архитектура	1227
Трехуровневая архитектура	1229
Робототехнические языки программирования	1230
25.8. Прикладные области	1231
25.9. Резюме	1235
Библиографические и исторические заметки	1237
Упражнения	1241

Часть VIII. Заключение 1247

Глава 26. Философские основания 1248

26.1. Слабый искусственный интеллект: могут ли машины действовать интеллектуально?	1249
Довод, исходящий из неспособности	1250
Возражения, основанные на принципах математики	1251
Довод, исходящий из неформализуемости	1253
26.2. Сильный искусственный интеллект: могут ли машины по-настоящему мыслить?	1255
Проблема разума и тела	1258
Эксперимент “мозг в колбе”	1260
Эксперимент с протезом мозга	1261
Китайская комната	1263
26.3. Этические и моральные последствия разработки искусственного интеллекта	1266
26.4. Резюме	1271
Библиографические и исторические заметки	1272
Упражнения	1275

ГЛАВА 27. НАСТОЯЩЕЕ И БУДУЩЕЕ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА	1277
27.1. Компоненты агента	1278
27.2. Архитектуры агентов	1281
27.3. Оценка правильности выбранного направления	1283
27.4. Перспективы развития искусственного интеллекта	1285
 ПРИЛОЖЕНИЕ А. МАТЕМАТИЧЕСКИЕ ОСНОВЫ	 1288
А.1. Анализ сложности и система обозначений $O()$	1288
Асимптотический анализ	1288
Изначально сложные и недетерминированные полиномиальные задачи	1290
А.2. Векторы, матрицы и линейная алгебра	1291
А.3. Распределения вероятностей	1293
Библиографические и исторические заметки	1295
 ПРИЛОЖЕНИЕ Б. ОБЩИЕ СВЕДЕНИЯ О ЯЗЫКАХ И АЛГОРИТМАХ, ИСПОЛЬЗУЕМЫХ В КНИГЕ	 1297
Б.1. Определение языков с помощью формы Бэкуса–Наура	1297
Б.2. Описание алгоритмов с помощью псевдокода	1298
Б.3. Оперативная помощь	1299
 ЛИТЕРАТУРА	 1302
 ПРЕДМЕТНЫЙ УКАЗАТЕЛЬ	 1373

Посвящается Лой, Гордону и Люси

С.Дж. Рассел

Посвящается Крису, Изабелле и Джульетте

П. Норвиг

ПРЕДИСЛОВИЕ

Искусственный интеллект (ИИ) — широкая область знаний; именно поэтому данная книга имеет такой большой объем. Авторы попытались достаточно полно описать теоретические основы искусственного интеллекта, включая математическую логику, теорию вероятностей и теорию непрерывных функций, раскрыть суть таких понятий, как восприятие, рассуждение, обучение и действие, а также описать все технические средства, созданные в рамках этого научного направления, начиная с микроэлектронных устройств и заканчивая межпланетными автоматическими зондами. Большой объем данной книги обусловлен также тем, что авторы стремились представить достигнутые результаты достаточно глубоко, хотя в основной части каждой главы они старались охватить только самые важные идеи, касающиеся рассматриваемой темы. Указания, позволяющие получить более полные сведения, приведены в библиографических заметках в конце каждой главы.

Эта книга имеет подзаголовок “Современный подход”. С помощью этого довольно-таки малосодержательного названия авторы хотели подчеркнуть, что пытались представить в ней в рамках единого способа изложения все современные достижения в области искусственного интеллекта, а не описать каждое отдельное его направление в его собственном историческом контексте. Авторы приносят свои извинения представителям тех направлений, которые стали выглядеть не столь значимыми, как они заслуживают, лишь из-за того, что для их описания принят такой подход.

Главной объединяющей темой этой книги является идея *интеллектуального агента*. Авторы определяют *искусственный интеллект* как науку об агентах, которые получают результаты актов восприятия из своей среды и выполняют действия, причем каждый такой агент реализует функцию, которая отображает последовательности актов восприятия в действия. В данной книге рассматриваются различные способы представления этих функций, в частности продукционные системы, реактивные агенты, условные планировщики в реальном масштабе времени, нейронные сети и системы, действующие на основе теории решений. Авторы трактуют роль *обучения* как распространения сферы деятельности проектировщика на неизвестную среду и показывают, какие ограничения налагает указанный подход к обучению на проект агента, способствуя применению явного представления знаний и таких же способов формирования рассуждений. Кроме того, авторы рассматривают робототехнику и системы технического зрения не как независимо определяемые научные направления, а как области знаний, позволяющие обеспечить более успешное достижение целей, стоящих перед агентами, и подчеркивают важность учета того, в какой среде агент решает поставленные перед ним задачи, при определении соответствующего проекта агента.

Основная цель авторов состояла в том, чтобы изложить идеи, которые были сформулированы в исследованиях по искусственному интеллекту, проводившихся в течение последних пятидесяти лет, а также собраны на протяжении последних двух тысячелетий в тех областях знаний, которые стали стимулом к развитию искусственного интеллекта. Мы старались избегать чрезмерного формализма при изложении этих идей, сохраняя при этом необходимую точность. При любой возможности мы приводили алгоритмы на псевдокоде, чтобы конкретизировать излагаемые идеи;

краткие сведения о применяемом нами псевдокоде содержатся в приложении Б. Реализации этих алгоритмов на нескольких языках программирования можно найти на сопровождающем Web-узле книги (aima.cs.berkeley.edu).

Настоящая книга прежде всего предназначена для использования в базовом университетском курсе или в последовательности курсов. Она может также использоваться в курсе по специальности (возможно, с добавлением материала из некоторых основных источников, предложенных в библиографических заметках). Кроме того, данная книга характеризуется всесторонним охватом тематики и большим количеством подробных алгоритмов, поэтому применима в качестве основного справочника для аспирантов, специализирующихся в области искусственного интеллекта, а также будет небезынтесна профессионалам, желающим выйти за пределы избранной ими специальности. При этом единственным требованием является знакомство с основными понятиями информатики (алгоритмы, структуры данных, классы сложности) на уровне студента-второкурсника. Для понимания материала по нейронным сетям и подробного ознакомления со сведениями о статистическом обучении полезно освоить исчисление на уровне студента первого курса. Часть необходимых сведений из области математики приведена в приложении А.

Краткий обзор книги

Данная книга разделена на восемь частей. В части I, “Искусственный интеллект”, предлагается общий обзор тематики искусственного интеллекта, базирующейся на идее интеллектуального агента — системы, которая способна принять решение о том, что делать, а затем выполнить это решение. В части II, “Решение проблем”, изложение сосредоточено на методах принятия решений по выбору оптимальных действий в тех условиях, когда необходимо продумывать наперед несколько этапов, например при поиске маршрута проезда через всю страну или при игре в шахматы. В части III, “Знания и рассуждения”, обсуждаются способы представления знаний о мире — как он функционирует, каковы его основные особенности в настоящее время и к чему могут привести те или иные действия, а также способы формирования логических рассуждений на основе этих знаний. В части IV, “Планирование”, описывается, как использовать эти способы формирования рассуждений для принятия решений по выбору дальнейших действий, особенно при составлении планов. Часть V, “Неопределенные знания и рассуждения в условиях неопределенности”, аналогична частям III и IV, но в ней изложение в основном сосредоточивается на способах формирования рассуждений и принятия решений в условиях неопределенности знаний о мире, с чем обычно приходится сталкиваться, например, в системах медицинской диагностики и лечения.

Части II–V, вместе взятые, содержат описание тех компонентов интеллектуального агента, которые отвечают за выработку решений. В части VI, “Обучение”, описаны методы выработки знаний, необходимых для этих компонентов, которые обеспечивают принятие решений. В части VII, “Общение, восприятие и осуществление действий”, описаны способы, с помощью которых интеллектуальные агенты могут получать результаты восприятия из своей среды, чтобы узнать, что в ней происходит, либо с помощью систем технического зрения, осязания, слуха, либо на основе понимания языка, а также способы, с помощью которых интеллектуальные агенты могут претворять свои планы в реальные действия, такие как выполнение движений

робота или произнесение фрагментов речи на естественном языке. Наконец, в части VIII, “Заключение”, анализируется прошлое и будущее искусственного интеллекта и рассматриваются философские и этические последствия его развития.

Отличия от первого издания

Со времени публикации первого издания этой книги в 1995 году в искусственном интеллекте многое изменилось, поэтому внесены значительные изменения и в саму книгу. Каждая глава была в значительной степени переработана, чтобы в ней можно было отразить новейшие достижения в рассматриваемой области, дать иное толкование старым работам с той точки зрения, которая более согласована с новыми результатами, а также улучшить качество изложения рассматриваемых идей в соответствии с принципами педагогики. Активных пользователей методов искусственного интеллекта должно порадовать то, что представленные в настоящем издании методы стали намного более эффективными по сравнению с теми, которые описывались в издании 1995 года; например, алгоритмы планирования, которые рассматривались в первом издании, позволяли формировать планы, состоящие всего лишь из нескольких шагов, тогда как масштабы применения алгоритмов, описанные в настоящем издании, увеличились до десятков тысяч шагов. Подобные усовершенствования, измеряемые несколькими порядками величин, достигнуты и в областях вероятностного вывода и обработки лингвистической информации, а также в других вспомогательных областях. Наиболее существенные изменения, внесенные во второе издание, описаны ниже.




- В части I изложены факты, которые свидетельствуют о признании исторического вклада в развитие искусственного интеллекта со стороны теории управления, теории игр, экономики и неврологии. Это позволяет создать основу для более целостного описания идей, заимствованных из этих научных областей, в последующих главах.
- В части II описаны алгоритмы оперативного поиска и введена новая глава по удовлетворению ограничений, которая позволяет установить естественную связь между вычислительными методами и приведенными в данной книге материалами по логике.
- Теперь в части III пропозициональная логика, которая в первом издании была рекомендована читателям как промежуточная ступенька на пути к логике первого порядка, рассматривается как полезный сам по себе язык представления, для которого предусмотрены быстродействующие алгоритмы логического вывода и эффективные проекты агентов на основе схемы. Главы по логике первого порядка были реорганизованы для более наглядного изложения материала, а в качестве примера проблемной области приведено описание процесса осуществления покупок в Internet.
- В части IV приведены сведения о более новых методах планирования, таких как Graphplan и планирование на основе выполнимости. Кроме того, увеличен объем изложения, касающегося составления расписаний, условного планирования, иерархического планирования и мультиагентского планирования.
- В часть V включен дополнительный материал по байесовским сетям, в котором описаны новые алгоритмы, в частности алгоритмы устранения перемен-

ных и алгоритмы Монте-Карло на основе марковской цепи, а также введена новая глава по формированию неопределенных рассуждений с учетом времени и созданию покрытий скрытых марковских моделей, а также по применению фильтров Калмана и динамических байесовских сетей. Описание марковских процессов принятия решений стало еще более глубоким; введены новые разделы по теории игр и проектированию механизма.

- В части VI связаны воедино все результаты, достигнутые в области статистического и символического обучения, а также обучения нейронных сетей; кроме того, введены разделы, содержащие сведения об увеличении производительности алгоритмов, алгоритме ЕМ, обучении на основе экземпляра и о ядерных методах (о машинах поддерживающих векторов).
- В части VII к общему объему материала об обработке лингвистической информации добавлены разделы, касающиеся обработки речи и индуктивного вывода грамматики, а также глава по вероятностным языковым моделям, с учетом того, что областью применения этих сведений должны стать информационный поиск и машинный перевод. В ходе изложения вопросов робототехники подчеркнута необходимость применения методов обработки неопределенных сенсорных данных, а в главе по системам технического зрения приведены уточненные сведения по распознаванию объектов.
- В части VIII предусмотрен дополнительный раздел, касающийся этических последствий развития искусственного интеллекта.

Как использовать эту книгу

Книга состоит из 27 глав, причем для изучения каждой из них требуется примерно недельный объем лекций. Таким образом, для учебной проработки всей книги требуется последовательность курсов лекций, рассчитанная на два семестра. Еще один вариант состоит в том, что может быть составлен выборочный курс, удовлетворяющий интересы преподавателя и студента. Благодаря тому что в ней охвачена широкая тематика, эта книга может использоваться в качестве основы для многих курсов, начиная с коротких, вводных циклов лекций для начинающих и заканчивая специализированными курсами с углубленным изучением избранной темы для студентов последних лет обучения. На Web-узле, находящемся по адресу aima.cs.berkeley.edu, приведены программы курсов лекций, проводимых более чем в 600 университетах и колледжах, в основу которых было положено первое издание настоящей книги, а также даны рекомендации, позволяющие читателю найти программу курсов лекций, в наибольшей степени соответствующую его потребностям.

Книга включает 385 упражнений. Упражнения, требующие существенного объема программирования, отмечены значком в виде клавиатуры (). Проще всего эти упражнения можно выполнить, воспользовавшись архивом кода, который находится по адресу aima.cs.berkeley.edu. Некоторые из упражнений настолько велики, что их можно рассматривать как проекты с заданными сроками. Многие упражнения требуют проведения определенных исследований с помощью доступной литературы; они отмечены значком в виде книги (). Важные примечания отмечены значком в виде «указующего перста» () и выделены курсивным шрифтом. В книгу

включен обширный предметный указатель, состоящий из нескольких тысяч элементов, который поможет читателю найти нужную тему. Кроме того, значком с изображением руки, держащей карандаш (✍), и полужирным шрифтом отмечаются все новые термины, везде, где впервые приведено их определение.

Использование Web-узла

На Web-узле `aima.cs.berkeley.edu` приведено следующее:

- реализации алгоритмов, описанных в книге, на нескольких языках программирования;
- список более чем 600 учебных заведений, в которых используется данная книга, сопровождающийся многочисленными ссылками на материалы курсов, доступные в оперативном режиме;
- аннотированный список более чем 800 ссылок на Web-узлы с полезными сведениями по искусственному интеллекту;
- списки дополнительных материалов и ссылок, относящихся к каждой главе;
- инструкции с описанием того, как присоединяться к дискуссионной группе, посвященной данной книге;
- инструкции с описанием того, как обратиться к авторам, чтобы передать им свои вопросы или комментарии;
- инструкции с описанием того, как сообщить об ошибках, обнаруженных в книге;
- копии рисунков из оригинала книги, а также слайды и другие материалы для преподавателей.

Благодарности

Основная часть главы 24 (по системам технического зрения) написана Джитендрой Маликом (Jitendra Malik). Глава 25 (по робототехнике) в основном написана Себастьяном Траном (Sebastian Thrun) для настоящего издания и Джоном Кэнни (John Caplu) для первого издания. Дуг Эдвардс (Doug Edwards) провел исследование, на основании которого написаны исторические заметки для первого издания. Тим Хуанг (Tim Huang), Марк Паскин (Mark Paskin) и Синтия Бруинс (Cynthia Bruyns) оказали помощь при оформлении диаграмм и алгоритмов. Алан Апт (Alan Apt), Сондра Чавес (Sondra Chavez), Тони Хом (Toni Holm), Джейк Вард (Jake Warde), Ирвин Zucker (Irwin Zucker) и Камилла Трантакост (Camille Trentacoste), сотрудники издательства Prentice Hall, приложили большие усилия, чтобы помочь нам соблюсти намеченный график подготовки книги, и внесли много полезных предложений по оформлению и содержанию книги.

Стюарт хотел бы поблагодарить своих родителей за их постоянную помощь и поддержку, а также свою жену, Лой Шефлотт (Loy Sheflott), за ее бесконечное терпение и безграничную мудрость. Он надеется, что скоро эту книгу прочитают Гордон и Люси. Исключительно полезной для него была работа с RUGS (Russell's Unusual Group of Students — необыкновенная группа студентов Рассела).

Питер хотел бы поблагодарить своих родителей, Торстена и Герду, за то, что они очень помогли ему на первых порах, и свою жену Крис, детей и друзей за то, что

подбадривали его и терпели его отсутствие в течение тех долгих часов, когда он писал эту книгу, и тех еще более долгих часов, когда он снова ее переписывал.

Мы очень обязаны библиотекарям, работающим в университете г. Беркли, Станфордском университете, Массачусеттском технологическом институте и агентстве NASA, а также разработчикам узлов CiteSeer и Google, которые внесли революционные изменения в сам способ проведения исследований.

Мы буквально не с состоянием выразить свою признательность всем тем, кто использовал данную книгу и внес свои предложения, но хотели бы поблагодарить за особо полезные комментарии следующих: Кшиштофа Апта (Krzysztof Apt), Эллери Эзиела (Ellery Aziel), Джефа Ван Баалена (Jeff Van Baalen), Брайена Бейкера (Brian Baker), Дона Баркера (Don Barker), Тони Баррета (Tony Barrett), Джеймса Ньютона Баса (James Newton Bass), Дона Била (Don Beal), Говарда Бека (Howard Beck), Вольфганга Бибеля (Wolfgang Bibel), Джона Биндера (John Binder), Лэрри Букмана (Larry Bookman), Дэвида Р. Боксолла (David R. Boxall), Герхарда Бревку (Gerhard Brewka), Селмера Бринсйорда (Selmer Bringsjord), Карла Бродли (Carla Brodley), Криса Брауна (Chris Brown), Вильгельма Бургера (Wilhelm Burger), Лорен Берка (Lauren Burka), Жоао Кашпоро (Joao Cachopo), Меррея Кэмпбелла (Murray Campbell), Нормана Карвера (Norman Carver), Эммануэля Кастро (Emmanuel Castro), Анила Чакраварти (Anil Chakravarthy), Дэна Чизарика (Dan Chisarick), Роберто Сиполлу (Roberto Cipolla), Дэвида Коэна (David Cohen), Джеймса Коулмэна (James Coleman), Джули Энн Компарини (Julie Ann Comparini), Гэри Коттрелла (Gary Cottrell), Эрнеста Дэвиса (Ernest Davis), Рину Дехтер (Rina Dechter), Тома Диттерика (Tom Dietterich), Чака Дийера (Chuck Dyer), Барбару Энгельхардт (Barbara Engelhardt), Дуга Эдвардса (Doug Edwards), Кутлухана Эрола (Kutluhan Erol), Орена Этциони (Oren Etzioni), Хану Филипа (Hana Filip), Дугласа Фишера (Douglas Fisher), Джефффри Форбса (Jeffrey Forbes), Кена Форда (Ken Ford), Джона Фослера (John Fosler), Алекса Франца (Alex Franz), Боба Фатрелла (Bob Futrelle), Марека Галецки (Marek Galecki), Штефана Гербердинга (Stefan Gerberding), Стюарта Джилла (Stuart Gill), Сабину Глеснер (Sabine Glesner), Сета Голуба (Seth Golub), Госту Гранье (Gosta Grahne), Расса Грейнера (Russ Greiner), Эрика Гримсона (Eric Grimson), Барбару Грош (Barbara Grosz), Лэрри Холла (Larry Hall), Стива Хэнкса (Steve Hanks), Отара Хэнссона (Othar Hansson), Эрнста Хайнца (Ernst Heinz), Джима Эндлера (Jim Hendler), Кристофа Херманна (Christoph Herrmann), Вазанта Хонавара (Vasant Honavar), Тима Хуанга (Tim Huang), Сета Хатчинсона (Seth Hutchinson), Джуста Джейкоба (Joost Jacob), Магнуса Йоханссона (Magnus Johansson), Дэна Джурафски (Dan Jurafsky), Лесли Кэлблинга (Leslie Kaelbling), Кейдзи Канадзава (Keiji Kanazawa), Сурекха Касибхатла (Surekha Kasibhatla), Саймона Казифа (Simon Kasif), Генри Каутца (Henry Kautz), Гернота Кершбаумера (Gernot Kerschbaumer), Ричарда Кирби (Richard Kirby), Кевина Найта (Kevin Knight), Свена Кёнига (Sven Koenig), Дафну Коллер (Daphne Koller), Рича Корфа (Rich Korf), Джеймса Керина (James Kurien), Джона Лафферти (John Lafferty), Гуса Ларссона (Gus Larsson), Джона Лаццаро (John Lazzaro), Джона Лебланка (Jon LeBlanc), Джейсона Литермана (Jason Leatherman), Фрэнка Ли (Frank Lee), Эдварда Лима (Edward Lim), Пьера Луво (Pierre Louveaux), Дона Лавленда (Don Loveland), Сридхара Махадевана (Sridhar Mahadevan), Джима Мартина (Jim Martin), Энди Мейера (Andy Mayer), Дэвида Мак-Грэйна (David McGrane), Джей Менделсон (Jay Mendelsohn), Брайена Милча (Brian Milch), Стива Майнтона (Steve Minton), Вибху Миттала (Vibhu Mittal), Леору Моргенстерн (Leora Morgenstern), Стивена Мугглтона (Stephen Muggleton), Кевина Мэрфи (Kevin Murphy), Рона Мьюзика (Ron Musick), Санга Миаэнга (Sung Myaeng), Ли Нэйша (Lee Naish), Панду Найака (Pandu

Nayak), Бернхарда Небеля (Bernhard Nebel), Стюарта Нельсона (Stuart Nelson), Шуан-лонг Нгуэн (XuanLong Nguyen), Иллаха Нурбакша (Illah Nourbakhsh), Стива Омохандро (Steve Omohundro), Дэвида Пейджа (David Page), Дэвида Палмера (David Palmer), Дэвида Паркса (David Parkes), Рона Парра (Ron Parr), Марка Паскина (Mark Paskin), Тони Пасера (Tony Passera), Майкла Паззани (Michael Pazzani), Вима Пейлса (Wim Pijls), Иру Пол (Ira Pohl), Марту Поллак (Martha Pollack), Дэвида Пула (David Poole), Брюса Портера (Bruce Porter), Малкома Прадхана (Malcolm Pradhan), Билла Прингла (Bill Pringle), Лоррэн Прайор (Lorraine Prior), Грэга Прована (Greg Provan), Уильяма Рапапорта (William Rapaport), Филипа Ресника (Philip Resnik), Франческу Росси (Francesca Rossi), Джонатана Шеффера (Jonathan Schaeffer), Ричарда Шерла (Richard Scherl), Ларса Шустера (Lars Schuster), Сохейль Шамс (Soheil Shams), Стюарта Шапиро (Stuart Shapiro), Джюд Шавлик (Jude Shavlik), Сатиндера Сингха (Satinder Singh), Дэниела Слитора (Daniel Sleator), Дэвида Смита (David Smith), Брайена Соу (Bryan So), Роберта Спрула (Robert Sproull), Линн Стейн (Lynn Stein), Лэрри Стивенса (Larry Stephens), Андреаса Штолке (Andreas Stolcke), Пола Страдлинга (Paul Stradling), Девику Субраманиан (Devika Subramanian), Рича Саттона (Rich Sutton), Джонатана Тэша (Jonathan Tash), Остина Тэйта (Austin Tate), Майкла Тилшера (Michael Thielscher), Уильяма Томпсона (William Thompson), Себастьяна Трана (Sebastian Thrun), Эрика Тидеманна (Eric Tiedemann), Марка Торранса (Mark Torgance), Рэндалла Уфама (Randall Upham), Пола Утгоффа (Paul Utgoff), Питера ван Бека (Peter van Beek), Хала Вариана (Hal Varian), Сунила Вемури (Sunil Vemuri), Джима Уолдо (Jim Waldo), Бонни Веббер (Bonnie Webber), Дэна Вэлда (Dan Weld), Майкла Веллмана (Michael Wellman), Майкла Дина Уайта (Michael Dean White), Камина Уайтхауза (Kamin Whitehouse), Брайена Уильямса (Brian Williams), Дэвида Уолфа (David Wolfe), Билла Вудса (Bill Woods), Олдена Райта (Alden Wright), Ричарда Йэна (Richard Yen), Вейшионг Джанг (Weixiong Zhang), Шломо Зильберштейна (Shlomo Zilberstein), а также анонимных рецензентов, привлеченных издательством Prentice Hall.

Об обложке

Изображение на обложке было спроектировано авторами и выполнено Лайзой Мэри Сарденья (Lisa Marie Sardegna) и Мэриэнн Симмонс (Maryann Simmons) с использованием программ SGI Inventor™ и Adobe Photoshop™. На обложке показаны перечисленные ниже предметы, иллюстрирующие историю искусственного интеллекта.

1. Алгоритм планирования Аристотеля из книги *De Motu Animalium* (ок. 400 до н.э.).
2. Генератор понятий Раймунда Луллия из книги *Ars Magna* (ок. 1300).
3. Разностная машина Чарльза Бэббиджа, прототип первого универсального компьютера (1848).
4. Система обозначений Готтлоба Фреге для логики первого порядка (1789).
5. Диаграммы Льюиса Кэрролла для формирования логических рассуждений (1886).
6. Система обозначений вероятностной сети Сьюэлла Райта (1921).
7. Алан Тьюринг (1912–1954).
8. Робот Shakey (1969–1973).
9. Современная диагностическая экспертная система (1993).

ОБ АВТОРАХ

Стюарт Рассел родился в 1962 году в г. Портсмут, Англия. Он получил степень бакалавра искусств по физике с наградами первой степени в Оксфордском университете в 1982 году и степень доктора философии по информатике в Станфордском университете в 1986 году. Затем он перешел на факультет Калифорнийского университета в г. Беркли, где занял должность профессора компьютерных наук, директора предприятия Center for Intelligent Systems и заведующего кафедрой инженерного искусства Смита—Задэ. В 1990 году он получил от фонда National Science Foundation премию Presidential Young Investigator Award, а в 1995 разделил первое место в конкурсе Computers and Thought Award. В 1996 году он победил на конкурсе Miller Research Professor в Калифорнийском университете, а в 2000 году был выдвинут на премию Chancellor's Professorship. В 1998 году он прочитал мемориальные лекции в память Форсита в Станфордском университете. Рассел — член AAAI (American Association for Artificial Intelligence — Американская ассоциация специалистов по искусственному интеллекту) и бывший член исполнительного совета этой ассоциации. Им опубликовано больше 100 статей по широкому кругу проблем искусственного интеллекта. Он также является автором книг *Use of Knowledge in Analogy and Induction* и *Do the Right Thing: Studies in Limited Rationality* (написана в соавторстве с Эриком Вефолдом — Eric Wefald).

Питер Норvig — директор подразделения Search Quality компании Google, а также член AAAI и член исполнительного совета этой ассоциации. Перед этим он возглавлял подразделение Computational Sciences Division в исследовательском центре NASA Ames Research Center, а также руководил проводимыми в комитете NASA (National Aeronautics and Space Administration — Национальный комитет по авиации и исследованию космического пространства) исследованиями и разработками по искусственному интеллекту и робототехнике. Еще раньше он занимал должность руководителя исследовательских работ в компании Junglee, где под его руководством была разработана одна из первых служб извлечения информации в Internet, и должность старшего научного сотрудника в подразделении компании Sun Microsystems Laboratories, которое работало над интеллектуальным информационным поиском. Он получил степень бакалавра наук по прикладной математике в Университете Брауна и степень доктора философии по информатике в Калифорнийском университете, г. Беркли. Он занимал должность профессора в Университете Южной Калифорнии и работал на факультете научно-технических исследований в Беркли. Ему принадлежат больше 50 публикаций по компьютерным наукам, в том числе книги *Case Studies in Common Lisp*, *Verbmobil: A Translation System for Face-to-Face Dialog* и *Intelligent Help Systems for UNIX*.

ЖДЕМ ВАШИХ ОТЗЫВОВ!

Вы, уважаемый читатель, и есть главный критик и комментатор этой книги. Мы ценим ваше мнение и хотим знать, что было сделано нами правильно, что можно было сделать лучше и что еще вы хотели бы увидеть изданным нами. Нам интересно услышать и любые другие замечания, которые вам хотелось бы высказать в наш адрес.

Мы ждем ваших комментариев и надеемся на них. Вы можете прислать нам бумажное или электронное письмо либо просто посетить наш Web-сервер и оставить свои замечания. Одним словом, любым удобным для вас способом дайте нам знать, нравится или нет вам эта книга, а также выскажите свое мнение о том, как сделать наши книги более интересными для вас.

Посылая письмо или сообщение, не забудьте указать название книги и ее авторов, а также ваш обратный адрес. Мы внимательно ознакомимся с вашим мнением и обязательно учтем его при отборе и подготовке к изданию последующих книг. Наши координаты:

E-mail: info@williamspublishing.com

WWW: <http://www.williamspublishing.com>

Информация для писем:

из России: 115419, Москва, а/я 783

из Украины: 03150, Киев, а/я 152

Часть I

ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ

Введение	34
Интеллектуальные агенты	75

1 ВВЕДЕНИЕ

В этой главе авторы пытаются объяснить, почему они рассматривают искусственный интеллект как тему, в наибольшей степени заслуживающую изучения, а также определить, в чем именно заключается данная тема; эти задачи необходимо решить, прежде чем приступить к дальнейшей работе.

Люди называют себя *Homo sapiens* (человек разумный), поскольку для них мыслительные способности имеют очень важное значение. В течение тысяч лет человек пытается понять, как он думает, т.е. разобраться в том, как именно ему, сравнительно небольшому материальному объекту, удастся ощущать, понимать, предсказывать и управлять миром, намного более значительным по своим размерам и гораздо более сложным по сравнению с ним. В области **искусственного интеллекта (ИИ)** решается еще более ответственная задача: специалисты в этой области пытаются не только понять природу интеллекта, но и создать интеллектуальные сущности.

Искусственный интеллект — это одна из новейших областей науки. Первые работы в этой области начались вскоре после Второй мировой войны, а само ее название было предложено в 1956 году. Ученые других специальностей чаще всего указывают искусственный интеллект, наряду с молекулярной биологией, как “область, в которой я больше всего хотел бы работать”. Студенты-физики вполне обоснованно считают, что все великие открытия в их области уже были сделаны Галилеем, Ньютоном, Эйнштейном и другими учеными. Искусственный интеллект, с другой стороны, все еще открывает возможности для проявления талантов нескольких настоящих Эйнштейнов.

В настоящее время тематика искусственного интеллекта охватывает огромный перечень научных направлений, начиная с таких задач общего характера, как обучение и восприятие, и заканчивая такими специальными задачами, как игра в шахматы, доказательство математических теорем, сочинение поэтических произведений и диагностика заболеваний. В искусственном интеллекте систематизируются и автоматизируются интеллектуальные задачи и поэтому эта область касается любой сферы интеллектуальной деятельности человека. В этом смысле искусственный интеллект является поистине универсальной научной областью.

1.1. ОБЩЕЕ ОПРЕДЕЛЕНИЕ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Из сказанного выше можно сделать вывод, что искусственный интеллект представляет собой чрезвычайно интересную научную область. Но определение этого научного направления в настоящей книге еще не было дано. В табл. 1.1 приведены определения искусственного интеллекта, взятые из восьми научных работ. Эти определения можно классифицировать по двум основным категориям. Грубо говоря, формулировки, приведенные в верхней части таблицы, касаются мыслительных процессов и способов рассуждения, а в нижней части таблицы находятся формулировки, касающиеся поведения. В определениях, приведенных слева, успех измеряется в терминах достоверного воспроизведения способностей человека, а формулировки, находящиеся справа, характеризуют конечные достижения в той области трактовки идеальной концепции интеллектуальности, которую авторы настоящей книги предпочитают называть **рациональностью**. Система является рациональной, если она “все действия выполняет правильно”, при условии, что система обладает знаниями о том, что является правильным.

Таблица 1.1. Некоторые определения искусственного интеллекта, распределенные по четырем категориям	
Системы, которые думают подобно людям	Системы, которые думают рационально
“Новое захватывающее направление работ по созданию компьютеров, способных думать, ...машин, обладающих разумом, в полном и буквальном смысле этого слова” [631]	“Изучение умственных способностей с помощью вычислительных моделей” [239]
“[Автоматизация] действий, которые мы ассоциируем с человеческим мышлением, т.е. таких действий, как принятие решений, решение задач, обучение...” [95]	“Изучение таких вычислений, которые позволяют чувствовать, рассуждать и действовать” [1603]
Системы, которые действуют подобно людям	Системы, которые действуют рационально
“Искусство создания машин, которые выполняют функции, требующие интеллектуальности при их выполнении людьми” [871]	“Вычислительный интеллект — это наука о проектировании интеллектуальных агентов” [1227]
“Наука о том, как научить компьютеры делать то, в чем люди в настоящее время их превосходят” [1285]	“Искусственный интеллект... — это наука, посвященная изучению интеллектуального поведения артефактов ¹ ” [1146]

История развития искусственного интеллекта показывает, что интенсивные исследования проводились по всем четырем направлениям. Вполне можно предположить, что между теми учеными, которые в основном исходят из способностей людей, и теми, кто занимается главным образом решением проблемы рациональности, существуют определенные разногласия². Подход, ориентированный на изучение человека, должен представлять собой эмпирическую научную область, развитие кото-

¹ Артефакт — искусственный объект.

² Необходимо указать, что авторы, проводя различие между человеческим и рациональным поведением, отнюдь не имеют в виду то, что люди обязательно действуют “нерационально” в том смысле этого слова, который характеризуется как “эмоциональная неустойчивость” или “неразумность”. Просто следует всегда помнить о том, что люди не идеальны, например, не все они становятся шахматными гроссмейстерами, даже если досконально изучили правила игры в шахматы, и, к сожалению, далеко не каждый получает высшие оценки на экзаменах. Некоторые систематические ошибки в человеческих рассуждениях были изучены и описаны в [762].

рой происходит по принципу выдвижения гипотез и их экспериментального подтверждения. С другой стороны, подход, основанный на понятии рациональности, представляет собой сочетание математики и техники. Каждые из этих групп ученых действуют разрозненно, но вместе с тем помогают друг другу. Ниже четыре указанных подхода рассматриваются более подробно.

Проверка того, способен ли компьютер действовать подобно человеку: подход, основанный на использовании теста Тьюринга

✎ **Тест Тьюринга**, предложенный Аланом Тьюрингом [1520], был разработан в качестве удовлетворительного функционального определения интеллекта. Тьюринг решил, что нет смысла разрабатывать обширный список требований, необходимых для создания искусственного интеллекта, который к тому же может оказаться противоречивым, и предложил тест, основанный на том, что поведение объекта, обладающего искусственным интеллектом, в конечном итоге нельзя будет отличить от поведения таких бесспорно интеллектуальных сущностей, как человеческие существа. Компьютер успешно пройдет этот тест, если человек-экспериментатор, задавший ему в письменном виде определенные вопросы, не сможет определить, получены ли письменные ответы от другого человека или от некоторого устройства. В главе 26 подробно обсуждается этот тест и рассматривается вопрос о том, действительно ли можно считать интеллектуальным компьютер, который успешно прошел подобный тест. На данный момент просто отметим, что решение задачи по составлению программы для компьютера для того, чтобы он прошел этот тест, требует большого объема работы. Запрограммированный таким образом компьютер должен обладать перечисленными ниже возможностями.

- Средства ✎ **обработки текстов на естественных языках** (Natural Language Processing —NLP), позволяющие успешно общаться с компьютером, скажем на английском языке.
- Средства ✎ **представления знаний**, с помощью которых компьютер может записать в память то, что он узнает или прочитает.
- Средства ✎ **автоматического формирования логических выводов**, обеспечивающие возможность использовать хранимую информацию для поиска ответов на вопросы и вывода новых заключений.
- Средства ✎ **машинного обучения**, которые позволяют приспосабливаться к новым обстоятельствам, а также обнаруживать и экстраполировать признаки стандартных ситуаций.

В тесте Тьюринга сознательно исключено непосредственное физическое взаимодействие экспериментатора и компьютера, поскольку для создания искусственного интеллекта не требуется физическая имитация человека. Но в так называемом ✎ **полном тесте Тьюринга** предусмотрено использование видеосигнала для того, чтобы экспериментатор мог проверить способности испытуемого объекта к восприятию, а также имел возможность представить физические объекты “в неполном виде” (пропустить их “через штриховку”). Чтобы пройти полный тест Тьюринга, компьютер должен обладать перечисленными ниже способностями.

- ✂ **Машинное зрение** для восприятия объектов.
- Средства ✂ **робототехники** для манипулирования объектами и перемещения в пространстве.

Шесть направлений исследований, перечисленных в данном разделе, составляют основную часть искусственного интеллекта, а Тьюринг заслуживает нашей благодарности за то, что предложил такой тест, который не потерял своей значимости и через 50 лет. Тем не менее исследователи искусственного интеллекта практически не занимаются решением задачи прохождения теста Тьюринга, считая, что гораздо важнее изучить основополагающие принципы интеллекта, чем продублировать одного из носителей естественного интеллекта. В частности, проблему “искусственного полета” удалось успешно решить лишь после того, как братья Райт и другие исследователи перестали имитировать птиц и приступили к изучению аэродинамики. В научных и технических работах по воздухоплаванию цель этой области знаний не определяется как “создание машин, которые в своем полете настолько напоминают голубей, что даже могут обмануть настоящих птиц”.

Как мыслить по-человечески: подход, основанный на когнитивном моделировании

Прежде чем утверждать, что какая-то конкретная программа мыслит, как человек, требуется иметь некоторый способ определения того, как же мыслят люди. Необходимо проникнуть в сам фактически происходящий процесс работы человеческого разума. Для этого могут использоваться два способа: интроспекция (попытка проследить за ходом собственных мыслей) и психологические эксперименты. Только после создания достаточно точной теории мышления появится возможность представить формулы этой теории в виде компьютерной программы. И если входные и выходные данные программы, а также распределение выполняемых ею действий во времени будут точно соответствовать поведению человека, это может свидетельствовать о том, что некоторые механизмы данной программы могут также действовать в человеческом мозгу. Например, Аллен Ньюэлл (Allen Newell) и Герберт Саймон (Herbert Simon), которые разработали программу GPS (“General Problem Solver” — универсальный решатель задач) [1129], не стремились лишь к тому, чтобы эта программа правильно решала поставленные задачи. Их в большей степени заботило, чтобы запись этапов проводимых ею рассуждений совпадала с регистрацией рассуждений людей, решающих такие же задачи. В междисциплинарной области ✂ **когнитологии** совместно используются компьютерные модели, взятые из искусственного интеллекта, и экспериментальные методы, взятые из психологии, для разработки точных и обоснованных теорий работы человеческого мозга.

Такая область знаний, как когнитология, является весьма увлекательной и настолько обширной, что ей вполне может быть посвящена отдельная энциклопедия [1599]. В данной книге авторы не пытаются описать все, что известно о человеческом познании. В ней лишь в некоторых местах комментируются аналогии или различия между методами искусственного интеллекта и человеческим познанием. Тем не менее настоящая научная когнитология обязательно должна быть основана на экспериментальном исследовании реальных людей или животных, а авторы данной книги предполагают, что ее читатель имеет доступ для экспериментирования только к компьютеру.

На начальных стадиях развития искусственного интеллекта часто возникала путаница между описанными выше подходами, например, иногда приходилось сталкиваться с такими утверждениями некоторых авторов, что предложенный ими алгоритм хорошо справляется с определенной задачей и поэтому является хорошей моделью способностей человека, или наоборот. Современные авторы излагают результаты своих исследований в этих двух областях отдельно; такое разделение позволяет развиваться быстрее как искусственному интеллекту, так и когнитологии. Но эти две научные области продолжают обогащать друг друга, особенно в таких направлениях, как зрительное восприятие и понимание естественного языка. В последнее время особенно значительные успехи достигнуты в области зрительного восприятия благодаря использованию интегрированного подхода, в котором применяются и нейрофизиологические экспериментальные данные, и вычислительные модели.

Как мыслить рационально: подход, основанный на использовании “законов мышления”

Греческий философ Аристотель был одним из первых, кто попытался определить законы “правильного мышления”, т.е. процессы формирования неопровержимых рассуждений. Его **силлогизмы** стали образцом для создания процедур доказательства, которые всегда позволяют прийти к правильным заключениям, если даны правильные предпосылки, например “Сократ — человек; все люди смертны; следовательно, Сократ смертен”. В основе этих исследований лежало предположение, что такие законы мышления управляют работой ума; на их основе развилось научное направление, получившее название **логики**.

В XIX столетии ученые, работавшие в области логики, создали точную систему логических обозначений для утверждений о предметах любого рода, которые встречаются в мире, и об отношениях между ними. (Сравните ее с обычной системой арифметических обозначений, которая предназначена в основном для формирования утверждений о равенстве и неравенстве чисел.) К 1965 году были уже разработаны программы, которые могли в принципе решить любую разрешимую проблему, описанную в системе логических обозначений³. Исследователи в области искусственного интеллекта, придерживающиеся так называемых традиций **логицизма**, надеются, что им удастся создать интеллектуальные системы на основе подобных программ.

Но при осуществлении указанного подхода возникают два серьезных препятствия. Во-первых, довольно сложно взять любые неформальные знания и выразить их в формальных терминах, требуемых для системы логических обозначений, особенно если эти знания не являются полностью достоверными. Во-вторых, возможность сравнительно легко решить проблему “в принципе” отнюдь не означает, что это действительно удастся сделать на практике. Даже такие задачи, в основе которых лежит несколько десятков фактов, могут исчерпать вычислительные ресурсы любого компьютера, если не используются определенные методы управления тем, какие этапы проведения рассуждений должны быть опробованы в первую очередь. Хотя с обоими этими препятствиями приходится сталкиваться при любой попытке создания вычислительных систем для автоматизации процесса проведения рассуждений, они были впервые обнаружены в рамках традиций логицизма.

³ Если же решение не существует, программа может так и не остановиться в процессе его поиска.

Как мыслить рационально: подход, основанный на использовании рационального агента

✎ **Агентом** считается все, что действует (слово *агент* произошло от из латинского слова *agere* — действовать). Но предполагается, что компьютерные агенты обладают некоторыми другими атрибутами, которые отличают их от обычных “программ”, такими как способность функционировать под автономным управлением, воспринимать свою среду, существовать в течение продолжительного периода времени, адаптироваться к изменениям и обладать способностью взять на себя достижение целей, поставленных другими. ✎ **Рациональным агентом** называется агент, который действует таким образом, чтобы можно было достичь наилучшего результата или, в условиях неопределенности, наилучшего ожидаемого результата.

В подходе к созданию искусственного интеллекта на основе “законов мышления” акцент был сделан на формировании правильных логических выводов. Безусловно, иногда формирование правильных логических выводов становится и частью функционирования рационального агента, поскольку один из способов рациональной организации своих действий состоит в том, чтобы логическим путем прийти к заключению, что данное конкретное действие позволяет достичь указанных целей, а затем действовать в соответствии с принятым решением. С другой стороны, правильный логический вывод не исчерпывает понятия рациональности, поскольку часто возникают ситуации, в которых невозможно однозначно выбрать какие-либо правильные действия, но все равно надо что-то делать. Кроме того, существуют способы рациональной организации действий, в отношении которых нельзя утверждать, что в них используется логический вывод. Например, отдергивание пальца от горячей печи — это рефлекторное действие, которое обычно является более успешным по сравнению с более медленным движением, сделанным после тщательного обдумывания всех обстоятельств.

Таким образом, все навыки, требуемые для прохождения теста Тьюринга, позволяют также осуществлять рациональные действия. Итак, прежде всего необходимо иметь возможность представлять знания и проводить на основании них рассуждения, поскольку это позволяет вырабатывать приемлемые решения в самых различных ситуациях. Необходимо обладать способностью формировать понятные предложения на естественном языке, поскольку в сложный социум принимают только тех, кто способен правильно высказывать свои мысли. Необходимо учиться не только ради приобретения эрудиции, но и в связи с тем, что лучшее представление о том, как устроен мир, позволяет вырабатывать более эффективные стратегии действий в этом мире. Нужно обладать способностью к зрительному восприятию не только потому, что процесс визуального наблюдения позволяет получать удовольствие, но и потому, что зрение подсказывает, чего можно достичь с помощью определенного действия, например тот, кто сумеет быстрее всех разглядеть лакомый кусочек, получит шанс подобраться к нему раньше других.

По этим причинам подход к исследованию искусственного интеллекта как области проектирования рациональных агентов имеет, по меньшей мере, два преимущества. Во-первых, этот подход является более общим по сравнению с подходом, основанном на использовании “законов мышления”, поскольку правильный логический вывод — это просто один из нескольких возможных механизмов достижения рациональности. Во-вторых, он является более перспективным для научной разработки по

сравнению с подходами, основанными на изучении человеческого поведения или человеческого мышления, поскольку стандарт рациональности четко определен и полностью обобщен. Человеческое поведение, с другой стороны, хорошо приспособлено лишь для одной определенной среды и отчасти является продуктом сложного и в основном неизученного эволюционного процесса, который, как оказалось, отнюдь не позволяет формировать существа, идеальные во всех отношениях.

☞ Поэтому данная книга в основном посвящена описанию общих принципов работы рациональных агентов и компонентов, необходимых для их создания. Из изложенного в ней станет очевидно, что несмотря на кажущуюся простоту формулировки этой проблемы, при попытке ее решения возникает невероятное количество трудностей. Некоторые из этих трудностей более подробно описываются в главе 2.

Следует всегда учитывать одно важное замечание: нужно неизменно исходить из того, что в сложной среде задача достижения идеальной рациональности, при которой всегда выполняются правильные действия, не осуществима. Дело в том, что при этом предъявляются слишком высокие требования к вычислительным ресурсам. Но в основной части данной книги применяется рабочая гипотеза, согласно которой идеальная рациональность является хорошей отправной точкой для анализа. Такой подход позволяет упростить задачу создания рационального агента и предоставляет подходящую основу для описания большей части теоретического материала в этой области. В главах 6 и 17 речь идет непосредственно о проблеме **ограниченной рациональности** — организации приемлемых действий в тех ситуациях, когда не хватает времени на выполнение всех вычислений, которые действительно могли бы потребоваться.

1.2. ПРЕДЫСТОРИЯ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

В данном разделе кратко описана история развития научных дисциплин, которые внесли свой вклад в область искусственного интеллекта в виде конкретных идей, воззрений и методов. Как и в любом историческом очерке, поневоле приходится ограничиваться описанием небольшого круга людей, событий и открытий, игнорируя все остальные факты, которые были не менее важными. Авторы построили этот исторический экскурс вокруг ограниченного круга вопросов. Безусловно, они не хотели бы, чтобы у читателя создалось такое впечатление, будто эти вопросы являются единственными, которые рассматриваются в указанных научных дисциплинах, или что сами эти дисциплины развивались исключительно ради того, чтобы их конечным итогом стало создание искусственного интеллекта.

Философия (период с 428 года до н.э. по настоящее время)

- Могут ли использоваться формальные правила для вывода правильных заключений?
- Как такой идеальный объект, как мысль, рождается в таком физическом объекте, как мозг?
- Каково происхождение знаний?
- Каким образом знания ведут к действиям?

Точный свод законов, руководящих рациональной частью мышления, был впервые сформулирован Аристотелем (384–322 годы до н.э.). Он разработал неформализованную систему силлогизмов, предназначенную для проведения правильных рассуждений, которая позволяла любому вырабатывать логические заключения механически, при наличии начальных предпосылок. Гораздо позднее Раймунд Луллий (умер в 1315 году) выдвинул идею, что полезные рассуждения можно фактически проводить с помощью механического артефакта. Предложенные им “концептуальные колеса” показаны на обложке данной книги. Томас Гоббс (1588–1679) предположил, что рассуждения аналогичны числовым расчетам и что “в наших неслышимых мыслях мы поневоле складываем и вычитаем”. В то время автоматизация самих вычислений уже шла полным ходом; примерно в 1500 году Леонардо да Винчи (1452–1519) спроектировал, но не построил механический калькулятор; недавно проведенная реконструкция показала, что его проект является работоспособным. Первая известная вычислительная машина была создана примерно в 1623 году немецким ученым Вильгельмом Шиккардом (1592–1635), хотя более известна машина Паскалина, построенная в 1642 году Блезом Паскалем (1623–1662). Паскаль писал, что “арифметическая машина производит эффект, который кажется более близким к мышлению по сравнению с любыми действиями животных”. Готтфрид Вильгельм Лейбниц (1646–1716) создал механическое устройство, предназначенное для выполнения операций над понятиями, а не над числами, но область его действия была довольно ограниченной.

После того как человечество осознало, каким должен быть набор правил, способных описать формальную, рациональную часть мышления, следующим этапом оказалось то, что разум стал рассматриваться как физическая система. Рене Декарт (1596–1650) впервые опубликовал результаты обсуждения различий между разумом и материей, а также возникающих при этом проблем. Одна из проблем, связанных с чисто физическими представлениями о разуме, состоит в том, что они, по-видимому, почти не оставляют места для свободной воли: ведь если разум руководствуется исключительно физическими законами, то человек проявляет не больше свободной воли по сравнению с булыжником, “решившим” упасть в направлении к центру земли. Несмотря на то что Декарт был убежденным сторонником взглядов, признающих только власть разума, он был также приверженцем **дуализма**. Декарт считал, что существует такая часть человеческого разума (душа, или дух), которая находится за пределами естества и не подчиняется физическим законам. С другой стороны, животные не обладают таким дуалистическим свойством, поэтому их можно рассматривать как своего рода машины. Альтернативой дуализму является **материализм**, согласно которому разумное поведение складывается из операций, выполняемых мозгом в соответствии с законами физики. *Свободная воля* — это просто форма, в которую в процессе выбора преобразуется восприятие доступных вариантов.

Если предположить, что знаниями манипулирует физический разум, то возникает следующая проблема — установить источник знаний. Такое научное направление, как **эмпиризм**, родоначальником которого был Фрэнсис Бекон (1561–1626), автор *Нового Органона*⁴, можно охарактеризовать высказыванием Джона Локка (1632–1704): “В человеческом понимании нет ничего, что не проявлялось бы прежде всего в ощущениях”. Дэвид Юм (1711–1776) в своей книге *A Treatise of Human Nature*

⁴ Эта книга была выпущена как новая версия *Органона* (или инструмента мышления) Аристотеля.

(Трактат о человеческой природе) [705] предложил метод, известный теперь под названием **принципа индукции**, который состоит в том, что общие правила вырабатываются путем изучения повторяющихся ассоциаций между элементами, которые рассматриваются в этих правилах. Основываясь на работе Людвиг Виттгенштейна (1889–1951) и Бертрана Рассела (1872–1970), знаменитый Венский кружок, возглавляемый Рудольфом Карнапом (1891–1970), разработал доктрину **логического позитивизма**. Согласно этой доктрине все знания могут быть охарактеризованы с помощью логических теорий, связанных в конечном итоге с **констатирующими предложениями**, которые соответствуют входным сенсорным данным⁵. В **теории подтверждения** Рудольфа Карнапа и Карла Хемпеля (1905–1997) предпринята попытка понять, как знания могут быть приобретены из опыта. В книге Карнапа *The Logical Structure of the World* [223] определена явно заданная вычислительная процедура для извлечения знаний из результатов элементарных опытов. По-видимому, это — первая теория мышления как вычислительного процесса.

Заключительным элементом в этой картине философских исследований проблемы разума является связь между знаниями и действиями. Данный вопрос для искусственного интеллекта является жизненно важным, поскольку интеллектуальность требует не только размышлений, но и действий. Кроме того, только поняв способы обоснования действий, можно понять, как создать агента, действия которого будут обоснованными (или рациональными). Аристотель утверждал, что действия обоснованы логической связью между целями и знаниями о результатах данного конкретного действия (последняя часть приведенной ниже цитаты Аристотеля на языке оригинала размещена также на обложке данной книги). Характерным примером рассуждений о рациональных действиях являются следующие.

Но почему происходит так, что размышления иногда сопровождаются действием, а иногда — нет, иногда за ними следует движение, а иногда — нет? Создается впечатление, как будто почти то же самое происходит и в случае построения рассуждений и формирования выводов о неизменных объектах. Но в таком случае целью умственной деятельности оказывается умозрительное суждение..., тогда как заключением, которое следует из данных двух предпосылок, является действие... Мне нужна защита от дождя; защитой может послужить плащ. Мне нужен плащ. Я должен сам изготовить то, в чем я нуждаюсь; я нуждаюсь в плаще. Я должен изготовить плащ. И заключение “я должен изготовить плащ” становится действием ([1151, с. 40]).

В книге *Никомахова этика* (том III. 3, 1112b) Аристотеля можно найти более подробные рассуждения на эту тему, где также предложен алгоритм.

Нам предоставляется право выбора не целей, а средств достижения цели, ведь врач рассуждает не о том, должен ли он лечить, а оратор — не о том, станет ли он убеждать... Поставив цель, он размышляет, как и какими средствами ее достичь; а если окажется несколько средств, то определяет, какое из них самое простое и наилучшее; если же достижению цели служит одно средство, думает, как ее достичь при помощи этого средства и что будет средством для этого средства, пока не дойдет до первой причины, которую находит последней... и то, что было последним в порядке анализа, обычно становится первым в порядке осуществления... Если же он приходит к выводу, что цель недостижима, отступает-

⁵ В данной картине мира все осмысленные утверждения можно подтвердить или опровергнуть либо с помощью анализа смысла слов, либо путем проведения экспериментов. Поскольку при этом основная часть метафизики остается за бортом, в чем и состояло намерение создателей данного направления, логический позитивизм в некоторых кругах встретил неодобрительное отношение.

ся, например, если нужны деньги, а достать их нельзя; но если достижение цели кажется возможным, то пытаются ее достичь.

Алгоритм Аристотеля был реализован через 2300 лет Ньюэллом и Саймоном в программе GPS. Теперь то, что создано на его базе, принято называть *регрессивной системой планирования* (см. главу 11).


Анализ на основе цели является полезным, но не дает ответа на то, что делать, если к цели ведет несколько вариантов действий или ни один вариант действий не позволяет достичь ее полностью. Антуан Арно (1612–1694) правильно описал количественную формулу для принятия решения о том, какое действие следует предпринять в подобных случаях (см. главу 16). В книге *Utilitarianism* приверженца утилитаризма Джона Стюарта Милла (1806–1873) [1050] провозглашена идея о том, что критерии принятия рациональных решений должны применяться во всех сферах человеческой деятельности. Более формальная теория принятия решений рассматривается в следующем разделе.

Математика (период примерно с 800 года по настоящее время)

- Каковы формальные правила формирования правильных заключений?
- Как определить пределы вычислимости?
- Как проводить рассуждения с использованием недостоверной информации?

Философы сформулировали наиболее важные идеи искусственного интеллекта, но для преобразования его в формальную науку потребовалось достичь определенного уровня математической формализации в трех фундаментальных областях: логика, вычисления и вероятность.

Истоки идей формальной логики можно найти в работах философов древней Греции (см. главу 7), но ее становление как математической дисциплины фактически началась с трудов Джорджа Буля (1815–1864), который детально разработал логику высказываний, или булеву логику [149]. В 1879 году Готтлоб Фреге (1848–1925) расширил булеву логику для включения в нее объектов и отношений, создав логику первого порядка, которая в настоящее время используется как наиболее фундаментальная система представления знаний⁶. Альфред Тарский (1902–1983) впервые ввел в научный обиход теорию ссылок, которая показывает, как связать логические объекты с объектами реального мира. Следующий этап состоял в определении пределов того, что может быть сделано с помощью логики и вычислений.

Первым нетривиальным  алгоритмом считается алгоритм вычисления наибольшего общего знаменателя, предложенный Евклидом. Исследование алгоритмов как самостоятельных объектов было начато аль-Хорезми, среднеазиатским математиком IX столетия, благодаря работам которого Европа познакомилась с арабскими цифрами и алгеброй. Буль и другие ученые широко обсуждали алгоритмы логического вывода, а к концу XIX столетия уже предпринимались усилия по формализации общих принципов проведения математических рассуждений как логического вывода. В 1900 году Давид Гильберт (1862–1943) представил список из 23 проблем и правильно предсказал, что эти проблемы будут занимать математиков почти до кон-

⁶ Предложенная Готтлобом Фреге система обозначений для логики первого порядка так и не нашла широкого распространения по причинам, которые становятся сразу же очевидными из примера, приведенного на первой странице обложки.

ца XX века. Последняя из этих проблем представляет собой вопрос о том, существует ли алгоритм для определения истинности любого логического высказывания, в состав которого входят натуральные числа. Это — так называемая знаменитая проблема поиска решения (*Entscheidungsproblem*). По сути, этот вопрос, заданный Гильбертом, сводился к определению того, есть ли фундаментальные пределы, ограничивающие мощь эффективных процедур доказательств. В 1930 году Курт Гёдель (1906–1978) показал, что существует эффективная процедура доказательства любого истинного высказывания в логике первого порядка Фреге и Рассела, но при этом логика первого порядка не позволяет выразить принцип математической индукции, необходимый для представления натуральных чисел. В 1931 году Гёдель показал, что действительно существуют реальные пределы вычислимости. Предложенная им **теорема о неполноте** показывает, что в любом языке, достаточно выразительном для описания свойств натуральных чисел, существуют истинные высказывания, которые являются недоказуемыми, в том смысле, что их истинность невозможно установить с помощью какого-либо алгоритма.

Этот фундаментальный результат может также рассматриваться как демонстрация того, что имеются некоторые функции от целых чисел, которые не могут быть представлены с помощью какого-либо алгоритма, т.е. они не могут быть вычислены. Это побудило Алана Тьюринга (1912–1954) попытаться точно охарактеризовать, какие функции способны быть вычисленными. Этот подход фактически немного проблематичен, поскольку в действительности понятию вычисления, или эффективной процедуры вычисления, не может быть дано формальное определение. Но общепризнано, что вполне удовлетворительное определение дано в тезисе Чёрча–Тьюринга, который указывает, что машина Тьюринга [1518] способна вычислить любую вычислимую функцию. Кроме того, Тьюринг показал, что существуют некоторые функции, которые не могут быть вычислены машиной Тьюринга. Например, вообще говоря, ни одна машина не способна определить, возвратит ли данная конкретная программа ответ на конкретные входные данные или будет работать до бесконечности.

Хотя для понимания возможностей вычисления очень важны понятия недоказуемости и невычислимости, гораздо большее влияние на развитие искусственного интеллекта оказало понятие **неразрешимости**. Грубо говоря, задача называется *неразрешимой*, если время, требуемое для решения отдельных экземпляров этой задачи, растёт экспоненциально с увеличением размеров этих экземпляров. Различие между полиномиальным и экспоненциальным ростом сложности было впервые подчеркнуто в середине 1960-х годов в работах Кобхэма [272] и Эдмондса [430]. Важность этого открытия состоит в следующем: экспоненциальный рост означает, что даже экземпляры задачи умеренной величины не могут быть решены за какое-либо приемлемое время. Поэтому, например, приходится заниматься разделением общей задачи выработки интеллектуального поведения на разрешимые подзадачи, а не пытаться решать неразрешимую задачу.

Как можно распознать неразрешимую проблему? Один из приемлемых методов такого распознавания представлен в виде теории **NP-полноты**, впервые предложенной Стивеном Куком [289] и Ричардом Карпом [772]. Кук и Карп показали, что существуют большие классы канонических задач комбинаторного поиска и формирования рассуждений, которые являются NP-полными. Существует вероятность того, что любой класс задач, к которому сводится этот класс NP-полных задач, является неразрешимым. (Хотя еще не было доказано, что NP-полные задачи обязатель-

но являются неразрешимыми, большинство теоретиков считают, что дело обстоит именно так.) Эти результаты контрастируют с тем оптимизмом, с которым в популярных периодических изданиях приветствовалось появление первых компьютеров под такими заголовками, как “Электронные супермозги”, которые думают “быстрее Эйнштейна!” Несмотря на постоянное повышение быстродействия компьютеров, характерной особенностью интеллектуальных систем является экономное использование ресурсов. Грубо говоря, наш мир, в котором должны освоиться системы ИИ, — это чрезвычайно крупный экземпляр задачи. В последние годы методы искусственного интеллекта помогли разобраться в том, почему некоторые экземпляры NP-полных задач являются сложными, а другие простыми [244].

Кроме логики и теории вычислений, третий по величине вклад математиков в искусственный интеллект состоял в разработке **теории вероятностей**. Идея вероятности была впервые сформулирована итальянским математиком Джероламо Кардано (1501–1576), который описал ее в терминах результатов событий с несколькими исходами, возникающих в азартных играх. Теория вероятностей быстро стала неотъемлемой частью всех количественных наук, помогая использовать недостоверные результаты измерений и неполные теории. Пьер Ферма (1601–1665), Блез Паскаль (1623–1662), Джеймс Бернулли (1654–1705), Пьер Лаплас (1749–1827) и другие ученые внесли большой вклад в эту теорию и ввели новые статистические методы. Томас Байес (1702–1761) предложил правило обновления вероятностей с учетом новых фактов. Правило Байеса и возникшее на его основе научное направление, называемое *байесовским анализом*, лежат в основе большинства современных подходов к проведению рассуждений с учетом неопределенности в системах искусственного интеллекта.

Экономика (период с 1776 года по настоящее время)

- Как следует организовать принятие решений для максимизации вознаграждения?
- Как действовать в таких условиях, когда другие могут препятствовать осуществлению намеченных действий?
- Как действовать в таких условиях, когда вознаграждение может быть предоставлено лишь в отдаленном будущем?

Экономика как наука возникла в 1776 году, когда шотландский философ Адам Смит (1723–1790) опубликовал свою книгу *An Inquiry into the Nature and Causes of the Wealth of Nations* (Исследование о природе и причинах богатства народов). Важный вклад в экономику был сделан еще древнегреческими учеными и другими предшественниками Смита, но только Смит впервые сумел оформить эту область знаний как науку, используя идею, что любую экономику можно рассматривать как состоящую из отдельных агентов, стремящихся максимизировать свое собственное экономическое благосостояние. Большинство людей считают, что экономика посвящена изучению денежного оборота, но любой экономист ответит на это, что в действительности он изучает то, как люди делают выбор, который ведет к предпочтительным для них результатам. Математическая трактовка понятия “предпочтительных результатов”, или *полезности*, была впервые формализована Леоном Валрасом (1834–1910), уточнена Фрэнком Рамсеем [1265], а затем усовершенствована Джоном фон Нейманом и Оскаром Моргенштерном в книге *The Theory of Games and Economic Behavior* (Теория игр и экономического поведения) [1546].

✎ **Теория решений**, которая объединяет в себе теорию вероятностей и теорию полезности, предоставляет формальную и полную инфраструктуру для принятия решений (в области экономики или в другой области) в условиях неопределенности, т.е. в тех случаях, когда среда, в которой действует лицо, принимающее решение, наиболее адекватно может быть представлена лишь с помощью вероятностных описаний. Она хорошо подходит для “крупных” экономических образований, где каждый агент не обязан учитывать действия других агентов как индивидуумов. А в “небольших” экономических образованиях ситуация в большей степени напоминает **игру**, поскольку действия одного игрока могут существенно повлиять на полезность действий другого (или положительно, или отрицательно). ✎ **Теория игр**, разработанная фон Нейманом и Моргенштерном (см. также [963]), позволяет сделать неожиданный вывод, что в некоторых играх рациональный агент должен действовать случайным образом или, по крайней мере, таким образом, который кажется случайным для соперников.

Экономисты чаще всего не стремятся найти ответ на третий вопрос, приведенный выше, т.е. не пытаются выработать способ принятия рациональных решений в тех условиях, когда вознаграждение в ответ на определенные действия не предоставляется немедленно, а становится результатом нескольких действий, выполненных в определенной последовательности. Изучению этой темы посвящена область ✎ **исследования операций**, которая возникла во время Второй мировой войны в результате усилий, которые были предприняты в Британии по оптимизации работы радарных установок, а в дальнейшем нашла применение и в гражданском обществе при выработке сложных управленческих решений. В работе Ричарда Беллмана [97] формализован определенный класс последовательных задач выработки решений, называемых **марковскими процессами принятия решений** (Markov Decision Process — MDP), которые рассматриваются в главах 17 и 21.

Работы в области экономики и исследования операций оказали большое влияние на сформулированное в этой книге понятие рациональных агентов, но в течение многих лет исследования в области искусственного интеллекта проводились совсем по другим направлениям. Одной из причин этого была кажущаяся **сложность** задачи выработки рациональных решений. Тем не менее Герберт Саймон (1916–2001) в некоторых из своих ранних работ показал, что лучшее описание фактического поведения человека дают модели, основанные на ✎ **удовлетворении** (принятии решений, которые являются “достаточно приемлемыми”), а не модели, предусматривающие трудоемкий расчет оптимального решения [1414], и стал одним из первых исследователей в области искусственного интеллекта, получившим Нобелевскую премию по экономике (это произошло в 1978 году). В 1990-х годах наблюдалось возрождение интереса к использованию методов теории решений для систем агентов [1576].

Неврология (период с 1861 года по настоящее время)

- Как происходит обработка информации в мозгу?

✎ **Неврология** — это наука, посвященная изучению нервной системы, в частности мозга. Одной из величайших загадок, не поддающихся научному описанию, остается определение того, как именно мозг обеспечивает мышление. Понимание того, что мышление каким-то образом связано с мозгом, существовало в течение тысяч лет, поскольку люди обнаружили, что сильные удары по голове могут привести к ум-

ственному расстройству. Кроме того, уже давно было известно, что человеческий мозг обладает какими-то важными особенностями; еще примерно в 335 до н.э. Аристотель⁷ писал: “Из всех животных только человек имеет самый крупный мозг по сравнению с его размерами”. Тем не менее широкое признание того, что мозг является вместилищем сознания, произошло только в середине XVIII столетия. До этого в качестве возможных источников сознания рассматривались сердце, селезенка и шишковидная железа (эпифиз).

Исследования афазии (нарушения речи) у пациентов с повреждением мозга, проведенные Полем Брока (1824–1880) в 1861 году, снова пробудили интерес к этой научной области и послужили для многих представителей медицины доказательством существования в мозгу локализованных участков, ответственных за конкретные познавательные функции. Например, этот ученый показал, что функции формирования речи сосредоточены в той части левого полушария, которая теперь называется *зоной Брока*⁸. К тому времени уже было известно, что мозг состоит из нервных клеток, или *нейронов*, но только в 1873 году Камилло Гольджи (1843–1926) сумел разработать надежный метод, позволяющий наблюдать за отдельными нейронами в мозгу (рис. 1.1). Этот метод использовал Сантьяго Рамон и Кахал (1852–1934) в своих пионерских исследованиях нейронных структур мозга⁹.

Теперь ученые располагают некоторыми данными о том, как связаны между собой отдельные области мозга и те части тела, которыми они управляют или от которых получают сенсорные данные. Оказалось, что подобная привязка может коренным образом измениться в течение нескольких недель, а у некоторых животных, по видимому, имеется несколько вариантов такой привязки. Более того, еще не совсем понятно, как другие области могут взять на себя функции поврежденных областей. К тому же почти полностью отсутствуют обоснованные теории того, как осуществляется хранение информации в памяти индивидуума.

Измерение активности неповрежденного мозга началось в 1929 году с изобретения электроэнцефалографа (ЭЭГ) Гансом Бергером. Разработки в области получения изображений на основе функционального магнитного резонанса [1152] позволили неврологам получать исключительно подробные изображения активности мозга, что дает возможность проводить измерения характеристик физиологических процессов, которые связаны с происходящими познавательными процессами какими-то интересными способами. Эти возможности для исследований становятся еще более широкими благодаря прогрессу в области регистрации нейронной активности отдельной клетки. Но, несмотря на эти успехи, ученые еще очень далеки от понимания того, как действительно осуществляется любой из этих познавательных процессов.

⁷ С тех пор было обнаружено, что некоторые виды дельфинов и китов имеют относительно более крупный мозг. Современные ученые считают, что большие размеры человеческого мозга отчасти обусловлены усовершенствованием системы его охлаждения на последних этапах эволюции человека.

⁸ Многие цитируют в качестве возможного более раннего источника работу Александра Гуда [673].

⁹ Гольджи упорно отстаивал свое мнение, что функции мозга осуществляются в основном непрерывной средой, в которую включены нейроны, тогда как Кахал проповедовал “нейронную доктрину”. Эти ученые совместно получили Нобелевскую премию в 1906 году, но в роли лауреатов произнесли речи, содержащие довольно антагонистичные взаимные выпады.

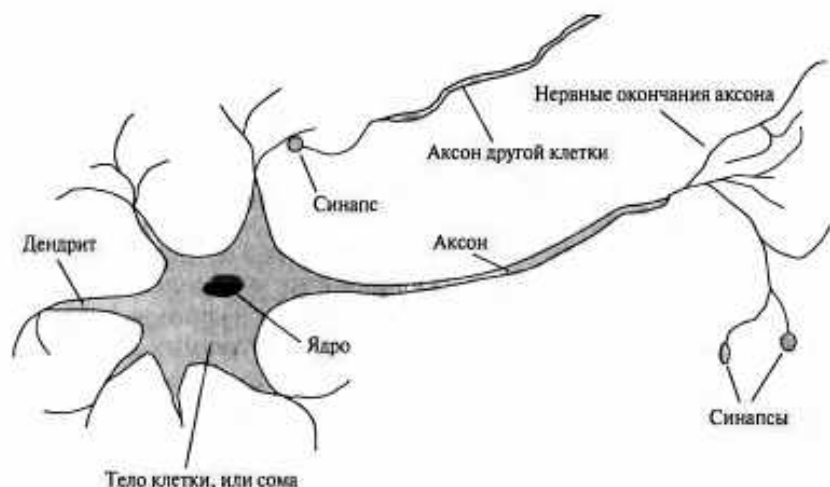


Рис. 1.1. Части нервной клетки, или нейрона. Каждый нейрон состоит из тела клетки (или соммы), которое содержит ядро клетки. От тела клетки ответвляется множество коротких волокон, называемых дендритами, и одно длинное волокно, называемое аксоном. Аксон растягивается на большое расстояние, намного превышающее то, что показано в масштабе этого рисунка. Обычно аксоны имеют длину 1 см (что превышает в 100 раз диаметр тела клетки), но могут достигать 1 метра. Нейрон создает соединения с другими нейронами, количество которых может составлять от 10 до 100 000 в точках сопряжения, называемых синапсами. Сигналы распространяются от одного нейрона к другому с помощью сложной электрохимической реакции. Эти сигналы управляют активностью мозга в течение короткого интервала, а также становятся причиной долговременных изменений состояния самих нейронов и их соединений. Считается, что эти механизмы служат в мозгу основой для обучения. Обработка информации главным образом происходит в коре головного мозга, которая представляет собой самый внешний слой нейронов мозга. По-видимому, основной структурной единицей является столбец ткани, имеющий диаметр около 0,5 мм и протяженность на всю глубину коры, толщина которой в человеческом мозгу составляет около 4 мм. Каждый столбец содержит примерно 20 000 нейронов

Тем не менее работы в области неврологии позволяют сделать поистине удивительное заключение о том, что *совместная работа простых клеток может приводить к появлению мышления, действия и сознания* или, другими словами, что мозг порождает разум [1379]. После этого открытия единственной реально существующей альтернативной теорией остается мистицизм, приверженцы которого провозглашают, что существует некое мистическое пространство, находящееся за пределами физического опыта, в котором функционирует разум.

Мозг и цифровой компьютер выполняют совершенно разные задачи и имеют различные свойства. В табл. 1.2 показано, что в типичном мозгу человека имеется в 1000 раз больше нейронов, чем логических элементов в процессоре типичного компьютера высокого класса. В соответствии с законом Мура¹⁰ может быть сделан про-

¹⁰ Закон Мура указывает, что плотность транзисторов в расчете на единицу площади удваивается примерно через каждые 1–1,5 года. Количество нейронов в мозгу человека, по расчетам, должно удваиваться примерно через каждые 2–4 миллиона лет.

гноз, что количество логических элементов в процессоре станет равным количеству нейронов в мозгу примерно к 2020 году. Безусловно, эти прогнозы мало о чем говорят; кроме того, это различие в отношении количества элементов является незначительным по сравнению с различием в скорости переключения и степени распараллеливания. Микросхемы компьютера способны выполнить отдельную команду меньше чем за наносекунду, тогда как нейроны действуют в миллионы раз медленнее. Но мозг сторицей восполняет этот свой недостаток, поскольку все его нейроны и синапсы действуют одновременно, тогда как большинство современных компьютеров имеет только один процессор или небольшое количество процессоров. Таким образом, *даже несмотря на то, что компьютер обладает преимуществом более чем в миллион раз в физической скорости переключения, оказывается, что мозг по сравнению с ним выполняет все свои действия примерно в 100 000 раз быстрее.*

Таблица 1.2. Грубое сравнение физических вычислительных ресурсов, имеющихся в компьютере (приблизительные оценки по состоянию на 2003 год) и в мозгу. Со времени выпуска первого издания данной книги показатели компьютера выросли, по меньшей мере, в 10 раз и будут, как ожидается, продолжать расти в течение текущего десятилетия. А показатели мозга за последние 10 000 лет не изменились

	Компьютер	Человеческий мозг
Вычислительные модули	Один центральный процессор, 10 ⁸ логических элементов	10 ¹¹ нейронов
Модули памяти	Оперативная память на 10 ¹⁰ битов	10 ¹¹ нейронов
	Диск емкостью 10 ¹¹ битов	10 ¹⁴ синапсов
Продолжительность цикла обработки	10 ⁻⁹ секунды	10 ⁻³ секунды
Пропускная способность	10 ¹⁰ бит/с	10 ¹⁴ бит/с
Количество обновлений памяти в секунду	10 ⁹	10 ¹⁴

Психология (период с 1879 года по настоящее время)

- Как думают и действуют люди и животные?

Истоки научной психологии обычно прослеживаются до работ немецкого физика Германа фон Гельмгольца (1821–1894) и его студента Вильгельма Вундта (1832–1920). Гельмгольц применил научный метод для изучения зрения человека, и выпущенная им книга *Handbook of Physiological Optics* даже в наши дни характеризуется как “непревзойденный по своей важности вклад в изучение физики и физиологии зрения человека” [1111, с. 15]. В 1879 году Вундт открыл первую лабораторию по экспериментальной психологии в Лейпцигском университете. Вундт настаивал на проведении тщательно контролируемых экспериментов, в которых его сотрудники выполняли задачи по восприятию или формированию ассоциаций, проводя интроспективные наблюдения за своими мыслительными процессами. Такой тщательный контроль позволил ему сделать очень многое для превращения психологии в науку, но из-за субъективного характера данных вероятность того, что экспериментатор будет стремиться опровергнуть выдвинутые им теории, оставалась очень низкой. С другой стороны, биологи, изучающие поведение животных, не пользовались интроспективными данными и разработали объективную методологию, как показал Г.С. Дженнингс [733] в своей важной работе *Behavior of the Lower Organisms*. Распространяя этот подход на людей, сторонники *бихевиористского* движения, возглавляемые Джоном Уотсоном


(1878–1958), отвергали любую теорию, учитывающую мыслительные процессы, на том основании, что интроспекция не может предоставлять надежные свидетельства. Бихевиористы настаивали на том, что следует изучать только объективные меры восприятия (или стимулы), предъявленные животному, и вытекающие из этого действия (или отклики на стимулы). Такие мыслительные конструкции, как знания, убеждения, цели и последовательные рассуждения, отвергались как ненаучная “обывательская психология”. Бихевиоризм позволил многое узнать о крысах и голубях, но оказался менее успешным при изучении людей. Тем не менее это научное направление сохраняло за собой мощные позиции в области психологии (особенно в Соединенных Штатах Америки) в период примерно с 1920 по 1960 годы.

Взгляды, согласно которым мозг рассматривается как устройство обработки информации, характерные для представителей ~~э~~ когнитивной психологии, прослеживаются, по крайней мере, до работ Уильяма Джеймса¹¹ (1842–1910). Гельмгольц также утверждал, что восприятие связано с определенной формой подсознательного логического вывода. В Соединенных Штатах такой подход к изучению познавательных процессов был в основном отвергнут из-за широкого распространения бихевиористских взглядов, но на факультете прикладной психологии Кембриджского университета, возглавляемом Фредериком Бартлеттом (1886–1969), удалось организовать проведение широкого спектра работ в области когнитивного моделирования. В своей книге *The Nature of Explanation* студент и последователь Бартлетта, Кеннет Крэг [306], привел весомые доводы в пользу допустимости применения таких “мыслительных” терминов, как убеждения и цели, доказав, что они являются не менее научными, чем, скажем, такие термины, применяемые в рассуждениях о газах, как давление и температура, несмотря на то, что речь в них идет о молекулах, которые сами не обладают этими характеристиками. Крэг обозначил следующие три этапа деятельности агента, основанного на знаниях: во-первых, действующий стимул должен быть преобразован во внутреннее представление, во-вторых, с этим представлением должны быть выполнены манипуляции с помощью познавательных процессов для выработки новых внутренних представлений, и, в-третьих, они должны быть, в свою очередь, снова преобразованы в действия. Он наглядно объяснил, почему такой проект является приемлемым для любого агента.

Если живой организм несет в своей голове “модель в уменьшенном масштабе” внешней реальности и своих возможных действий, то обладает способностью проверять различные варианты, приходить к заключению, какой из них является наилучшим, реагировать на будущие ситуации, прежде чем они возникнут, использовать знания о прошлых событиях, сталкиваясь с настоящим и будущим, и во всех отношениях реагировать на опасности, встречаясь с ними, гораздо полнее, безопаснее для себя, а также в более компетентной форме [306].

В 1945 году, после смерти Крэга в результате несчастного случая во время катания на велосипеде, его работа была продолжена Дональдом Броудбентом, книга *Perception and Communication* [188] которого включила некоторые из первых моделей информационной обработки психологических феноменов. Между тем в Соединенных Штатах работы в области компьютерного моделирования привели к созданию такого

¹¹ Уильям Джеймс был братом писателя Генри Джеймса. Говорили, что Генри пишет свои романы так, как если бы они были трудами по психологии, а Уильям сочиняет свои труды по психологии так, как если бы это были романы.

научного направления, как  **когнитология**. Существует такое мнение, что зарождение этого направления произошло на одном из семинаров в Массачусеттском технологическом институте в сентябре 1956 года. (Ниже показано, что это событие произошло всего лишь через два месяца после проведения конференции, на которой “родился” сам искусственный интеллект.) На этом семинаре Джордж Миллер представил доклад *The Magic Number Seven*, Ноам Хомский прочитал доклад *Three Models of Language*, а Аллен Ньюэлл и Герберт Саймон представили свою работу *The Logic Theory Machine*. В этих трех работах, получивших широкую известность, было показано, как можно использовать компьютерные модели для решения задач в области психологии, запоминания, обработки естественного языка и логического мышления. В настоящее время среди психологов находят широкое признание взгляды на то, что “любая теория познания должна напоминать компьютерную программу” [30], т.е. она должна подробно описывать механизм обработки информации, с помощью которого может быть реализована некоторая познавательная функция.

Вычислительная техника (период с 1940 года по настоящее время)

- Каким образом можно создать эффективный компьютер?

Для успешного создания искусственного интеллекта требуется, во-первых, интеллект и, во-вторых, артефакт. Наиболее предпочтительным артефактом в этой области всегда был компьютер. Современный цифровой электронный компьютер был изобретен независимо и почти одновременно учеными трех стран, участвующих во Второй мировой войне. Первым операционным компьютером было электромеханическое устройство Heath Robinson¹², созданное в 1940 году группой Алана Тьюринга для единственной цели — расшифровки сообщений, передаваемых немецкими войсками. В 1943 году та же группа разработала мощный компьютер общего назначения, получивший название Colossus, в конструкции которого применялись электронные лампы¹³. Первым операционным программируемым компьютером был компьютер Z-3, изобретенный Конрадом Цузе в Германии в 1941 году. Цузе изобрел также числа с плавающей точкой и создал первый язык программирования высокого уровня, Plankalkül. Первый электронный компьютер, ABC, был собран Джоном Атанасовым и его студентом Клиффордом Берри в период с 1940 по 1942 год в университете штата Айова. Исследования Атанасова почти не получили поддержки или признания; как оказалось, наибольшее влияние на развитие современных компьютеров оказал компьютер ENIAC, разработанный в составе секретного военного проекта в Пенсильванском университете группой специалистов, в состав которой входили Джон Мочли и Джон Экерт.

За прошедшее с тех пор столетие появилось несколько поколений компьютерного аппаратного обеспечения, причем каждое из них характеризовалось увеличением скорости и производительности, а также снижением цены. Производительность компьютеров, созданных на основе кремниевых микросхем, удваивается примерно через каждые 18 месяцев, и такая скорость роста наблюдается уже в течение

¹² Хет Робинсон (Heath Robinson), в честь которого названо это устройство, был карикатуристом, знаменитым тем, что изображал причудливые и абсурдно усложненные картины таких повседневных действий, как намазывание тостов маслом.

¹³ В послевоенный период Тьюринг высказал пожелание применить эти компьютеры для исследований в области искусственного интеллекта, например для разработки одной из первых шахматных программ [1521], но его усилия были заблокированы британским правительством.

двух десятилетий. После достижения пределов этого роста потребуется молекулярная инженерия или какая-то другая, новая технология.

Безусловно, вычислительные устройства существовали и до появления электронного компьютера. Одно из первых автоматизированных устройств, появившееся еще в XVII столетии, рассматривалось на с. 41. Первым программируемым устройством был ткацкий станок, изобретенный в 1805 году Жозефом Марией Жаккардом (1752–1834), в котором использовались перфокарты для хранения инструкций по плетению узоров ткани. В середине XIX столетия Чарльз Бэббидж (1792–1871) разработал две машины, но ни одну из них не успел закончить. Его “разностная машина”, которая показана на обложке данной книги, предназначалась для вычисления математических таблиц, используемых в инженерных и научных проектах. В дальнейшем эта машина была построена и ее работа продемонстрирована в 1991 году в лондонском Музее науки [1481]. Другой замысел Бэббиджа, проект “аналитической машины”, был гораздо более амбициозным: в этой машине предусмотрено использование адресуемой памяти, хранимых программ и условных переходов, и она была первым артефактом, способным выполнять универсальные вычисления. Коллега Бэббиджа Ада Лавлейс, дочь поэта Лорда Байрона, была, возможно, первым в мире программистом. (В ее честь назван язык программирования Ada.) Она писала программы для незаконченной аналитической машины и даже размышляла над тем, что эта машина сможет играть в шахматы или сочинять музыку.

Искусственный интеллект во многом обязан также тем направлениям компьютерных наук, которые касаются программного обеспечения, поскольку именно в рамках этих направлений создаются операционные системы, языки программирования и инструментальные средства, необходимые для написания современных программ (и статей о них). Но эта область научной деятельности является также одной из тех, где искусственный интеллект в полной мере возмещает свои долг: работы в области искусственного интеллекта стали источником многих идей, которые затем были воплощены в основных направлениях развития компьютерных наук, включая разделение времени, интерактивные интерпретаторы, персональные компьютеры с оконными интерфейсами и поддержкой позиционирующих устройств, применение среды ускоренной обработки, создание типов данных в виде связанных списков, автоматическое управление памятью и ключевые концепции символического, функционального, динамического и объектно-ориентированного программирования.

Теория управления и кибернетика (период с 1948 года по настоящее время)

- Каким образом артефакты могут работать под своим собственным управлением?

Первое самоуправляемое устройство было построено Ктесибием из Александрии (примерно в 250 году до н.э.); это были водяные часы с регулятором, который поддерживал поток воды, текущий через эти часы с постоянным, предсказуемым расходом. Это изобретение изменило представление о том, на что могут быть способны устройства, созданные человеком. До его появления считалось, что только живые существа способны модифицировать свое поведение в ответ на изменения в окружающей среде. К другим примерам саморегулирующихся систем управления с обратной связью относятся регулятор паровой машины, созданный Джеймсом Уаттом (1736–1819), и термостат, изобретенный Корнелисом Дреббелем (1572–1633), кото-

рый изобрел также подводную лодку. Математическая теория устойчивых систем с обратной связью была разработана в XIX веке.

Центральной фигурой в создании науки, которая теперь именуется **теорией управления**, был Норберт Винер (1894–1964). Винер был блестящим математиком, который совместно работал со многими учеными, включая Бертрانا Рассела, под влиянием которых у него появился интерес к изучению биологических и механических систем управления и их связи с познанием. Как и Крэг (который также использовал системы управления в качестве психологических моделей), Винер и его коллеги Артуро Розенблют и Джулиан Бигелоу бросили вызов ортодоксальным бихевиористским взглядам [1306]. Они рассматривали целенаправленное поведение как обусловленное действием регуляторного механизма, пытающегося минимизировать “ошибку” — различие между текущим и целевым состоянием. В конце 1940-х годов Винер совместно с Уорреном Мак-Каллоком, Уолтером Питтсом и Джоном фон Нейманом организовал ряд конференций, на которых рассматривались новые математические и вычислительные модели познания; эти конференции оказали большое влияние на взгляды многих других исследователей в области наук о поведении. Книга Винера *Cybernetics* [1589], в которой было впервые дано определение **кибернетики** как науки, стала бестселлером и убедила широкие круги общественности в том, что мечта о создании машин, обладающих искусственным интеллектом, воплотилась в реальность.

Предметом современной теории управления, особенно той ее ветви, которая получила название *стохастического оптимального управления*, является проектирование систем, которые максимизируют **целевую функцию** во времени. Это примерно соответствует представлению авторов настоящей книги об искусственном интеллекте как о проектировании систем, которые действуют оптимальным образом. Почему же в таком случае искусственный интеллект и теория управления рассматриваются как две разные научные области, особенно если учесть, какие тесные взаимоотношения связывали их основателей? Ответ на этот вопрос состоит в том, что существует также тесная связь между математическими методами, которые были знакомы участникам этих разработок, и соответствующими множествами задач, которые были охвачены в каждом из этих подходов к описанию мира. Дифференциальное и интегральное исчисление, а также алгебра матриц, являющиеся инструментами теории управления, в наибольшей степени подходят для анализа систем, которые могут быть описаны с помощью фиксированных множеств непрерывно изменяющихся переменных; более того, точный анализ, как правило, осуществим только для линейных систем. Искусственный интеллект был отчасти основан как способ избежать ограничений математических средств, применявшихся в теории управления в 1950-х годах. Такие инструменты, как логический вывод и вычисления, позволили исследователям искусственного интеллекта успешно рассматривать некоторые проблемы (например, понимание естественного языка, зрение и планирование), полностью выходящие за рамки исследований, предпринимавшихся теоретиками управления.

Лингвистика (период с 1957 года по настоящее время)

- Каким образом язык связан с мышлением?

В 1957 году Б.Ф. Скиннер опубликовал свою книгу *Verbal Behavior*. Это был всеобъемлющий, подробный отчет о результатах исследований по изучению языка, проведенных в рамках бихевиористского подхода, который был написан наиболее

выдающимся экспертом в этой области. Но весьма любопытно то, что рецензия к этой книге стала не менее известной, чем сама книга, и послужила причиной почти полного исчезновения интереса к бихевиоризму. Автором этой рецензии был Ноам Хомский, который сам только что опубликовал книгу с изложением своей собственной теории, *Syntactic Structures*. Хомский показал, что бихевиористская теория не позволяет понять истоки творческой деятельности, осуществляемой с помощью языка, — она не объясняет, почему ребенок способен понимать и складывать предложения, которые он до сих пор никогда еще не слышал. Теория Хомского, основанная на синтаксических моделях, восходящих к работам древнеиндийского лингвиста Панини (примерно 350 год до н.э.), позволяла объяснить этот феномен, и, в отличие от предыдущих теорий, оказалась достаточно формальной для того, чтобы ее можно было реализовать в виде программ.

Таким образом, современная лингвистика и искусственный интеллект, которые “родились” примерно в одно и то же время и продолжают вместе расти, пересекаются в гибридной области, называемой **вычислительной лингвистикой** или **обработкой естественного языка**. Вскоре было обнаружено, что проблема понимания языка является гораздо более сложной, чем это казалось в 1957 году. Для понимания языка требуется понимание предмета и контекста речи, а не только анализ структуры предложений. Это утверждение теперь кажется очевидным, но сам данный факт не был широко признан до 1960-х годов. Основная часть ранних работ в области **представления знаний** (науки о том, как преобразовать знания в такую форму, с которой может оперировать компьютер) была привязана к языку и подпитывалась исследованиями в области лингвистики, которые, в свою очередь, основывались на результатах философского анализа языка, проводившегося в течение многих десятков лет.

1.3. ИСТОРИЯ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

После ознакомления с изложенным выше материалом о предыстории искусственного интеллекта перейдем к изучению процесса развития самого искусственного интеллекта.

Появление предпосылок искусственного интеллекта (период с 1943 года по 1955 год)

Первая работа, которая теперь по общему признанию считается относящейся к искусственному интеллекту, была выполнена Уорреном Мак-Каллоком и Уолтером Питтсом [1017]. Они черпали вдохновение из трех источников: знание основ физиологии и назначения нейронов в мозгу; формальный анализ логики высказываний, взятый из работ Рассела и Уайтхеда; а также теория вычислений Тьюринга. Мак-Каллок и Питтс предложили модель, состоящую из искусственных нейронов, в которой каждый нейрон характеризовался как находящийся во “включенном” или “выключенном” состоянии, а переход во “включенное” состояние происходил в ответ на стимуляцию достаточного количества соседних нейронов. Состояние нейрона рассматривалось как “фактически эквивалентное высказыванию, в котором предлагается адекватное количество стимулов”. Работы этих ученых показали, например, что любая вычислимая функция может быть вычислена с по-

мощью некоторой сети из соединенных нейронов и что все логические связи (“И”, “ИЛИ”, “НЕ” и т.д.) могут быть реализованы с помощью простых сетевых структур. Кроме того, Мак-Каллок и Питтс выдвинули предположение, что сети, структурированные соответствующим образом, способны к обучению. Дональд Хебб [638] продемонстрировал простое правило обновления для модификации количества соединений между нейронами. Предложенное им правило, называемое теперь **правилом хеббовского обучения**, продолжает служить основой для моделей, широко используемых и в наши дни.

Два аспиранта факультета математики Принстонского университета, Марвин Минский и Дин Эдмондс, в 1951 году создали первый сетевой компьютер на основе нейронной сети. В этом компьютере, получившем название Snarc, использовалось 3000 электронных ламп и дополнительный механизм автопилота с бомбардировщика В-24 для моделирования сети из 40 нейронов. Аттестационная комиссия, перед которой Минский защищал диссертацию доктора философии, выразила сомнение в том, может ли работа такого рода рассматриваться как математическая, на что фон Нейман, по словам современников, возразил: “Сегодня — нет, но когда-то будет”. В дальнейшем Минский доказал очень важные теоремы, показывающие, с какими ограничениями должны столкнуться исследования в области нейронных сетей.

Кроме того, можно привести большое количество примеров других ранних работ, которые можно охарактеризовать как относящиеся к искусственному интеллекту, но именно Алан Тьюринг впервые выразил полное представление об искусственном интеллекте в своей статье *Computing Machinery and Intelligence*, которая была опубликована в 1950 году. В этой статье он описал тест Тьюринга, принципы машинного обучения, генетические алгоритмы и обучение с подкреплением.

Рождение искусственного интеллекта (1956 год)

В Принстонском университете проводил свои исследования еще один авторитетный специалист в области искусственного интеллекта, Джон Маккарти. После получения ученой степени Маккарти перешел в Дартмутский колледж, который и стал официальным местом рождения этой области знаний. Маккарти уговорил Марвина Минского, Клода Шеннона и Натаниэля Рочестера, чтобы они помогли ему собрать всех американских исследователей, проявляющих интерес к теории автоматов, нейронным сетям и исследованиям интеллекта. Они организовывали двухмесячный семинар в Дартмуте летом 1956 года. Всего на этом семинаре присутствовали 10 участников, включая Тренчарда Мура из Принстонского университета, Артура Самюэла из компании IBM, а также Рея Соломонова и Оливера Селфриджа из Массачусетского технологического института (Massachusetts Institute of Technology — MIT).

Два исследователя из технологического института Карнеги¹⁴, Аллен Ньюэлл и Герберт Саймон, буквально монополизировали все это представление. Тогда как другие могли лишь поделиться своими идеями и в некоторых случаях показать программы для таких конкретных приложений, как шашки, Ньюэлл и Саймон уже мог-

¹⁴ Теперь это учебное заведение называется Университет Карнеги–Меллона (Carnegie–Mellon University — CMU).

ли продемонстрировать программу, проводящую рассуждения, Logic Theorist (LT)¹⁵, или логик-теоретик, в отношении которой Саймон заявил: “Мы изобрели компьютерную программу, способную мыслить в нечисловых терминах и поэтому решили почтенную проблему о соотношении духа и тела”. Вскоре после этого семинара программа показала свою способность доказать большинство теорем из главы 2 труда Рассела и Уайтхеда *Principia Mathematica*. Сообщали, что Рассел пришел в восторг, когда Саймон показал ему, что эта программа предложила доказательство одной теоремы, более короткое, чем в *Principia*. Редакторы *Journal of Symbolic Logic* оказались менее подверженными эмоциям; они отказались принимать статью, в качестве соавторов которой были указаны Ньюэлл, Саймон и программа Logic Theorist.

Дартмутский семинар не привел к появлению каких-либо новых крупных открытий, но позволил познакомиться всем наиболее важным деятелям в этой научной области. Они, а также их студенты и коллеги из Массачусеттского технологического института, Университета Карнеги–Меллона, Станфордского университета и компании IBM занимали ведущее положение в этой области в течение следующих 20 лет. Возможно, дольше всего сохранившимся результатом данного семинара было соглашение принять новое название для этой области, предложенное Маккарти, — **искусственный интеллект**. Возможно, лучше было бы назвать эту научную область “вычислительная рациональность”, но за ней закрепилось название “искусственный интеллект”.

Анализ предложений по тематике докладов для Дартмутского семинара [1014] позволяет понять, с чем связана необходимость преобразовать искусственный интеллект в отдельную область знаний. Почему нельзя было бы публиковать все работы, выполненные в рамках искусственного интеллекта, под флагом теории управления, или исследования операций, или теории решений, которые в конечном итоге имеют цели, аналогичные искусственному интеллекту? Или почему искусственный интеллект не рассматривается как область математики? Ответом на эти вопросы, во-первых, является то, что искусственный интеллект с самого начала впитал идею моделирования таких человеческих качеств, как творчество, самосовершенствование и использование естественного языка. Эти задачи не рассматриваются ни в одной из указанных областей. Во-вторых, еще одним ответом является методология. Искусственный интеллект — это единственная из перечисленных выше областей, которая, безусловно, является одним из направлений компьютерных наук (хотя в исследовании операций также придается большое значение компьютерному моделированию), кроме того, искусственный интеллект — это единственная область, в которой предпринимаются попытки создания машин, действующих автономно в сложной, изменяющейся среде.

Ранний энтузиазм, большие ожидания (период с 1952 года по 1969 год)

Первые годы развития искусственного интеллекта были полны успехов, хотя и достаточно скромных. Если учесть, какими примитивными были в то время компьютеры и инструментальные средства программирования, и тот факт, что лишь за несколько лет до этого компьютеры рассматривались как устройства, способные вы-

¹⁵ Для написания программы LT Ньюэлл и Саймон разработали также язык обработки списков IPL. У них не было компилятора, поэтому эти ученые транслировали программы на своем языке в машинный код вручную. Чтобы избежать ошибок, они работали параллельно, называя друг другу двоичные числа после записи каждой команды, чтобы убедиться в том, что они совпадают.

полнять только арифметические, а не какие-либо иные действия, можно лишь удивляться тому, как удалось заставить компьютер выполнять операции, хоть немного напоминающие разумные. Интеллектуальное сообщество в своем большинстве продолжало считать, что “ни одна машина не сможет выполнить действие X ”. (Длинный список таких X , собранный Тьюрингом, приведен в главе 26.) Вполне естественно, что исследователи в области искусственного интеллекта отвечали на это, демонстрируя способность решать одну задачу X за другой. Джон Маккарти охарактеризовал этот период как эпоху восклицаний: “Гляди, мама, что я умею!”


За первыми успешными разработками Ньюэлла и Саймона последовало создание программы общего решателя задач (General Problem Solver — GPS). В отличие от программы Logic Theorist, эта программа с самого начала была предназначена для моделирования процедуры решения задач человеком. Как оказалось, в пределах того ограниченного класса головоломок, которые была способна решать эта программа, порядок, в котором она рассматривала подцели и возможные действия, был аналогичен тому подходу, который применяется людьми для решения таких же проблем. Поэтому программа GPS была, по-видимому, самой первой программой, в которой был воплощен подход к “организации мышления по такому же принципу, как и у человека”. Результаты успешного применения GPS и последующих программ в качестве модели познания позволили сформулировать знаменитую гипотезу **физической символической системы** ([1131]), в которой утверждается, что существует “физическая символическая система, которая имеет необходимые и достаточные средства для интеллектуальных действий общего вида”. Под этим подразумевается, что любая система, проявляющая интеллект (человек или машина), должна действовать по принципу манипулирования структурами данных, состоящими из символов. Ниже будет показано, что эта гипотеза во многих отношениях оказалась уязвимой для критики.

Работая в компании IBM, Натаниэль Rochester и его коллеги создали некоторые из самых первых программ искусственного интеллекта. Герберт Гелернтер [532] сконструировал программу Geometry Theorem Prover (программа автоматического доказательства геометрических теорем), которая была способна доказывать такие теоремы, которые показались бы весьма сложными многим студентам-математикам. Начиная с 1952 года Артур Самюэл написал ряд программ для игры в шашки, которые в конечном итоге научились играть на уровне хорошо подготовленного любителя. В ходе этих исследований Самюэл опроверг утверждение, что компьютеры способны выполнять только то, чему их учили: одна из его программ быстро научилась играть лучше, чем ее создатель. Эта программа была продемонстрирована по телевидению в феврале 1956 года и произвела очень сильное впечатление на зрителей. Как и Тьюринг, Самюэл с трудом находил машинное время. Работая по ночам, он использовал компьютеры, которые все еще находились на испытательной площадке производственного предприятия компании IBM. Проблема ведения игр рассматривается в главе 6, а в главе 21 описаны и дополнены методы обучения, которые использовались Самюэлом.

Джон Маккарти перешел из Дартмутского университета в Массачусетский технологический институт и здесь в течение одного исторического 1958 года внес три крайне важных вклада в развитие искусственного интеллекта. В документе *MIT AI Lab Memo No. 1* Джон Маккарти привел определение нового языка высокого уровня **Lisp**, которому суждено было стать доминирующим языком программирования для искусственного интеллекта. Lisp остается одним из главных языков высокого

уровня, применяемых в настоящее время, будучи вместе с тем вторым по очередности появления языком такого типа, который был создан всего на один год позже чем Fortran. Разработав язык Lisp, Маккарти получил необходимый для него инструмент, но доступ к ограниченным и дорогостоящим компьютерным ресурсам продолжал оставаться серьезной проблемой. В связи с этим он совместно с другими сотрудниками Массачусеттского технологического института изобрел режим разделения времени. В том же 1958 году Маккарти опубликовал статью под названием *Programs with Common Sense*, в которой он описал гипотетическую программу Advice Taker, которая может рассматриваться как первая полная система искусственного интеллекта. Как и программы Logic Theorist и Geometry Theorem Prover, данная программа Маккарти была предназначена для использования знаний при поиске решений задач. Но в отличие от других программ она была предназначена для включения общих знаний о мире. Например, Маккарти показал, что некоторые простые аксиомы позволяют этой программе разработать план оптимального маршрута автомобильной поездки в аэропорт, чтобы можно было успеть на самолет. Данная программа была также спроектирована таким образом, что могла принимать новые аксиомы в ходе обычной работы, а это позволяло ей приобретать компетентность в новых областях без перепрограммирования. Таким образом, в программе Advice Taker были воплощены центральные принципы представления знаний и проведения рассуждений, которые заключаются в том, что всегда полезно иметь формальное, явное представление о мире, а также о том, как действия агента влияют на этот мир и как приобрести способность манипулировать подобными представлениями с помощью дедуктивных процессов. Замечательной особенностью указанной статьи, которая вышла в 1958 году, является то, что значительная ее часть не потеряла своего значения и в наши дни.

Знаменитый 1958 год отмечен также тем, что именно в этот год Марвин Минский перешел в Массачусеттский технологический институт. Но успешно складывавшееся на первых порах его сотрудничество с Маккарти продолжалось недолго. Маккарти настаивал на том, что нужно изучать способы представления и проведения рассуждений в формальной логике, тогда как Минский в большей степени интересовался тем, как довести программы до рабочего состояния, и в конечном итоге у него сформировалось отрицательное отношение к логике. В 1963 году Маккарти открыл лабораторию искусственного интеллекта в Станфордском университете. Разработанный им план использования логики для создания окончательной версии программы Advice Taker выполнялся еще быстрее, чем было задумано, благодаря открытию Дж.А. Робинсоном метода резолюции (полного алгоритма доказательства теорем для логики первого порядка; см. главу 9). Работы, выполненные в Станфордском университете, подчеркнули важность применения методов общего назначения для проведения логических рассуждений. В число логических приложений вошли системы формирования ответов на вопросы и планирования Корделла Грина [592], а также робототехнический проект Shakey, разрабатываемый в новом Станфордском научно-исследовательском институте (Stanford Research Institute — SRI). Последний проект, который подробно рассматривается в главе 25, впервые продемонстрировал полную интеграцию логических рассуждений и физической активности.

Минский руководил работой ряда студентов, выбравших для себя задачи ограниченных масштабов, для решения которых, как в то время казалось, требовалась интеллектуальность. Эти ограниченные проблемные области получили название  микромиров.

Программа Saint Джеймса Слэгла [1426] оказалась способной решать задачи интеграции в исчислении замкнутой формы, типичные для первых курсов колледжей. Программа Analogy Тома Эванса [448] решала задачи выявления геометрических аналогий, применяемые при проверке показателя интеллекта, аналогичные приведенной на рис. 1.2. Программа Student Дэниэла Боброва [142] решала изложенные в виде рассказа алгебраические задачи, подобные приведенной ниже.

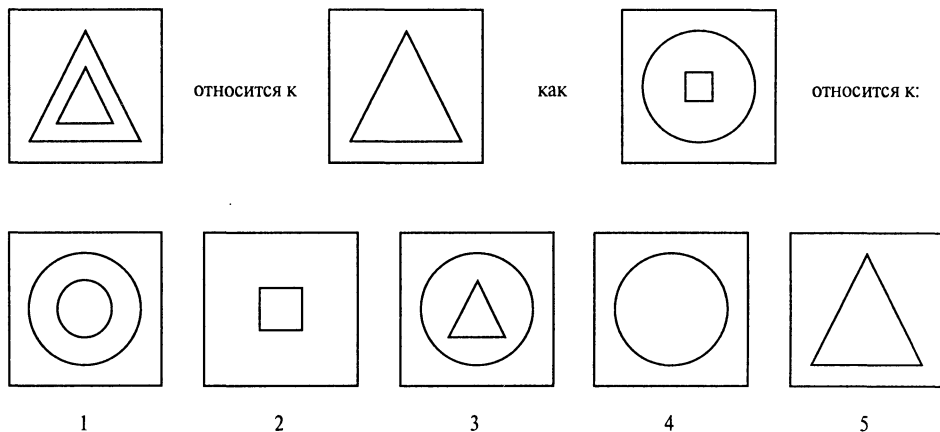


Рис. 1.2. Пример задачи, решаемой программой Analogy Эванса

Если количество заказов, полученных Томом, вдвое превышает квадратный корень из 20% опубликованных им рекламных объявлений, а количество этих рекламных объявлений равно 45, то каково количество заказов, полученных Томом?

Наиболее известным примером микромира был мир блоков, состоящий из множества цельных блоков, размещенных на поверхности стола (или, что более часто, на имитации стола), как показано на рис. 1.3. Типичной задачей в этом мире является изменение расположения блоков определенным образом с использованием манипулятора робота, который может захватывать по одному блоку одновременно. Мир блоков стал основой для проекта системы технического зрения Дэвида Хаффмена [702], работы по изучению зрения и распространения (удовлетворения) ограничений Дэвида Уолтса [1552], теории обучения Патрика Уинстона [1602], программы понимания естественного языка Тэрри Винограда [1601] и планировщика в мире блоков Скотта Фалмана [450].

Бурно продвигались также исследования, основанные на ранних работах по созданию нейронных сетей Мак-Каллока и Питтса. В работе Винограда и Коуэна [1600] было показано, как нужно представить отдельную концепцию с помощью коллекции, состоящей из большого количества элементов, соответственно увеличивая надежность и степень распараллеливания их работы. Методы обучения Хебба были усовершенствованы в работах Берни Видроу [1587], [1586], который называл свои сети **адалинами**, а также Френка Розенблатта [1304], создателя **перцептронов**. Розенблатт доказал **теорему сходимости перцептрона**, которая подтверждает, что предложенный им алгоритм обучения позволяет корректировать количество соединений перцептрона в соответствии с любыми входными данными, при условии, что такое соответствие существует. Эта тема рассматривается в главе 20.

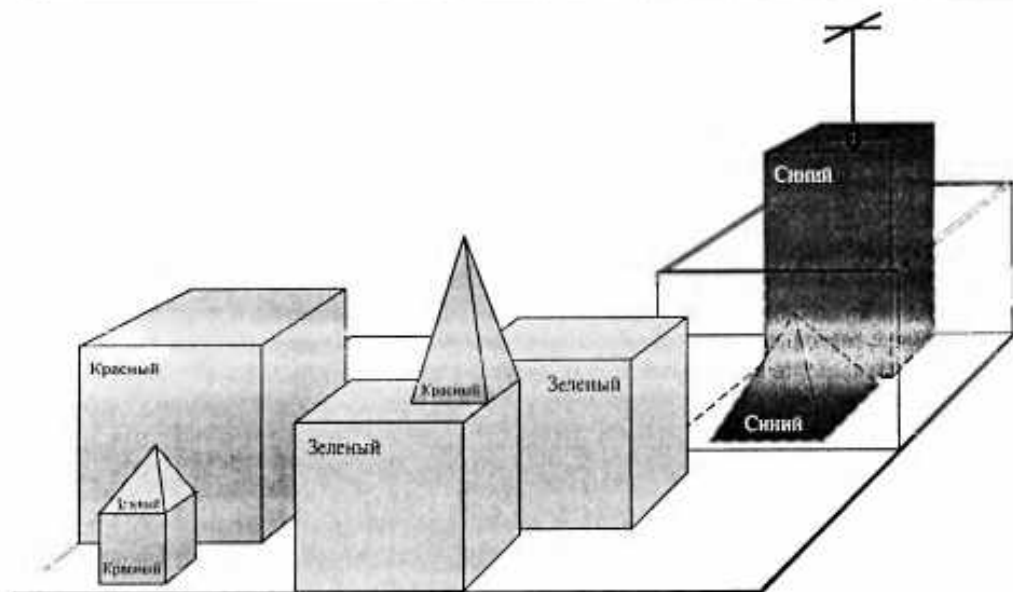


Рис. 1.3. Сцена из мира блоков. Программа Shrdlu [1601] только что завершила выполнение команды "Найти блок, более высокий по сравнению с тем, который находится в манипуляторе, и поместить его в ящик"

Столкновение с реальностью (период с 1966 года по 1973 год)

С самого начала исследователи искусственного интеллекта не отличались сдержанностью, высказывая прогнозы в отношении своих будущих успехов. Например, часто цитировалась приведенное ниже предсказание Герберта Саймона, опубликованное им в 1957 году.

Я не ставлю перед собой задачу удивить или шокировать вас, но проще всего я могу подвести итог, сказав, что теперь мы живем в таком мире, где машины могут думать, учиться и создавать. Более того, их способность выполнять эти действия будет продолжать расти до тех пор, пока (в обозримом будущем) круг проблем, с которыми смогут справиться машины, будет сопоставим с тем кругом проблем, где до сих пор нужен человеческий мозг.

Такие выражения, как "обозримое будущее", могут интерпретироваться по-разному, но Саймон сделал также более конкретный прогноз, что через десять лет компьютер станет чемпионом мира по шахматам и что машиной будут доказаны все важные математические теоремы. Эти предсказания сбылись (или почти сбылись) не через десять лет, а через сорок. Чрезмерный оптимизм Саймона был обусловлен тем, что первые системы искусственного интеллекта демонстрировали многообещающую производительность, хотя и на простых примерах. Но почти во всех случаях эти ранние системы терпели сокрушительное поражение, сталкиваясь с более широким кругом проблем или с более трудными проблемами.

Сложности первого рода были связаны с тем, что основная часть ранних программ не содержала знаний или имела лишь небольшой объем знаний о своей предметной области; их временные успехи достигались за счет простых синтаксических манипуляций. Типичная для этого периода история произошла при проведении

первых работ по машинному переводу текста на естественном языке, которые щедро финансировались Национальным научно-исследовательским советом США (U.S. National Research Council) в попытке ускорить перевод советских научных статей во время того периода бурной деятельности, который начался вслед за запуском в СССР первого искусственного спутника Земли в 1957 году. Вначале считалось, что для сохранения точного смысла предложений достаточно провести простые синтаксические преобразования, основанные на грамматиках русского и английского языков, и замену слов с использованием электронного словаря. Но дело в том, что для устранения неоднозначности и определения смысла предложения в процессе перевода необходимо обладать общими знаниями о предметной области. Возникающие при этом сложности иллюстрируются знаменитым обратным переводом фразы “the spirit is willing but the flesh is weak” (дух полон желаний, но плоть слаба), в результате которого получилось следующее: “the vodka is good but the meat is rotten” (водка хороша, но мясо испорчено). В 1966 году в отчете одного консультативного комитета было отмечено, что “машинный перевод научного текста общего характера не осуществлен и не будет осуществлен в ближайшей перспективе”. Все финансирование академических проектов машинного перевода правительством США было свернуто. В настоящее время машинный перевод является несовершенным, но широко применяемым инструментальным средством обработки технических, коммерческих, правительственных документов, а также документов, опубликованных в Internet.

Сложности второго рода были связаны с неразрешимостью многих проблем, решение которых пытались найти с помощью искусственного интеллекта. В большинстве ранних программ искусственного интеллекта решение задач осуществлялось по принципу проверки различных комбинаций возможных шагов, которая проводилась до тех пор, пока не будет найдено решение. На первых порах такая стратегия приводила к успеху, поскольку микромиры содержали очень небольшое количество объектов, поэтому предусматривали лишь незначительный перечень возможных действий и позволяли находить очень короткие последовательности решения. До того как была разработана теория вычислительной сложности, было широко распространено такое мнение, что для “масштабирования” задач до уровня более крупных проблем достаточно просто применить более быстродействующие аппаратные средства с большим объемом памяти. Например, оптимизм, с которым были встречены сообщения о разработке метода доказательства теорем с помощью резолюции, быстро угас, когда исследователи не смогли доказать таким образом теоремы, которые включали чуть больше нескольких десятков фактов. Как оказалось, *то, что программа может найти решение в принципе, не означает, что эта программа действительно содержит все механизмы, позволяющие найти данное решение на практике.*

Иллюзия неограниченной вычислительной мощи распространялась не только на программы решения задач. Ранние эксперименты в области *эволюции машин* (которая теперь известна под названием **разработка генетических алгоритмов**) [502], [503] были основаны на уверенности в том, что внесение соответствующего ряда небольших изменений в машинный код программы позволяет создать программу решения любой конкретной простой задачи, обладающую высокой производительностью. Безусловно, что сам этот подход является вполне обоснованным. Поэтому общая идея состояла в том, что необходимо проверять случайные мутации (изменения в коде) с помощью процесса отбора для сохранения мутаций, которые кажутся полезными. На эти эксперименты было потрачено тысячи часов процессорного време-

ни, но никаких признаков прогресса не было обнаружено. В современных генетических алгоритмах используются лучшие способы представления, которые показывают более успешные результаты.

Одним из основных критических замечаний в адрес искусственного интеллекта, содержащихся в отчете Лайтхилла [930], который лег в основу решения британского правительства прекратить поддержку исследований в области искусственного интеллекта во всех университетах, кроме двух, была неспособность справиться с “комбинаторным взрывом” — стремительным увеличением сложности задачи. (Это — официальная версия событий, а в устном изложении рисуется немного иная и более красочная картина, в которой проявляются политические амбиции и личные интересы, описание которых выходит за рамки данного изложения.)

Сложности третьего рода возникли в связи с некоторыми фундаментальными ограничениями базовых структур, которые использовались для выработки интеллектуального поведения. Например, в книге Минского и Пейперта *Perceptrons* [1054] было доказано, что перцептроны (простая форма нейронной сети) могут продемонстрировать способность изучить все, что возможно представить с их помощью, но, к сожалению, они позволяют представить лишь очень немногое. В частности, перцептрон с двумя входами нельзя обучить распознаванию такой ситуации, при которой на два его входа подаются разные сигналы. Хотя полученные этими учеными результаты не распространяются на более сложные, многослойные сети, вскоре было обнаружено, что финансы, выделенные на поддержку исследований в области нейронных сетей, почти не приносят никакой отдачи. Любопытно отметить, что новые алгоритмы обучения путем обратного распространения для многослойных сетей, которые стали причиной возрождения необычайного интереса к исследованиям в области нейронных сетей в конце 1980-х годов, фактически были впервые открыты в 1969 году [201].

Системы, основанные на знаниях: могут ли они стать ключом к успеху (период с 1969 года по 1979 год)

Основной подход к решению задач, сформированный в течение первого десятилетия исследований в области искусственного интеллекта, представлял собой механизм поиска общего назначения, с помощью которого предпринимались попытки связать в единую цепочку элементарные этапы проведения рассуждений для формирования полных решений. Подобные подходы получили название **слабых методов**, поскольку они не позволяли увеличить масштабы своего применения до уровня более крупных или более сложных экземпляров задач, несмотря на то, что были общими. Альтернативным по сравнению со слабыми методами стал подход, предусматривающий использование более содержательных знаний, относящихся к проблемной области, который позволяет создавать более длинные цепочки шагов логического вывода и дает возможность проще справиться с теми проблемными ситуациями, которые обычно возникают в специализированных областях знаний. Как известно, чтобы решить достаточно сложную задачу, необходимо уже почти полностью знать ответ.

Одним из первых примеров реализации такого подхода была программа Dendral [205]. Она была разработана в Станфордском университете группой ученых, в которую вошли Эд Фейгенбаум (бывший студент Герберта Саймона), Брюс Бьюкенен

(философ, который сменил специальность и стал заниматься компьютерными науками) и Джошуа Ледерберг (лауреат Нобелевской премии в области генетики). Эта группа занималась решением проблемы определения структуры молекул на основе информации, полученной от масс-спектрометра. Вход этой программы состоял из химической формулы соединения (например, $C_6H_{13}NO_2$) и спектра масс, позволяющего определять массы различных фрагментов молекулы, который формировался при бомбардировке молекулы потоком электронов. Например, спектр масс может содержать пик в точке $m=15$, соответствующий массе метилового фрагмента (CH_3).

Первая, примитивная версия этой программы предусматривала выработку всех возможных структур, совместимых с данной формулой, после чего предсказывала, какой спектр масс должен наблюдаться для каждой из этих структур, сравнивая его с фактическим спектром. Вполне можно ожидать, что такая задача применительно к молекулам более крупных размеров становится неразрешимой. Поэтому разработчики программы Dendral проконсультировались с химиками-аналитиками и пришли к выводу, что следует попытаться организовать работу по принципу поиска широко известных картин расположения пиков в спектре, которые указывают на наличие общих подструктур в молекуле. Например, для распознавания кетоновых подгрупп ($C=O$) с атомными весами 28 может использоваться приведенное ниже правило.

if имеются два пика в точках x_1 и x_2 , такие, что:

- а) $x_1 + x_2 = M + 28$ (где M — масса всей молекулы);
- б) в точке $x_1 - 28$ — высокий пик;
- в) в точке $x_2 - 28$ — высокий пик;
- г) по меньшей мере в одной из точек x_1 и x_2 — высокий пик,

then существует кетоновая подгруппа.

Применение способа, предусматривающего распознавание того, что молекула содержит какие-то конкретные подструктуры, позволило весьма значительно сократить количество возможных кандидатов, подлежащих проверке. В конечном итоге программа Dendral оказалась очень мощной, и причины этого описаны ниже.

Все относящиеся к делу теоретические знания, требуемые для решения указанных проблем, были преобразованы в [компоненте предсказания спектра] из наиболее общей формы (из “исходных принципов”) в эффективные специальные формы (в “рецепты поваренной книги”) [458].


Значение программы Dendral состояло в том, что это была первая успешно созданная экспертная система, основанная на широком использовании знаний: ее способность справляться с поставленными задачами была обусловлена применением большого количества правил специального назначения. В более поздних системах также широко применялся основной принцип подхода, реализованного Маккарти в программе Advice Taker, — четкое отделение знаний (в форме правил) от компонента, обеспечивающего проведение рассуждений.

Руководствуясь этим опытом, Фейгенбаум и другие специалисты из Станфордского университета приступили к разработке проекта эвристического программирования (Heuristic Programming Project — HPP), целью которого было исследование того, в какой степени созданная ими новая методология ~~э~~ экспертных систем может быть применена в других областях интеллектуальной деятельности человека. На очередном этапе основные усилия были сосредоточены в области медицинской диагностики. Фейгенбаум, Бьюкенен и доктор Эдвард Шортлифф разработали

программу Musin для диагностики инфекционных заболеваний кровеносной системы. После ввода в нее примерно 450 правил программа Musin приобрела способность работать на уровне некоторых экспертов, а также показала значительно более лучшие результаты по сравнению с врачами, имеющими не такой большой стаж. Она также обладала двумя важными отличительными особенностями по сравнению с программой Dendral. Во-первых, в отличие от правил Dendral не существовала общая теоретическая модель, на основании которой мог бы осуществляться логический вывод правил Musin. Для выявления этих правил приходилось широко применять знания, полученные от экспертов, которые, в свою очередь, приобретали эти знания с помощью учебников, других экспертов и непосредственного опыта, накопленного путем изучения практических случаев. Во-вторых, в этих правилах приходилось учитывать ту степень неопределенности, которой характеризуются знания в области медицины. В программе Musin применялось исчисление неопределенностей на основе так называемых **коэффициентов уверенности** (см. главу 13), которое (в то время) казалось вполне соответствующим тому, как врачи оценивают влияние объективных данных на диагноз.

Важность использования знаний в проблемной области стала также очевидной и для специалистов, которые занимались проблемами понимания естественного языка. Хотя система понимания естественного языка Shrdlu, разработанная Тэрри Виноградом, стала в свое время предметом всеобщего восхищения, ее зависимость от результатов синтаксического анализа вызвала появление примерно таких же проблем, которые обнаружились в ранних работах по машинному переводу. Эта система была способна преодолеть неоднозначность и правильно понимала ссылки, выраженные с помощью местоимений, но это в основном было связано с тем, что она специально предназначалась только для одной области — для мира блоков. Некоторые исследователи, включая Юджина Чарняка, коллегу и аспиранта Винограда в Массачусеттском технологическом институте, указывали, что для обеспечения надежного понимания языка потребуются общие знания о мире и общий метод использования этих знаний.

Работавший в Йельском университете Роджер Шенк, лингвист, ставший исследователем в области искусственного интеллекта, еще более ярко выразил эту мысль, заявив, что “такого понятия, как синтаксис, не существует”. Это заявление вызвало возмущение многих лингвистов, но послужило началом полезной дискуссии. Шенк со своими студентами создал ряд интересных программ [425], [1358], [1359], [1590]. Задача всех этих программ состояла в обеспечении понимания естественного языка. Но в них основной акцент был сделан в меньшей степени на языке как таковом и в большей степени на проблемах представления и формирования рассуждений с помощью знаний, требуемых для понимания языка. В число рассматриваемых проблем входило представление стереотипных ситуаций [314], описание организации человеческой памяти [829], [1287], а также понимание планов и целей [1591].

В связи с широким ростом количества приложений, предназначенных для решения проблем реального мира, столь же широко возрастали потребности в создании работоспособных схем представления знаний. Было разработано большое количество различных языков для представления знаний и проведения рассуждений. Некоторые из них были основаны на логике, например, в Европе получил распространение язык Prolog, а в Соединенных Штатах широко применялось семейство языков Planner. В других языках, основанных на выдвинутой Минским идее  **Фреймов**

[1053], был принят более структурированный подход, предусматривающий сбор фактов о конкретных типах объектов и событий, а также упорядочение этих типов в виде крупной таксономической иерархии, аналогичной биологической таксономии.

Превращение искусственного интеллекта в индустрию (период с 1980 года по настоящее время)

Первая успешно действующая коммерческая экспертная система, R1, появилась в компании DEC (Digital Equipment Corporation) [1026]. Эта программа помогала составлять конфигурации для выполнения заказов на новые компьютерные системы; к 1986 году она позволяла компании DEC экономить примерно 40 миллионов долларов в год. К 1988 году группой искусственного интеллекта компании DEC было развернуто 40 экспертных систем, а в планах дальнейшего развертывания было предусмотрено еще большее количество таких систем. В компании Du Pont применялось 100 систем, в разработке находилось еще 500, а достигнутая экономия составляла примерно 10 миллионов долларов в год. Почти в каждой крупной корпорации США была создана собственная группа искусственного интеллекта и либо применялись экспертные системы, либо проводились их исследования.

В 1981 году в Японии было объявлено о развертывании проекта создания компьютера “пятого поколения” — 10-летнего плана по разработке интеллектуальных компьютеров, работающих под управлением языка Prolog. В ответ на это в Соединенных Штатах была сформирована корпорация Microelectronics and Computer Technology Corporation (MCC) как научно-исследовательский консорциум, предназначенный для обеспечения конкурентоспособности американской промышленности. И в том и в другом случае искусственный интеллект стал частью общего плана, включая его применение для проектирования микросхем и проведения исследований в области человеко-машинного интерфейса. Но амбициозные цели, поставленные перед специалистами в области искусственного интеллекта в проектах MCC и компьютеров пятого поколения, так и не были достигнуты. Тем не менее в Британии был выпущен отчет Олви (Alvey)¹⁶, в котором предусматривалось возобновление финансирования, урезанного на основании отчета Лайтхилла.

В целом в индустрии искусственного интеллекта произошел бурный рост, начиная с нескольких миллионов долларов в 1980 году и заканчивая миллиардами долларов в 1988 году. Однако вскоре после этого наступил период, получивший название “зимы искусственного интеллекта”, в течение которого пострадали многие компании, поскольку не сумели выполнить своих заманчивых обещаний.

Возвращение к нейронным сетям (период с 1986 года по настоящее время)

Хотя основная часть специалистов по компьютерным наукам прекратила исследования в области нейронных сетей в конце 1970-х годов, работу в этой области продолжили специалисты из других научных направлений. Такие физики, как Джон Хопфилд [674], использовали методы из статистической механики для анализа

¹⁶ Чтобы не ставить себя в затруднительное положение, авторы этого отчета изобрели новую научную область, получившую название “интеллектуальные системы, основанные на знаниях” (Intelligent Knowledge-Based Systems — IKBS), поскольку термин “искусственный интеллект” был уже официально отменен.

свойств хранения данных и оптимизации сетей, рассматривая коллекции узлов как коллекции атомов. Психологи, включая Дэвида Румельхарта и Джефа Хинтона, продолжали исследовать модели памяти на основе нейронных сетей. Как будет описано в главе 20, настоящий прорыв произошел в середине 1980-х годов, когда по меньшей мере четыре разные группы снова открыли алгоритм обучения путем обратного распространения, впервые предложенный в 1969 году Брайсоном и Хо [201]. Этот алгоритм был применен для решения многих проблем обучения в компьютерных науках и психологии, а после публикации результатов его использования в сборнике статей *Parallel Distributed Processing* [1318] всеобщее внимание привлек тот факт, насколько разнообразными оказались области его применения.

Эти так называемые **коннекционистские** (основанные на соединениях) модели интеллектуальных систем многими рассматривались как непосредственно конкурирующие и с символическими моделями, разрабатываемыми Ньюэллом и Саймоном, и с логицистским подходом, предложенным Маккарти и другими [1442]. Повидимому, не следует отрицать, что на некотором уровне мышления люди манипулируют символами; и действительно, в книге Терренса Дикона под названием *The Symbolic Species* [354] указано, что способность манипулировать символами — определяющая характеристика человека, но наиболее горячие сторонники коннекционизма поставили под сомнение то, что на основании манипулирования символами действительно можно полностью объяснить какие-то познавательные процессы в подробных моделях познания. Вопрос остается открытым, но современный взгляд на эту проблему состоит в том, что коннекционистский и символический подходы являются взаимодополняющими, а не конкурирующими.

Преобразование искусственного интеллекта в науку (период с 1987 года по настоящее время)

В последние годы произошла буквально революция как в содержании, так и в методологии работ в области искусственного интеллекта¹⁷. В настоящее время гораздо чаще встречаются работы, которые основаны на существующих теориях, а не содержат описания принципиально новых открытий; утверждения, изложенные в этих работах, основаны на строгих теоремах или надежных экспериментальных свидетельствах, а не на интуиции; при этом обоснованность сделанных выводов подтверждается на реальных практических приложениях, а не на игрушечных примерах.

Появление искусственного интеллекта отчасти стало результатом усилий по преодолению ограничений таких существующих научных областей, как теория управления и статистика, но теперь искусственный интеллект включил в себя и эти области. В одной из своих работ Дэвид Макаллестер [1006] выразил эту мысль следующим образом.

В ранний период развития искусственного интеллекта казалось вероятным, что в результате появления новых форм символических вычислений, например фреймов и семантиче-

¹⁷ Некоторые охарактеризовали эту смену подходов как победу **теоретиков** (тех, кто считает, что теории искусственного интеллекта должны быть основаны на строгих математических принципах) над **экспериментаторами** (теми, кто предпочитает проверить множество идей, написать какие-то программы, а затем оценить те из них, которые кажутся работоспособными). Оба подхода являются важными. А смещение акцентов в пользу теоретической обоснованности свидетельствует о том, что данная область достигла определенного уровня стабильности и зрелости. Будет ли когда-либо такая стабильность нарушена новой идеей, родившейся в экспериментах, — это другой вопрос.

ских сетей, основная часть классической теории станет устаревшей. Это привело к определенной форме самоизоляции, характеризовавшейся тем, что искусственный интеллект в значительной степени отделился от остальной части компьютерных наук. В настоящее время такой изоляционизм преодолен. Появилось признание того, что машинное обучение не следует отделять от теории информации, что проведение рассуждений в условиях неопределенности нельзя изолировать от стохастического моделирования, что поиск не следует рассматривать отдельно от классической оптимизации и управления и что автоматизированное формирование рассуждений не должно трактоваться как независимое от формальных методов и статистического анализа.

С точки зрения методологии искусственный интеллект наконец-то твердо перешел на научные методы. Теперь, для того чтобы быть принятыми, гипотезы должны подвергаться проверке в строгих практических экспериментах, а значимость результатов должна подтверждаться данными статистического анализа [275]. Кроме того, в настоящее время имеется возможность воспроизводить эксперименты с помощью Internet, а также совместно используемых репозитариев тестовых данных и кода.

Именно по этому принципу развивается область распознавания речи. В 1970-е годы было опробовано широкое разнообразие различных архитектур и подходов. Многие из них оказались довольно надуманными и недолговечными и были продемонстрированы только на нескольких специально выбранных примерах. В последние годы доминирующее положение в этой области заняли подходы, основанные на использовании **скрытых марковских моделей** (Hidden Markov Model — HMM). Описанное выше современное состояние искусственного интеллекта подтверждается двумя особенностями моделей HMM. Во-первых, они основаны на строгой математической теории. Это позволяет исследователям речи использовать в своей работе математические результаты, накопленные в других областях за несколько десятилетий. Во-вторых, они получены в процессе обучения программ на крупном массиве реальных речевых данных. Это гарантирует обеспечение надежных показателей производительности, а в строгих слепых испытаниях модели HMM неизменно улучшают свои показатели. Технология распознавания речи и связанная с ней область распознавания рукописных символов уже совершают переход к созданию широко применяемых индустриальных и потребительских приложений.

Нейронные сети также следуют этой тенденции. Основная часть работ по нейронным сетям, осуществленных в 1980-х годах, была проведена в попытке оценить масштабы того, что должно быть сделано, а также понять, в чем нейронные сети отличаются от “традиционных” методов. В результате использования усовершенствованной методологии и теоретических основ исследователи в этой области достигли такого уровня понимания, что теперь нейронные сети стали сопоставимыми с соответствующими технологиями из области статистики, распознавания образов и машинного обучения, а наиболее перспективная методология может быть применена к каждому из этих приложений. В результате этих разработок была создана так называемая технология **анализа скрытых закономерностей в данных** (data mining), которая легла в основу новой, быстро растущей отрасли информационной индустрии.

Знакомство широких кругов специалистов с книгой Джуди Перла *Probabilistic Reasoning in Intelligent Systems* [1191] привело к признанию важности теории вероятностей и теории решений для искусственного интеллекта, что последовало за возрождением интереса к этой теме, вызванной статьей Питера Чизмана *In Defense of Probability* [242]. Для обеспечения эффективного представления неопределенных

знаний и проведения на их основе строгих рассуждений были разработаны формальные средства **байесовских сетей**. Этот подход позволил преодолеть многие проблемы систем вероятностных рассуждений, возникавшие в 1960–1970-х гг.; теперь он стал доминирующим в таких направлениях исследований искусственного интеллекта, как формирование рассуждений в условиях неопределенности и экспертные системы. Данный подход позволяет организовать обучение на основе опыта и сочетает в себе лучшие достижения классического искусственного интеллекта и нейронных сетей. В работах Джуди Перла [1186], а также Эрика Горвица и Дэвида Хекермана [688], [689] была развита идея *нормативных экспертных систем*. Таковыми являются системы, которые действуют рационально, в соответствии с законами теории решений, а не пытаются имитировать мыслительные этапы в работе людей-экспертов. Операционная система Windows™ включает несколько нормативных диагностических экспертных систем, применяемых для устранения нарушений в работе. Эта область рассматривается в главах 13–16.

Аналогичные бескровные революции произошли в области робототехники, компьютерного зрения и представления знаний. Благодаря лучшему пониманию исследовательских задач и свойств, обуславливающих их сложность, в сочетании с всевозрастающим усложнением математического аппарата, удалось добиться формирования реальных планов научных исследований и перейти к использованию более надежных методов. Но во многих случаях формализация и специализация привели также к фрагментации направлений, например, такие темы, как машинное зрение и робототехника, все больше отделяются от “основного направления” работ по искусственному интеллекту. Снова добиться объединения этих разрозненных областей можно на основе единого взгляда на искусственный интеллект как науку проектирования рациональных агентов.

Появление подхода, основанного на использовании интеллектуальных агентов (период с 1995 года по настоящее время)

Вдохновленные успехами в решении указанных проблем искусственного интеллекта, исследователи также вновь приступили к решению проблемы “целостного агента”. Наиболее широко известным примером создания полной архитектуры агента является работа Аллена Ньюэлла, Джона Лэрда и Пола Розенблума [880], [1125] над проектом Soar. Для того чтобы проще было разобраться в работе агентов, внедренных в реальную среду с непрерывным потоком сенсорных входных данных, были применены так называемые *ситуационные движения*. Одним из наиболее важных примеров среды для интеллектуальных агентов может служить Internet. Системы искусственного интеллекта стали настолько распространенными в приложениях для Web, что суффикс “-бот” (сокращение от робот) вошел в повседневный язык. Более того, технологии искусственного интеллекта легли в основу многих инструментальных средств Internet, таких как машины поиска, системы, предназначенные для выработки рекомендаций, и системы создания Web-узлов.

Пересмотру с учетом нового представления о роли агентов было подвергнуто не только первое издание данной книги [1328], но и другие новейшие труды по этой теме [1146], [1227]. Одним из следствий попыток создания полных агентов стало понимание того, что ранее изолированные подобласти искусственного интеллекта могут потребовать определенной реорганизации, когда возникнет необходимость

снова связать воедино накопленные в них результаты. В частности, теперь широко признано, что сенсорные системы (системы машинного зрения, эхолокации, распознавания речи и т.д.) не способны предоставить абсолютно надежную информацию о среде. Поэтому системы проведения рассуждений и планирования должны быть приспособленными к работе в условиях неопределенности. Вторым важным следствием изменения взглядов на роль агентов является то, что исследования в области искусственного интеллекта теперь необходимо проводить в более тесном контакте с другими областями, такими как теория управления и экономика, которые также имеют дело с агентами.

1.4. СОВРЕМЕННОЕ СОСТОЯНИЕ РАЗРАБОТОК

Какие возможности предоставляет искусственный интеллект в наши дни? Краткий ответ на этот вопрос сформулировать сложно, поскольку в этом научном направлении существует слишком много подобластей, в которых выполняется очень много исследований. Ниже в качестве примеров перечислено лишь несколько приложений; другие будут указаны в следующих главах.

- **Автономное планирование и составление расписаний.** Работающая на удалении в сотни миллионов километров от Земли программа Remote Agent агентства NASA стала первой бортовой автономной программой планирования, предназначенной для управления процессами составления расписания операций для космического аппарата [744]. Программа Remote Agent вырабатывала планы на основе целей высокого уровня, задаваемых с Земли, а также контролировала работу космического аппарата в ходе выполнения планов: обнаруживала, диагностировала и устраняла неполадки по мере их возникновения.
- **Ведение игр.** Программа Deep Blue компании IBM стала первой компьютерной программой, которой удалось победить чемпиона мира в шахматном матче, после того как она обыграла Гарри Каспарова со счетом 3,5:2,5 в показательном матче [577]. Каспаров заявил, что ощущал напротив себя за шахматной доской присутствие “интеллекта нового типа”. Журнал *Newsweek* описал этот матч под заголовком “Последний оборонительный рубеж мозга”. Стоимость акций IBM выросла на 18 миллиардов долларов.
- **Автономное управление.** Система компьютерного зрения Alvinn была обучена вождению автомобиля, придерживаясь определенной полосы движения. В университете CMU эта система была размещена в микроавтобусе, управляемом компьютером NavLab, и использовалось для проезда по Соединенным Штатам; на протяжении 2850 миль (4586,6 км) система обеспечивала рулевое управление автомобилем в течение 98% времени. Человек брал на себя управление лишь в течение остальных 2%, главным образом на выездных пандусах. Компьютер NavLab был оборудован видеокамерами, которые передавали изображения дороги в систему Alvinn, а затем эта система вычисляла наилучшее направление движения, основываясь на опыте, полученном в предыдущих учебных пробегах.
- **Диагностика.** Медицинские диагностические программы, основанные на вероятностном анализе, сумели достичь уровня опытного врача в нескольких

областях медицины. Хекерман [640] описал случай, когда ведущий специалист в области патологии лимфатических узлов не согласился с диагнозом программы в особо сложном случае. Создатели программы предложили, чтобы этот врач запросил у компьютера пояснения по поводу данного диагноза. Машина указала основные факторы, повлиявшие на ее решение, и объяснила нюансы взаимодействия нескольких симптомов, наблюдавшихся в данном случае. В конечном итоге эксперт согласился с решением программы.

- **Планирование снабжения.** Во время кризиса в Персидском заливе в 1991 году в армии США была развернута система DART (Dynamic Analysis and Replanning) [311] для обеспечения автоматизированного планирования поставок и составления графиков перевозок. Работа этой системы охватывала одновременно до 50 000 автомобилей, единиц груза и людей; в ней приходилось учитывать пункты отправления и назначения, маршруты, а также устранять конфликты между всеми параметрами. Методы планирования на основе искусственного интеллекта позволяли вырабатывать в течение считанных часов такие планы, для составления которых старыми методами потребовались бы недели. Представители агентства DARPA (Defense Advanced Research Project Agency — Управление перспективных исследовательских программ) заявили, что одно лишь это приложение сторицей окупило тридцатилетние инвестиции в искусственный интеллект, сделанные этим агентством.
- **Робототехника.** Многие хирурги теперь используют роботов-ассистентов в микрохирургии. Например, HipNav [398] — это система, в которой используются методы компьютерного зрения для создания трехмерной модели анатомии внутренних органов пациента, а затем применяется робототехническое управление для руководства процессом вставки протеза, заменяющего тазобедренный сустав.
- **Понимание естественного языка и решение задач.** Программа Proverb [938] — это компьютерная программа, которая решает кроссворды намного лучше, чем большинство людей; в ней используются ограничения, определяющие состав возможных заполнителей слов, большая база с данными о встречающихся ранее кроссвордах, а также множество различных источников информации, включая словари и оперативные базы данных, таких как списки кинофильмов и актеров, которые играли в этих фильмах. Например, эта программа способна определить, что одним из решений, подходящих для ключа “Nice Story”, является слово “ETAGE”, поскольку ее база данных содержит пару ключ—решение “Story in France/ETAGE”, а сама программа распознает, что шаблоны “Nice X” и “X in France” часто имеют одно и то же решение. Программа не знает, что Nice (Ницца) — город во Франции, но способна разгадать эту головоломку.

Выше приведено лишь несколько примеров систем искусственного интеллекта, которые существуют в настоящее время. Искусственный интеллект — это не магия и не научная фантастика, а сплав методов науки, техники и математики, вводное описание которых приведено в данной книге.

1.5. РЕЗЮМЕ

В настоящей главе дано определение искусственного интеллекта и описан исторический контекст, в котором развивалась эта область науки. Ниже приведены некоторые важные темы, которые рассматривались в этой главе.

- Взгляды ученых на искусственный интеллект не совпадают. Для того чтобы определить наиболее приемлемый для себя подход, необходимо ответить на два важных вопроса: “Интересует ли вас в основном мышление или поведение?” и “Стремитесь ли вы моделировать способности людей или строить свою работу исходя из идеального стандарта?”
- В данной книге принят подход, согласно которому интеллектуальность в основном связана с **рациональной деятельностью**. В идеальном случае **интеллектуальный агент** в любой ситуации предпринимает наилучшее возможное действие. В дальнейшем изложении рассматривается проблема создания агентов, которые являются интеллектуальными именно в этом смысле.
- Философы (начиная с 400 года до н.э.) заложили основы искусственного интеллекта, сформулировав идеи, что мозг в определенных отношениях напоминает машину, что он оперирует знаниями, закодированными на каком-то внутреннем языке, и что мышление может использоваться для выбора наилучших предпринимаемых действий.
- Математики предоставили инструментальные средства для манипулирования высказываниями, обладающими логической достоверностью, а также недостоверными вероятностными высказываниями. Кроме того, они заложили основу не только понимания того, что представляют собой вычисления, но и формирования рассуждений об алгоритмах.
- Экономисты формализовали проблему принятия решений, максимизирующих ожидаемый выигрыш для лица, принимающего решение.
- Психологи подтвердили идею, что люди и животные могут рассматриваться как машины обработки информации. Лингвисты показали, что процессы использования естественного языка укладываются в эту модель.
- Компьютерные инженеры предоставили артефакты, благодаря которым стало возможным создание приложений искусственного интеллекта. Обычно программы искусственного интеллекта имеют большие размеры, и не могли бы работать без тех значительных достижений в повышении быстродействия и объема памяти, которые были достигнуты в компьютерной индустрии.
- Теория управления посвящена проектированию устройств, которые действуют оптимально на основе обратной связи со средой. Первоначально математические инструментальные средства теории управления весьма отличались от применяемых в искусственном интеллекте, но эти научные области все больше сближаются.
- История искусственного интеллекта характеризуется периодами успеха и неоправданного оптимизма, за которыми следовало снижение интереса и сокращение финансирования. В ней также были периоды, когда появлялись новые творческие подходы, а затем лучшие из них систематически совершенствовались.

- Искусственный интеллект развивался быстрее, чем обычно, в прошлое десятилетие, поскольку в этой области стали шире применяться научные методы экспериментирования и сравнения подходов.
- Последние достижения на пути понимания теоретических основ интеллектуальности неразрывно связаны с расширением возможностей реальных систем. Отдельные подобласти искусственного интеллекта стали в большей степени интегрированными, а сам искусственный интеллект успешно находит общую почву с другими научными дисциплинами.

БИБЛИОГРАФИЧЕСКИЕ И ИСТОРИЧЕСКИЕ ЗАМЕТКИ

Состояние методологических основ искусственного интеллекта исследовано Гербертом Саймоном в его работе *The Sciences of the Artificial* [1417], в которой обсуждаются области научных исследований, относящиеся к сложным артефактам. В этой книге дано объяснение того, почему искусственный интеллект может рассматриваться и как прикладная, и как теоретическая наука. В [275] дан краткий обзор методологии проведения экспериментов, применяемой в искусственном интеллекте. В [480] изложено определенное мнение о полезности теста Тьюринга, отражающее собственную оценку этого теста некоторыми учеными.

В книге Джона Хоглэнда *Artificial Intelligence: The Very Idea* [631] приведено интересное описание философских и практических проблем искусственного интеллекта. Когнитология хорошо описана в нескольких недавно опубликованных книгах [740], [1465], [1502] и *Encyclopedia of the Cognitive Sciences* [1599]. В [61] рассматривается синтаксическая часть современной лингвистики, в [249] предметом изложения является семантика, а в [756] описана компьютерная лингвистика.




Ранний период развития искусственного интеллекта описан в книге Фейгенбаума и Фельдмана *Computers and Thought* [459], в книге Минского *Semantic Information Processing* [1052] и в серии книг *Machine Intelligence*, изданной под редакцией Дональда Мичи. Большое количество важных статей было выпущено в виде антологий [964] и [1562]. Ранние работы по нейронным сетям собраны в работе *Neurocomputing* [29]. В книге *Encyclopedia of AI* [1397] содержатся обзорные статьи почти по каждой теме в искусственном интеллекте. Обычно эти статьи становятся хорошим введением перед ознакомлением с научно-исследовательской литературой по каждой теме.

Результаты новейших работ публикуются в трудах основных конференций по искусственному интеллекту: проводимой один раз в два года конференции *International Joint Conference on AI* (IJCAI), ежегодной конференции *European Conference on AI* (ECAI) и конференции *National Conference on AI*, более часто упоминаемой под названием AAAI (так сокращенно называется организация American Association for AI, под эгидой которой проводится эта конференция). Главными журналами по общим направлениям искусственного интеллекта являются *Artificial Intelligence*, *Computational Intelligence*, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *IEEE Intelligent Systems* и электронный *Journal of Artificial Intelligence Research*. Имеется также много конференций и журналов, посвященных определенным областям, которые будут указаны в соответствующих главах. Основными профессиональными обществами по искусственному интеллекту являются *American Association for Artificial*

Intelligence (AAAI), *ACM Special Interest Group in Artificial Intelligence* (SIGART) и *Society for Artificial Intelligence and Simulation of Behaviour* (AISB). В журнале *AI Magazine* организации AAAI можно найти много тематических и учебных статей, а на Web-узле этого общества (aaai.org) публикуются новости и основная информация.

УПРАЖНЕНИЯ

Эти упражнения предназначены для организации на их основе творческой дискуссии, а некоторые из них могут быть определены как проекты с заданными сроками. Еще один вариант состоит в том, чтобы предпринять предварительные попытки их решения сейчас, а затем пересмотреть результаты этих попыток после завершения изучения книги.

- 1.1. Самостоятельно сформулируйте определения следующих понятий: а) интеллектуальность; б) искусственный интеллект; в) агент.
- 1.2.  Прочитайте оригинальную статью Тьюринга по искусственному интеллекту [1520]. В этой статье он обсуждает несколько потенциальных возражений против предложенного им подхода и теста интеллектуальности. Какие из этих возражений все еще остаются весомыми в определенной степени? Действительно ли приведенные им опровержения этих возражений являются правильными? Можете ли вы выдвинуть новые возражения, которые следуют из событий, происшедших с тех пор, как Тьюринг написал свою статью? В этой статье он предсказал, что к 2000 году компьютер с вероятностью 30% будет успешно проходить пятиминутный тест Тьюринга с участием слабо подготовленного экспериментатора. Какие шансы, по вашему мнению, имел бы компьютер сегодня? Еще через 50 лет?
- 1.3.  Каждый год происходит вручение приза Лебнера (Loebner) создателям программы, которая показывает наилучшие результаты при прохождении определенной версии теста Тьюринга. Проведите исследование и сообщите о последнем победителе в соревновании за приз Лебнера. Какие методы используются в этой программе? Какой вклад внесла эта программа в развитие искусственного интеллекта?
- 1.4. Существуют известные классы проблем, которые являются трудноразрешимыми для компьютеров, а в отношении других классов доказано, что они неразрешимы. Следует ли из этого вывод, что создание искусственного интеллекта невозможно?
- 1.5. Предположим, что программа Analogy Эванса будет настолько усовершенствована, что сможет получать 200 очков при стандартной проверке показателя интеллекта. Означает ли это, что при этом будет создана программа, более интеллектуальная, чем человек? Обоснуйте свой ответ.
- 1.6. Почему самоанализ (составление отчета о своих собственных сокровенных мыслях) может оказаться неточным? Как человек может оказаться неправ, обсуждая то, что он думает? Обоснуйте свой ответ.
- 1.7.  Изучите литературу по искусственному интеллекту, чтобы определить, могут ли следующие задачи в настоящее время быть решены компьютерами.

- а) Игра в настольный теннис (пинг-понг) на достаточно высоком уровне.
- б) Вождение автомобиля в центре Каира.
- в) Покупка в супермаркете недельного запаса продовольствия.
- г) Покупка недельного запаса продовольствия в Web.
- д) Участие в карточной игре бридж на конкурентоспособном уровне.
- е) Открытие и доказательство новых математических теорем.
- ж) Написание рассказа, который непременно должен быть смешным.
- а) Предоставление компетентной юридической консультации в специализированной области законодательства.
- и) Перевод в реальном времени разговорной речи с английского языка на шведский язык.
- к) Выполнение сложной хирургической операции.

В отношении задач, которые в настоящее время остаются неосуществимыми, попытайтесь узнать, в чем заключаются трудности, и предсказать, когда они будут преодолены (и произойдет ли это вообще).

- 1.8. Некоторые авторы утверждают, что самой важной частью интеллекта служат сенсорные способности и моторные навыки и что “высокоуровневые” возможности неизбежно остаются паразитическими, поскольку являются простыми дополнениями к этим основным возможностям. И действительно, не подлежит сомнению, что развитие способностей к восприятию и моторных навыков происходило на протяжении почти всей эволюции, а их поддержка осуществляется в большей части мозга, тогда как искусственный интеллект сосредоточился на таких задачах, как ведение игры и формирование логического вывода, которые во многом оказались значительно более простыми по сравнению с восприятием и осуществлением действий в реальном мире. Не кажется ли вам, что традиционная направленность искусственного интеллекта на изучение высокоуровневых познавательных способностей не совсем оправдана?
- 1.9. Почему результатом эволюции обычно становится появление систем, которые действуют рационально? Для достижения каких целей предназначены подобные системы?
- 1.10. Являются ли рефлекторные действия (такие как отдергивание руки от горячей печи) рациональными? Являются ли они интеллектуальными?
- 1.11. “Безусловно, компьютеры не могут быть интеллектуальными, ведь они способны выполнять только то, что диктуют им программисты”. Является ли последнее утверждение истинным и следует ли из него первое?
- 1.12. “Безусловно, животные не могут быть интеллектуальными, ведь они способны выполнять только то, что диктуют им гены”. Является ли последнее утверждение истинным и следует ли из него первое?
- 1.13. “Безусловно, животные, люди и компьютеры не могут быть интеллектуальными, ведь они способны выполнять только то, что диктуют атомам, из которых они состоят, законы физики”. Является ли последнее утверждение истинным и следует ли из него первое?

2 ИНТЕЛЛЕКТУАЛЬНЫЕ АГЕНТЫ

В этой главе рассматриваются характеристики агентов, идеальных или неидеальных, разнообразие вариантов среды и вытекающая из этого классификация типов агентов.

Как было указано в главе 1, понятие **рационального агента** является центральным в применяемом авторами данной книги подходе к искусственному интеллекту. В этой главе указанное понятие раскрывается более подробно. В ней показано, что концепция рациональности может применяться к самым различным агентам, действующим в любой среде, которую только можно себе представить. План авторов состоит в том, чтобы использовать эту концепцию в данной книге для разработки небольшого набора принципов проектирования для создания успешно действующих агентов — систем, которые вполне можно было бы назвать **интеллектуальными**.

Начнем с изучения агентов, вариантов среды и связей между ними. Наблюдая за тем, что некоторые агенты действуют лучше, чем другие, можно вполне обоснованно выдвинуть идею рационального агента; таковым является агент, который действует настолько успешно, насколько это возможно. Успехи, которых может добиться агент, зависят от характера среды; некоторые варианты среды являются более сложными, чем другие. В этой главе дана грубая классификация вариантов среды и показано, как свойства среды влияют на проектирование агентов, наиболее подходящих для данной среды. Здесь описан ряд основных, “скелетных” проектов агентов, которые будут облечены в плоть в остальной части книги.

2.1. АГЕНТЫ И ВАРИАНТЫ СРЕДЫ

Агентом является все, что может рассматриваться как воспринимающее свою **среду** с помощью **датчиков** и воздействующее на эту среду с помощью **исполнительных механизмов**. Эта простая идея иллюстрируется на рис. 2.1. Человек, рассматриваемый в роли агента, имеет глаза, уши и другие органы чувств, а исполнительными механизмами для него служат руки, ноги, рот и другие части тела. Робот, выполняющий функции агента, в качестве датчиков может иметь видеокамеры и инфракрасные дальномеры, а его исполнительными механизмами могут являться различные двигатели. Программное обеспечение, выступающее в роли аген-

та, в качестве входных сенсорных данных получает коды нажатия клавиш, содержимое файлов и сетевые пакеты, а его воздействие на среду выражается в том, что программное обеспечение выводит данные на экран, записывает файлы и передает сетевые пакеты. Мы принимаем общее допущение, что каждый агент может воспринимать свои собственные действия (но не всегда их результаты).

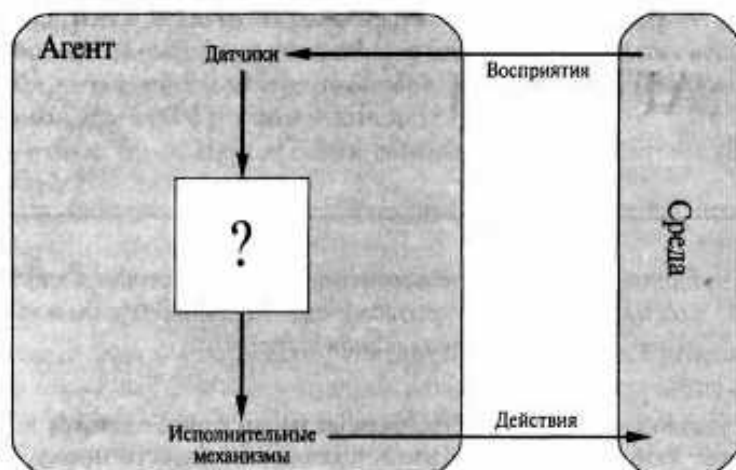


Рис. 2.1. Агент взаимодействует со средой с помощью датчиков и исполнительных механизмов

Мы используем термин Σ **восприятие** для обозначения полученных агентом сенсорных данных в любой конкретный момент времени. Σ **Последовательностью актов восприятия** агента называется полная история всего, что было когда-либо воспринято агентом. Вообще говоря, Φ **выбор агентом действия в любой конкретный момент времени может зависеть от всей последовательности актов восприятия, наблюдавшихся до этого момента времени**. Если существует возможность определить, какое действие будет выбрано агентом в ответ на любую возможную последовательность актов восприятия, то может быть дано более или менее точное определение агента. С точки зрения математики это равносильно утверждению, что поведение некоторого агента может быть описано с помощью Σ **функции агента**, которая отображает любую конкретную последовательность актов восприятия на некоторое действие.

Может рассматриваться задача табуляции функции агента, которая описывает любого конкретного агента; для большинства агентов это была бы очень большая таблица (фактически бесконечная), если не устанавливается предел длины последовательностей актов восприятия, которые должны учитываться в таблице. Проводя эксперименты с некоторым агентом, такую таблицу в принципе можно сконструировать, проверяя все возможные последовательности актов восприятия и регистрируя, какие действия в ответ выполняет агент¹. Такая таблица, безусловно, является внешним описанием агента. Внутреннее описание состоит в определении того, ка-

¹ Если агент для выбора своих действий использует определенную рандомизацию, то может потребоваться проверить каждую последовательность многократно, чтобы определить вероятность каждого действия. На первый взгляд кажется, что выбор действий случайным образом является довольно неразумным, но ниже в этой главе будет показано, что такая организация функционирования может оказаться весьма интеллектуальной.

кая функция агента для данного искусственного агента реализуется с помощью **программы агента**. Важно различать два последних понятия. *Функция агента* представляет собой абстрактное математическое описание, а *программа агента* — это конкретная реализация, действующая в рамках архитектуры агента.

Для иллюстрации изложенных идей воспользуемся очень простым примером: рассмотрим показанный на рис. 2.2 мир, в котором работает пылесос. Этот мир настолько прост, что существует возможность описать все, что в нем происходит; кроме того, это — мир, созданный человеком, поэтому можно изобрести множество вариантов его организации. В данном конкретном мире имеются только два местонахождения: квадраты *A* и *B*. Пылесос, выполняющий роль агента, воспринимает, в каком квадрате он находится и есть ли мусор в этом квадрате. Агент может выбрать такие действия, как переход влево, вправо, всасывание мусора или бездействие. Одна из очень простых функций агента состоит в следующем: если в текущем квадрате имеется мусор, то всосать его, иначе перейти в другой квадрат. Частичная табуляция данной функции агента показана в табл. 2.1. Простая программа агента для этой функции агента приведена ниже в этой главе, в листинге 2.2.

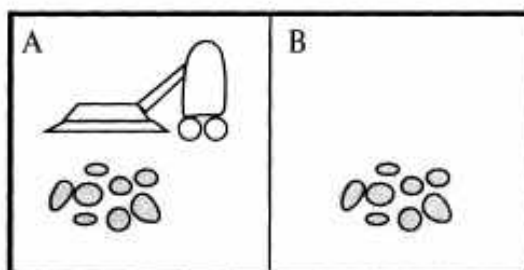


Рис. 2.2. Мир пылесоса, в котором имеются только два местонахождения

Таблица 2.1. Частичная табуляция функции простого агента для мира пылесоса, показанного на рис. 2.2

Последовательность актов восприятия	Действие
[A, Clean]	Right
[A, Dirty]	Suck
[B, Clean]	Left
[B, Dirty]	Suck
[A, Clean], [A, Clean]	Right
[A, Clean], [A, Dirty]	Suck
...	...
[A, Clean], [A, Clean], [A, Clean]	Right
[A, Clean], [A, Clean], [A, Dirty]	Suck
...	...

На основании табл. 2.1 можно сделать вывод, что для мира пылесоса можно определять различных агентов, заполняя разными способами правый столбец этой таблицы. Поэтому очевидный вопрос состоит в следующем: *«Какой способ заполнения этой таблицы является правильным?»* Иными словами, благодаря чему агент

становится хорошим или плохим, интеллектуальным или не соответствующим критериям интеллектуальности? Ответ на этот вопрос приведен в следующем разделе.

Прежде чем завершить этот раздел, необходимо отметить, что понятие агента рассматривается как инструмент для анализа систем, а не как абсолютная классификация, согласно которой мир делится на агентов и неагентов. Например, в качестве агента можно было бы рассматривать карманный калькулятор, который выбирает действие по отображению “4” после получения последовательности актов восприятия “ $2+2=$ ”, но подобный анализ вряд ли поможет понять работу калькулятора.

2.2. КАЧЕСТВЕННОЕ ПОВЕДЕНИЕ: КОНЦЕПЦИЯ РАЦИОНАЛЬНОСТИ

✎ **Рациональным агентом** является такой агент, который выполняет правильные действия; выражаясь более формально, таковым является агент, в котором каждая запись в таблице для функции агента заполнена правильно. Очевидно, что выполнение правильных действий лучше, чем осуществление неправильных действий, но что подразумевается под выражением “выполнение правильных действий”? В первом приближении можно сказать, что правильным действием является такое действие, которое обеспечивает наиболее успешное функционирование агента. Поэтому требуется определенный способ измерения успеха. Критерии успеха, наряду с описанием среды, а также датчиков и исполнительных механизмов агента, предоставляют полную спецификацию задачи, с которой сталкивается агент. Имея эти компоненты, мы можем определить более точно, что подразумевается под словом “рациональный”.

Показатели производительности

✎ **Показатели производительности** воплощают в себе критерии оценки успешного поведения агента. После погружения в среду агент вырабатывает последовательность действий, соответствующих полученным им восприятиям. Эта последовательность действий вынуждает среду пройти через последовательность состояний. Если такая последовательность соответствует желаемому, то агент функционирует хорошо. Безусловно, что не может быть одного постоянного показателя, подходящего для всех агентов. Можно было бы узнать у агента его субъективное мнение о том, насколько он удовлетворен своей собственной производительностью, но некоторые агенты не будут способны ответить, а другие склонны заниматься самообманом². Поэтому необходимо упорно добиваться применения объективных показателей производительности, и, как правило, проектировщик, конструирующий агента, предусматривает такие показатели.

Рассмотрим агент-пылесос, описанный в предыдущем разделе. Можно было бы предложить измерять показатели производительности по объему мусора, убранного за одну восьмичасовую смену. Но, безусловно, имея дело с рациональным агентом,

² Особенно известны тем, что недостигнутый успех для них — “зелен виноград”, такие агенты, как люди. Не получив кое-что для себя весьма ценное, они искренне считают, что и не стремились к этому: “Подумаешь! Мне и даром не нужна эта дурацкая Нобелевская премия!”


вы получаете то, что просите. Рациональный агент может максимизировать такой показатель производительности, убирая мусор, затем вываливая весь его на пол, затем снова убирая, и т.д. Поэтому более приемлемые критерии производительности должны вознаграждать агента за то, что пол остается чистым. Например, одно очко могло бы присуждаться за каждый чистый квадрат в каждом интервале времени (возможно, в сочетании со штрафом за потребляемую электроэнергию и создаваемый шум). ☞ *В качестве общего правила следует указать, что лучше всего разрабатывать показатели производительности в соответствии с тем, чего действительно необходимо добиться в данной среде, а не в соответствии с тем, как, по мнению проектировщика, должен вести себя агент.*

Задача выбора показателей производительности не всегда является простой. Например, понятие “чистого пола”, которое рассматривалось выше, основано на определении усредненной чистоты пола во времени. Но необходимо также учитывать, что одна и та же усредненная чистота может быть достигнута двумя различными агентами, один из которых постоянно, но неторопливо выполняет свою работу, а другой время от времени энергично занимается очисткой, но делает длинные перерывы. Может показаться, что определение того способа действий, который является в данном случае наиболее предпочтительным, относится к тонкостям домоводства, но фактически это — глубокий философский вопрос с далеко идущими последствиями. Что лучше — бесшабашная жизнь со взлетами и падениями или безопасное, но однообразное существование? Что лучше — экономика, в которой каждый живет в умеренной бедности, или такая экономика, в которой одни ни в чем не нуждаются, а другие еле сводят концы с концами? Оставляем задачу поиска ответов на эти вопросы в качестве упражнения для любознательного читателя.

Рациональность

В любой конкретный момент времени оценка рациональности действий агента зависит от четырех перечисленных ниже факторов.

- Показатели производительности, которые определяют критерии успеха.
- Знания агента о среде, приобретенные ранее.
- Действия, которые могут быть выполнены агентом.
- Последовательность актов восприятия агента, которые произошли до настоящего времени.

С учетом этих факторов можно сформулировать следующее  **определение рационального агента.**

☞ *Для каждой возможной последовательности актов восприятия рациональный агент должен выбрать действие, которое, как ожидается, максимизирует его показатели производительности, с учетом фактов, предоставленных данной последовательностью актов восприятия и всех встроенных знаний, которыми обладает агент.*

Рассмотрим пример простого агента-пылесоса, который очищает квадрат, если в нем имеется мусор, и переходит в другой квадрат, если мусора в нем нет; результаты частичной табуляции такой функции агента приведены в табл. 2.1. Является ли этот агент рациональным? Ответ на этот вопрос не так уж прост! Вначале необходимо определить, в чем состоят показатели производительности, что известно о среде и ка-

кие датчики и исполнительные механизмы имеет агент. Примем перечисленные ниже предположения.

- Применяемые показатели производительности предусматривают вознаграждение в одно очко за каждый чистый квадрат в каждом интервале времени в течение “срока существования” агента, состоящего из 1000 интервалов времени.
- “География” среды известна заранее (рис. 2.2), но распределение мусора и первоначальное местонахождение агента не определены. Чистые квадраты остаются чистыми, а всасывание мусора приводит к очистке текущего квадрата. Действия *Left* и *Right* приводят к перемещению агента соответственно влево и вправо, за исключением тех случаев, когда они могли бы вывести агента за пределы среды, и в этих случаях агент остается там, где он находится.
- Единственными доступными действиями являются *Left*, *Right*, *Suck* (всосать мусор) и *NoOp* (ничего не делать).
- Агент правильно определяет свое местонахождение и воспринимает показания датчика, позволяющие узнать, имеется ли мусор в этом местонахождении.

Авторы утверждают, что в этих обстоятельствах агент действительно является рациональным; его ожидаемая производительность, по меньшей мере, не ниже, чем у любых других агентов. В упр. 2.4 предложено доказать это утверждение.

Можно легко обнаружить, что в других обстоятельствах тот же самый агент может стать нерациональным. Например, после того как весь мусор будет очищен, агент станет совершать ненужные периодические перемещения вперед и назад; если показатели производительности предусматривают штраф в одно очко за каждое передвижение в том или ином направлении, то агент не сможет хорошо зарабатывать. В таком случае лучший агент должен был бы ничего не делать до тех пор, пока он уверен в том, что все квадраты остаются чистыми. Если же чистые квадраты могут снова стать грязными, то агент обязан время от времени проводить проверку и снова очищать их по мере необходимости. А если география среды неизвестна, то агенту может потребоваться исследовать ее, а не ограничиваться квадратами *A* и *B*. В упр. 2.4 предложено спроектировать агентов для подобных случаев.

Всезнание, обучение и автономность

Необходимо тщательно проводить различие между рациональностью и ~~всезнанием~~. Всезнающий агент знает фактический результат своих действий и может действовать соответствующим образом; но всезнание в действительности невозможно. Рассмотрим следующий пример: некий господин однажды гуляет в Париже по Елисейским Полям и видит на другой стороне улицы старого приятеля. Вблизи нет никаких машин, а наш господин никуда не спешит, поэтому, будучи рациональным агентом, он начинает переходить через дорогу. Между тем на высоте 10 000 метров у пролетающего самолета отваливается дверь грузового отсека³, и прежде чем несчастный успевает достичь другой стороны улицы, расплющивает его в лепешку. Было ли нерациональным именно то, что этот господин решил перейти на другую сторону улицы? Весьма маловероятно, что в его некрологе написали бы: “Жертва идиотской попытки перейти улицу”.

³ См. заметку Н. Гендерсона “New door latches urged for Boeing 747 jumbo jets” (На дверях аэробусов Boeing 747 необходимо срочно установить новые замки). — Washington Post, 24 августа 1989 года.

Этот пример показывает, что рациональность нельзя рассматривать как равнозначную совершенству. Рациональность — это максимизация ожидаемой производительности, а совершенство — максимизация фактической производительности. Отказываясь от стремления к совершенству, мы не только применяем к агентам справедливые критерии, но и учитываем реальность. Дело в том, что если от агента требуют, чтобы он выполнял действия, которые оказываются наилучшими после их совершения, то задача проектирования агента, отвечающего этой спецификации, становится невыполнимой (по крайней мере, до тех пор, пока мы не сможем повысить эффективность машин времени или хрустальных шаров, применяемых гадалками).

Поэтому наше определение рациональности не требует всезнания, ведь рациональный выбор зависит только от последовательности актов восприятия, сформированной к данному моменту. Необходимо также следить за тем, чтобы мы непреднамеренно не позволили бы агенту участвовать в действиях, которые, безусловно, не являются интеллектуальными. Например, если агент не оглядывается влево и вправо, прежде чем пересечь дорогу с интенсивным движением, то полученная им до сих пор последовательность актов восприятия не сможет подсказать, что к нему на большой скорости приближается огромный грузовик. Указывает ли наше определение рациональности, что теперь агент может перейти через дорогу? Отнюдь нет! Во-первых, агент не был бы рациональным, если бы попытался перейти на другую сторону, получив такую неинформативную последовательность актов восприятия: риск несчастного случая при подобной попытке перейти автомагистраль, не оглянувшись, слишком велик. Во-вторых, рациональный агент должен выбрать действие “оглянуться”, прежде чем ступить на дорогу, поскольку такой осмотр позволяет максимизировать ожидаемую производительность. Выполнение в целях модификации будущих восприятий определенных действий (иногда называемых **сбором информации**) составляет важную часть рациональности и подробно рассматривается в главе 16. Второй пример сбора информации выражается в том **исследовании ситуации**, которое должно быть предпринято агентом-пылесосом в среде, которая первоначально была для него неизвестной.

Наше определение требует, чтобы рациональный агент не только собирал информацию, но также **обучался** в максимально возможной степени на тех данных, которые он воспринимает. Начальная конфигурация агента может отражать некоторые предварительные знания о среде, но по мере приобретения агентом опыта эти знания могут модифицироваться и пополняться. Существуют крайние случаи, в которых среда полностью известна заранее. В подобных случаях агенту не требуется воспринимать информацию или обучаться; он просто сразу действует правильно. Безусловно, такие агенты являются весьма уязвимыми. Рассмотрим скромного навозного жука. Выкопав гнездо и отложив яйца, он скатывает шарик навоза, набрав его из ближайшей навозной кучи, чтобы заткнуть вход в гнездо. Если шарик навоза будет удален непосредственно перед тем, как жук его схватит, жук продолжает манипулировать им и изображает такую пантомиму, как будто он затыкает гнездо несуществующим шариком навоза, даже не замечая, что этот шарик отсутствует. В результате эволюции поведение этого жука было сформировано на основании определенного предположения, а если это предположение нарушается, то за этим следует безуспешное поведение. Немного более интеллектуальными являются осы-сфексы. Самка сфекса выкапывает норку, выходит из нее, жалит гусеницу и затаскивает ее в норку, затем снова выходит из норки, чтобы проверить, все ли в порядке, вытаски-

вает гусеницу наружу и откладывает в нее яйца. Гусеница служит в качестве источника питания во время развития яиц. До сих пор все идет хорошо, но если энтомолог переместит гусеницу на несколько дюймов в сторону, пока сфекс выполняет свою проверку, это насекомое снова возвращается к этапу “перетаскивания” своего плана и продолжает выполнять план без изменений, даже после десятков вмешательств в процедуру перемещения гусеницы. Оса-сфекс не способна обучиться действовать в такой ситуации, когда ее врожденный план нарушается, и поэтому не может его изменить.

В успешно действующих агентах задача вычисления функции агента разбивается на три отдельных периода: при проектировании агента некоторые вычисления осуществляются его проектировщиками; дополнительные вычисления агент производит, выбирая одно из своих очередных действий; а по мере того как агент учится на основании опыта, он осуществляет другие вспомогательные вычисления для принятия решения о том, как модифицировать свое поведение.

Если степень, в которой агент полагается на априорные знания своего проектировщика, а не на свои восприятия, слишком высока, то такой агент рассматривается как обладающий недостаточной **автономностью**. Рациональный агент должен быть автономным — он должен обучаться всему, что может освоить, для компенсации неполных или неправильных априорных знаний. Например, агент-пылесос, который обучается прогнозированию того, где и когда появится дополнительный мусор, безусловно, будет работать лучше, чем тот агент, который на это не способен. С точки зрения практики агенту редко предъявляется требование, чтобы он был полностью автономным с самого начала: если агент имеет мало опыта или вообще не имеет опыта, то вынужден действовать случайным образом, если проектировщик не оказал ему определенную помощь. Поэтому, как и эволюция предоставила животным достаточное количество врожденных рефлексов, позволяющих им прожить после рождения настолько долго, чтобы успеть обучиться самостоятельно, так и искусственному интеллектуальному агенту было бы разумно предоставить некоторые начальные знания, а не только наделить его способностью обучаться. После достаточного опыта существования в своей среде поведение рационального агента может по сути стать независимым от его априорных знаний. Поэтому включение в проект способности к обучению позволяет проектировать простых рациональных агентов, которые могут действовать успешно в исключительно разнообразных вариантах среды.

2.3. ОПРЕДЕЛЕНИЕ ХАРАКТЕРА СРЕДЫ

Теперь, после разработки определения рациональности, мы почти готовы приступить к созданию рациональных агентов. Но вначале необходимо определить, чем является **проблемная среда**, по сути представляющая собой “проблему”, для которой рациональный агент служит “решением”. Начнем с демонстрации того, как определить проблемную среду, и проиллюстрируем этот процесс на ряде примеров. Затем в этом разделе будет показано, что проблемная среда может иметь целый ряд разновидностей. Выбор проекта, наиболее подходящего для программы конкретного агента, непосредственно зависит от рассматриваемой разновидности проблемной среды.

Определение проблемной среды

В приведенном выше исследовании рациональности простого агента-пылесоса нам пришлось определить показатели производительности, среду, а также исполнительные механизмы и датчики агента. Сгруппируем описание всех этих факторов под заголовком **проблемная среда**. Для тех, кто любит аббревиатуры, авторы сокращенно обозначили соответствующее описание как **PEAS** (Performance, Environment, Actuators, Sensors — производительность, среда, исполнительные механизмы, датчики). Первый этап проектирования любого агента всегда должен состоять в определении проблемной среды с наибольшей возможной полнотой.

Пример, в котором рассматривался мир пылесоса, был несложным; теперь рассмотрим более сложную проблему — создание автоматизированного водителя такси. Этот пример будет использоваться во всей оставшейся части данной главы. Прежде чем читатель почувствует тревогу за безопасность будущих пассажиров, хотим сразу же отметить, что задача создания полностью автоматизированного водителя такси в настоящее время все еще выходит за пределы возможностей существующей технологии. (См. с. 69, где приведено описание существующего робота-водителя; с состоянием дел в этой области можно также ознакомиться по трудам конференций, посвященных интеллектуальным транспортным системам, в названиях которых есть слова *Intelligent Transportation Systems*.) Полное решение проблемы вождения автомобиля является чрезвычайно трудоемким, поскольку нет предела появлению все новых и новых комбинаций обстоятельств, которые могут возникать в процессе вождения; это еще одна из причин, по которой мы выбрали данную проблему для обсуждения. В табл. 2.2 приведено итоговое описание PEAS для проблемной среды вождения такси. Каждый из элементов этого описания рассматривается более подробно в настоящей главе.

Таблица 2.2. Описание PEAS проблемной среды для автоматизированного водителя такси

Тип агента	Показатели производительности	Среда	Исполнительные механизмы	Датчики
Водитель такси	Безопасная, быстрая, комфортная езда в рамках правил дорожного движения, максимизация прибыли	Дороги, другие транспортные средства, пешеходы, клиенты	Рулевое управление, акселератор, тормоз, световые сигналы, клаксон, дисплей	Видеокамеры, ультразвуковой дальномер, спидометр, глобальная система навигации и определения положения, одометр, акселерометр, датчики двигателя, клавиатура

Прежде всего необходимо определить **показатели производительности**, которыми мы могли бы стимулировать деятельность нашего автоматизированного водителя. К желаемым качествам относится успешное достижение нужного места назначения; минимизация потребления топлива, износа и старения; минимизация продолжительности и/или стоимости поездки; минимизация количества нарушений правил дорожного движения и помех другим водителям; максимизация безопасности и комфорта пассажиров; максимизация прибыли. Безусловно, что некоторые из этих целей конфликтуют, поэтому должны рассматриваться возможные компромиссы.

Затем рассмотрим, в чем состоит **среда вождения**, в которой действует такси. Любому водителю такси приходится иметь дело с самыми различными дорогами, начи-

ная с проселков и узких городских переулков и заканчивая автострадами с двенадцатью полосами движения. На дороге встречаются другие транспортные средства, беспризорные животные, пешеходы, рабочие, производящие дорожные работы, полицейские автомобили, лужи и выбоины. Водителю такси приходится также иметь дело с потенциальными и действительными пассажирами. Кроме того, имеется еще несколько важных дополнительных факторов. Таксисту может выпасть участь работать в Южной Калифорнии, где редко возникает такая проблема, как снег, или на Аляске, где снега на дорогах не бывает очень редко. Может оказаться, что водителю всю жизнь придется ездить по правой стороне или от него может потребоваться, чтобы он сумел достаточно успешно приспособиться к езде по левой стороне во время пребывания в Британии или в Японии. Безусловно, чем более ограниченной является среда, тем проще задача проектирования.

Исполнительные механизмы, имеющиеся в автоматизированном такси, должны быть в большей или меньшей степени такими же, как и те, которые находятся в распоряжении водителя-человека: средства управления двигателем с помощью акселератора и средства управления вождением с помощью руля и тормозов. Кроме того, для него могут потребоваться средства вывода на экран дисплея или синтеза речи для передачи ответных сообщений пассажирам и, возможно, определенные способы общения с водителями других транспортных средств, иногда вежливого, а иногда и не совсем.

Для достижения своих целей в данной среде вождения таксисту необходимо будет знать, где он находится, кто еще едет по этой дороге и с какой скоростью движется он сам. Поэтому в число его основных **датчиков** должны входить одна или несколько управляемых телевизионных камер, спидометр и одометр. Для правильного управления автомобилем, особенно на поворотах, в нем должен быть предусмотрен акселерометр; водителю потребуется также знать механическое состояние автомобиля, поэтому для него будет нужен обычный набор датчиков для двигателя и электрической системы. Автоматизированный водитель может также иметь приборы, недоступные для среднего водителя-человека: спутниковую глобальную систему навигации и определения положения (Global Positioning System — GPS) для получения точной информации о местонахождении по отношению к электронной карте, а также инфракрасные или ультразвуковые датчики для измерения расстояний до других автомобилей и препятствий. Наконец, ему потребуется клавиатура или микрофон для пассажиров, чтобы они могли указать место своего назначения.

В табл. 2.3 кратко перечислены основные элементы PEAS для целого ряда других типов агентов. Дополнительные примеры приведены в упр. 2.5. Некоторым читателям может показаться удивительным то, что авторы включили в этот список типов агентов некоторые программы, которые функционируют в полностью искусственной среде, ограничиваемой вводом с клавиатуры и выводом символов на экран. Кое-кто мог бы сказать: “Разумеется, это же не реальная среда, не правда ли?” В действительности суть состоит не в различиях между “реальными” и “искусственными” вариантами среды, а в том, какова сложность связей между поведением агента, последовательностью актов восприятия, вырабатываемой этой средой, и показателями производительности. Некоторые “реальные” варианты среды фактически являются чрезвычайно простыми. Например, для робота, предназначенного для контроля деталей, проходящих мимо него на ленточном конвейере, может использоваться целый ряд упрощающих допущений, например, что освещение всегда включено, что единственными предме-

тами на ленте конвейера являются детали того типа, который ему известен, и что существуют только два действия (принять изделие или забраковать его).

Таблица 2.3. Примеры типов агентов и их описаний PEAS

Тип агента	Показатели произ-водительности	Среда	Исполнительные механизмы	Датчики
Медицинская диагностическая система	Успешное излечение пациента, минимизация затрат, отсутствие поводов для судебных тяжб	Пациент, больница, персонал	Вывод вопросов, тестов, диагнозов, рекомендаций, направлений	Ввод с клавиатуры симптомов, результатов лабораторных исследований, ответов пациента
Система анализа изображений, полученных со спутника	Правильная классификация изображения	Канал передачи данных от орбитального спутника	Вывод на дисплей результатов классификации определенного фрагмента изображения	Массивы пикселей с данными о цвете
Робот-сортировщик деталей	Процентные показатели безошибочной сортировки по лоткам	Ленточный конвейер с движущимися на нем деталями; лотки	Шарнирный манипулятор и захват	Видеокамера, датчики углов поворота шарниров
Контроллер очистительной установки	Максимизация степени очистки, продуктивности, безопасности	Очистительная установка, операторы	Клапаны, насосы, нагреватели, дисплеи	Температура, давление, датчики химического состава
Интерактивная программа обучения английскому языку	Максимизация оценок студентов на экзаменах	Множество студентов, экзаменационное агентство	Вывод на дисплей упражнений, рекомендаций, исправлений	Ввод с клавиатуры

В отличие от этого некоторые **программные агенты** (называемые также **программными роботами** или **софтботами**) существуют в сложных, неограниченных проблемных областях. Представьте себе программный робот, предназначенный для управления тренажером, имитирующим крупный пассажирский самолет. Этот тренажер представляет собой очень детально промоделированную, сложную среду, в которой имитируются движения других самолетов и работа наземных служб, а программный агент должен выбирать в реальном времени наиболее целесообразные действия из широкого диапазона действий. Еще одним примером может служить программный робот, предназначенный для просмотра источников новостей в Internet и показа клиентам интересующих их сообщений. Для успешной работы ему требуются определенные способности к обработке текста на естественном языке, он должен в процессе обучения определять, что интересует каждого заказчика, а также должен уметь изменять свои планы динамически, допустим, когда соединение с каким-либо из источников новостей закрывается или в оперативный режим переходит новый источник новостей. Internet представляет собой среду, которая по своей сложности соперничает с физическим миром, а в число обитателей этой сети входит много искусственных агентов.

Свойства проблемной среды

Несомненно, что разнообразие вариантов проблемной среды, которые могут возникать в искусственном интеллекте, весьма велико. Тем не менее существует возможность определить относительно небольшое количество измерений, по которым могут быть классифицированы варианты проблемной среды. Эти измерения в значительной степени определяют наиболее приемлемый проект агента и применимость каждого из основных семейств методов для реализации агента. Вначале в этом разделе будет приведен список измерений, а затем проанализировано несколько вариантов проблемной среды для иллюстрации этих идей. Приведенные здесь определения являются неформальными; более точные утверждения и примеры вариантов среды каждого типа описаны в следующих главах.

- **Полностью наблюдаемая или частично наблюдаемая**

Если датчики агента предоставляют ему доступ к полной информации о состоянии среды в каждый момент времени, то такая проблемная среда называется полностью наблюдаемой⁴. По сути, проблемная среда является полностью наблюдаемой, если датчики выявляют все данные, которые являются релевантными для выбора агентом действия; релевантность, в свою очередь, зависит от показателей производительности. Полностью наблюдаемые варианты среды являются удобными, поскольку агенту не требуется поддерживать какое-либо внутреннее состояние для того, чтобы быть в курсе всего происходящего в этом мире. Среда может оказаться частично наблюдаемой из-за создающих шум и неточных датчиков или из-за того, что отдельные характеристики ее состояния просто отсутствуют в информации, полученной от датчиков; например, агент-пылесос, в котором имеется только локальный датчик мусора, не может определить, имеется ли мусор в других квадратах, а автоматизированный водитель такси не имеет сведений о том, какие маневры намереваются выполнить другие водители.

- **Детерминированная или стохастическая**

Если следующее состояние среды полностью определяется текущим состоянием и действием, выполненным агентом, то такая среда называется детерминированной; в противном случае она является стохастической. В принципе в полностью наблюдаемой детерминированной среде агенту не приходится действовать в условиях неопределенности. Но если среда — частично наблюдаемая, то может создаться впечатление, что она является стохастической. Это отчасти справедливо, если среда — сложная и агенту нелегко следить за всеми ее ненаблюдаемыми аспектами. В связи с этим часто бывает более удобно классифицировать среду как детерминированную или стохастическую с точки зрения агента. Очевидно, что при такой трактовке среда вождения такси является стохастической, поскольку никто не может точно предсказать поведение всех других транспортных средств; более того, в любом автомобиле совершенно неожиданно может произойти прокол шины или остановка двигателя.

⁴ В первом издании этой книги использовались термины *доступная среда* и *недоступная среда* вместо терминов *полностью наблюдаемая среда* и *частично наблюдаемая среда*; *недетерминированная* вместо *стохастической* и *неэпизодическая* вместо *последовательной*. Применяемая в этом издании новая терминология более совместима со сложившейся в этой области терминологией.

Описанный здесь мир пылесоса является детерминированным, но другие варианты этой среды могут включать стохастические элементы, такие как случайное появление мусора и ненадежная работа механизма всасывания (см. упр. 2.12). Если среда является детерминированной во всех отношениях, кроме действий других агентов, то авторы данной книги называют эту среду **стратегической**.

- **Эпизодическая или последовательная⁵**

В эпизодической проблемной среде опыт агента состоит из неразрывных эпизодов. Каждый эпизод включает в себя восприятие среды агентом, а затем выполнение одного действия. При этом крайне важно то, что следующий эпизод не зависит от действий, предпринятых в предыдущих эпизодах. В эпизодических вариантах среды выбор действия в каждом эпизоде зависит только от самого эпизода. Эпизодическими являются многие задачи классификации. Например, агент, который должен распознавать дефектные детали на сборочной линии, формирует каждое решение применительно к текущей детали, независимо от предыдущих решений; более того, от текущего решения не зависит то, будет ли определена как дефектная следующая деталь. С другой стороны, в последовательных вариантах среды текущее решение может повлиять на все будущие решения. Последовательными являются такие задачи, как игра в шахматы и вождение такси: в обоих случаях кратковременные действия могут иметь долговременные последствия. Эпизодические варианты среды гораздо проще по сравнению с последовательными, поскольку в них агенту не нужно думать наперед.

- **Статическая или динамическая**

Если среда может измениться в ходе того, как агент выбирает очередное действие, то такая среда называется динамической для данного агента; в противном случае она является статической. Действовать в условиях статической среды проще, поскольку агенту не требуется наблюдать за миром в процессе выработки решения о выполнении очередного действия, к тому же агенту не приходится беспокоиться о том, что он затрачивает на размышления слишком много времени. Динамические варианты среды, с другой стороны, как бы непрерывно спрашивают агента, что он собирается делать, а если он еще ничего не решил, то это рассматривается как решение ничего не делать. Если с течением времени сама среда не изменяется, а изменяются показатели производительности агента, то такая среда называется **полудинамической**. Очевидно, что среда вождения такси является динамической, поскольку другие автомобили и само такси продолжают движение и в ходе того, как алгоритм вождения определяет, что делать дальше. Игра в шахматы с контролем времени является полудинамической, а задача решения кроссворда — статической.

- **Дискретная или непрерывная**

Различие между дискретными и непрерывными вариантами среды может относиться к состоянию среды, способу учета времени, а также восприятиям и действиям агента. Например, такая среда с дискретными состояниями, как

⁵ Термин “последовательный” используется также в компьютерных науках как антоним термина “параллельный”. Соответствующие два толкования фактически не связаны друг с другом.

игра в шахматы, имеет конечное количество различных состояний. Кроме того, игра в шахматы связана с дискретным множеством восприятий и действий. Вождение такси — это проблема с непрерывно меняющимся состоянием и непрерывно текущим временем, поскольку скорость и местонахождение самого такси и других транспортных средств изменяются в определенном диапазоне непрерывных значений, причем эти изменения происходят во времени плавно. Действия по вождению такси также являются непрерывными (непрерывная регулировка угла поворота руля и т.д.). Строго говоря, входные данные от цифровых камер поступают дискретно, но обычно рассматриваются как представляющие непрерывно изменяющиеся скорости и местонахождения.

- ☒ **Одноагентная или ☒ мультиагентная**

Различие между одноагентными и мультиагентными вариантами среды на первый взгляд может показаться достаточно простым. Например, очевидно, что агент, самостоятельно решающий кроссворд, находится в одноагентной среде, а агент, играющий в шахматы, действует в двухагентной среде. Тем не менее при анализе этого классификационного признака возникают некоторые нюансы. Прежде всего, выше было описано, на каком основании некоторая сущность *может* рассматриваться как агент, но не было указано, какие сущности *должны* рассматриваться как агенты. Должен ли агент А (например, водитель такси) считать агентом объект В (другой автомобиль), или может относиться к нему просто как к стохастически действующему объекту, который можно сравнить с волнами, набегающими на берег, или с листьями, трепещущими на ветру? Ключевое различие состоит в том, следует ли или не следует описывать поведение объекта В как максимизирующее личные показатели производительности, значения которых зависят от поведения агента А. Например, в шахматах соперничающая сущность В пытается максимизировать свои показатели производительности, а это по правилам шахмат приводит к минимизации показателей производительности агента А. Таким образом, шахматы — это ☒ **конкурентная** мультиагентная среда. А в среде вождения такси, с другой стороны, предотвращение столкновений максимизирует показатели производительности всех агентов, поэтому она может служить примером частично ☒ **кооперативной** мультиагентной среды. Она является также частично конкурентной, поскольку, например, парковочную площадку может занять только один автомобиль. Проблемы проектирования агентов, возникающие в мультиагентной среде, часто полностью отличаются от тех, с которыми приходится сталкиваться в одноагентных вариантах среды; например, одним из признаков рационального поведения в мультиагентной среде часто бывает **поддержка связи**, а в некоторых вариантах частично наблюдаемой конкурентной среды рациональным становится **стохастическое поведение**, поскольку оно позволяет избежать *ловушек предсказуемости*.

Как и следует ожидать, наиболее сложными вариантами среды являются частично наблюдаемые, стохастические, последовательные, динамические, непрерывные и мультиагентные. Кроме того, часто обнаруживается, что многие реальные ситуации являются настолько сложными, что неясно даже, действительно ли их можно считать детерминированными. С точки зрения практики их следует рассматривать как стохастические. Проблема вождения такси является сложной во всех указанных отношениях.

В табл. 2.4 перечислены свойства многих известных вариантов среды. Следует отметить, что в отдельных случаях приведенные в ней описания являются слишком краткими и сухими. Например, в ней указано, что шахматы — это полностью наблюдаемая среда, но строго говоря, это утверждение является ложным, поскольку некоторые правила, касающиеся рокировки, взятия пешки на проходе и объявления ничьи при повторении ходов, требуют запоминания определенных фактов об истории игры, которые нельзя выявить из анализа позиции на доске. Но эти исключения из определения наблюдаемости, безусловно, являются незначительными по сравнению с теми необычными ситуациями, с которыми сталкивается автоматизированный водитель такси, интерактивная система преподавания английского языка или медицинская диагностическая система.

Таблица 2.4. Примеры вариантов проблемной среды и их характеристик

Проблемная среда	Наблюдаема полностью или частично	Детерминированная, стратегическая или стохастическая	Эпизодическая или последовательная	Статическая, динамическая или полудинамическая	Дискретная или непрерывная	Одноагентная или мультиагентная
Решение кроссворда	Полностью наблюдаемая	Детерминированная	Последовательная	Статическая	Дискретная	Одноагентная
Игра в шахматы с контролем времени	Полностью наблюдаемая	Стохастическая	Последовательная	Полудинамическая	Дискретная	Мультиагентная
Игра в покер	Частично наблюдаемая	Стохастическая	Последовательная	Статическая	Дискретная	Мультиагентная
Игра в нарды	Полностью наблюдаемая	Стохастическая	Последовательная	Статическая	Дискретная	Мультиагентная
Вожделение такси	Частично наблюдаемая	Стохастическая	Последовательная	Динамическая	Непрерывная	Мультиагентная
Медицинская диагностика	Частично наблюдаемая	Стохастическая	Последовательная	Динамическая	Непрерывная	Одноагентная
Анализ изображений	Полностью наблюдаемая	Детерминированная	Эпизодическая	Полудинамическая	Непрерывная	Одноагентная
Робот-сортировщик деталей	Частично наблюдаемая	Стохастическая	Эпизодическая	Динамическая	Непрерывная	Одноагентная
Контроллер очистительной установки	Частично наблюдаемая	Стохастическая	Последовательная	Динамическая	Непрерывная	Одноагентная
Интерактивная программа, обучающая английскому языку	Частично наблюдаемая	Стохастическая	Последовательная	Динамическая	Дискретная	Мультиагентная

Некоторые другие ответы в этой таблице зависят от того, как определена проблемная среда. Например, в ней задача медицинского диагноза определена как одноагентная, поскольку сам процесс развития заболевания у пациента нецелесообразно моделировать в качестве агента, но системе медицинской диагностики иногда приходится сталкиваться с пациентами, не желающими принимать ее рекомендации, и со скептически настроенным персоналом, поэтому ее среда может иметь мультиагентный аспект. Кроме того, медицинская диагностика является эпизодической, если она рассматривается как задача выбора диагноза на основе анализа перечня симптомов, но эта проблема становится последовательной, если решаемая при этом задача может включать выработку рекомендаций по выполнению ряда лабораторных исследований, оценку прогресса в ходе лечения и т.д. К тому же многие варианты среды являются эпизодическими на более высоких уровнях по сравнению с отдельными действиями агента. Например, шахматный турнир состоит из ряда игр; каждая игра является эпизодом, поскольку (вообще говоря) от ходов, сделанных в предыдущей игре, не зависит то, как повлияют на общую производительность агента ходы, сделанные им в текущей игре. С другой стороны, принятие решений в одной и той же игре, безусловно, происходит последовательно.

Репозиторий кода, который относится к данной книге (aima.cs.berkeley.edu), включает реализации многих вариантов среды, наряду с имитатором среды общего назначения, который помещает одного или нескольких агентов в моделируемую среду, наблюдает за их поведением в течение определенного времени и оценивает их действия в соответствии с заданными показателями производительности. Такие эксперименты часто выполняются применительно не к одному варианту среды, а ко многим вариантам, сформированным на основе некоторого **класса вариантов среды**. Например, чтобы оценить действия водителя такси в моделируемой ситуации дорожного движения, может потребоваться провести много сеансов моделирования с различными условиями трафика, освещения и погоды. Если бы мы спроектировали этого агента для одного сценария, то могли бы лучше воспользоваться специфическими свойствами данного конкретного случая, но не имели бы возможности определить приемлемый проект решения задачи автоматизированного вождения в целом. По этой причине репозиторий кода включает также **генератор вариантов среды** для каждого класса вариантов среды; этот генератор выбирает определенные варианты среды (с некоторой вероятностью), в которых выполняется проверка агента. Например, генератор вариантов среды пылесоса инициализирует случайным образом такие исходные данные, как распределение мусора и местонахождение агента. Дело в том, что наибольший интерес представляет то, какую среднюю производительность будет иметь данный конкретный агент в некотором классе вариантов среды. Рациональный агент для определенного класса вариантов среды максимизирует свою среднюю производительность. Процесс разработки класса вариантов среды и оценки в них различных агентов иллюстрируется в упр. 2.7–2.12.

2.4. СТРУКТУРА АГЕНТОВ

До сих пор в этой книге свойства агентов рассматривались на основании анализа их поведения — действий, выполняемых агентом после получения любой заданной последовательности актов восприятия. Теперь нам поневоле придется сменить тему

и перейти к описанию того, как организовано их внутреннее функционирование. Задача искусственного интеллекта состоит в разработке **программы агента**, которая реализует функцию агента, отображая восприятия на действия. Предполагается, что эта программа должна работать в своего рода вычислительном устройстве с физическими датчиками и исполнительными механизмами; в целом эти компоненты именуется в данной книге **архитектурой**, а структура агента условно обозначается следующей формулой:

$$\text{Агент} = \text{Архитектура} + \text{Программа}$$

Очевидно, что выбранная программа должна быть подходящей для этой архитектуры. Например, если в программе осуществляется выработка рекомендаций по выполнению таких действий, как *walk* (ходьба), то в архитектуре целесообразно предусмотреть использование опорно-двигательного аппарата. Архитектура может представлять собой обычный персональный компьютер или может быть воплощена в виде роботизированного автомобиля с несколькими бортовыми компьютерами, видеокамерами и другими датчиками. Вообще говоря, архитектура обеспечивает передачу в программу результатов восприятия, полученных от датчиков, выполнение программы и передачу исполнительным механизмам вариантов действий, выбранных программой, по мере их выработки. Основная часть данной книги посвящена проектированию программ агентов, а главы 24 и 25 касаются непосредственно датчиков и исполнительных механизмов.

Программы агентов

Все программы агентов, которые будут разработаны в этой книге, имеют одну и ту же структуру: они принимают от датчиков в качестве входных данных результаты текущего восприятия и возвращают исполнительным механизмам выбранный вариант действия⁶. Необходимо указать на различие между программой агента, которая принимает в качестве входных данных результаты текущего восприятия, и функцией агента, которая принимает на входе всю историю актов восприятия. Программа агента получает в качестве входных данных только результаты текущего восприятия, поскольку больше ничего не может узнать из своей среды; если действия агента зависят от всей последовательности актов восприятия, то агент должен сам запоминать результаты этих актов восприятия.

Для описания программ агентов будет применяться простой язык псевдокода, который определен в приложении Б. (Оперативный репозиторий кода содержит реализации на реальных языках программирования.) Например, в листинге 2.1 показана довольно несложная программа агента, которая регистрирует последовательность актов восприятия, а затем использует полученную последовательность для доступа по индексу к таблице действий и определения того, что нужно сделать. Таблица явно отображает функцию агента, воплощаемую данной программой агента. Чтобы создать рационального агента таким образом, проектировщики должны сформировать

⁶ Существуют и другие варианты структуры программы агента, например, можно было бы использовать в виде программ агентов **сопроцедуры**, которые действуют асинхронно со средой. Каждая такая сопроцедура имеет входной и выходной порты, а ее работа организована в виде цикла, в котором из входного порта считываются результаты восприятий, а в выходной порт записываются варианты действий.

таблицу, которая содержит подходящее действие для любой возможной последовательности актов восприятия.

Листинг 2.1. Программа Table-Driven-Agent, которая вызывается после каждого восприятия новых данных и каждый раз возвращает вариант действия; программа регистрирует последовательность актов восприятия с использованием своей собственной закрытой структуры данных

```

function Table-Driven-Agent(percept) returns действие action
  static: percepts, последовательность актов восприятия,
           первоначально пустая
           table, таблица действий, индексированная по
           последовательностям актов восприятия и
           полностью заданная с самого начала

  добавить результаты восприятия percept к концу
  последовательности percepts
  action ← Lookup(percepts, table)
  return action

```

Анализ того, почему такой подход к созданию агента, основанный на использовании таблицы, обречен на неудачу, является весьма поучительным. Допустим, что \mathcal{P} — множество возможных актов восприятия, а T — срок существования агента (общее количество актов восприятия, которое может быть им получено). Поисковая таблица будет содержать

$$\sum_{t=1}^T |\mathcal{P}|^t \text{ записей.}$$

Рассмотрим автоматизированное такси: визуальные входные данные от одной телекамеры поступают со скоростью примерно 27 мегабайтов в секунду (30 кадров в секунду, 640×480 пикселей с 24 битами информации о цвете). Согласно этим данным поисковая таблица, рассчитанная на 1 час вождения, должна содержать количество записей, превышающее $10^{250\,000\,000\,000}$. И даже поисковая таблица для шахмат (крошечного, хорошо изученного фрагмента реального мира) имела бы, по меньшей мере, 10^{150} записей. Ошеломляющий размер этих таблиц (при том что количество атомов в наблюдаемой вселенной не превышает 10^{80}) означает, что, во-первых, ни один физический агент в нашей вселенной не имеет пространства для хранения такой таблицы, во-вторых, проектировщик не сможет найти достаточно времени для создания этой таблицы, в-третьих, ни один агент никогда не сможет обучиться тому, что содержится во всех правильных записях этой таблицы, на основании собственного опыта, и, в-четвертых, даже если среда достаточно проста для того, чтобы можно было создать таблицу приемлемых размеров, все равно у проектировщика нет руководящих сведений о том, как следует заполнять записи подобной таблицы.

Несмотря на все сказанное, программа Table-Driven-Agent выполняет именно то, что от нее требуется: она реализует желаемую функцию агента. Основная сложность, стоящая перед искусственным интеллектом как научным направлением, состоит в том, чтобы узнать, как создавать программы, которые в рамках возможного вырабатывают рациональное поведение с использованием небольшого объема кода, а не большого количества записей таблицы. Существует множество примеров, показывающих, что такая задача может быть выполнена успешно в других областях; например, огромные таблицы квадратных корней, использовавшиеся инженерами и

школьниками до 1970-х годов, теперь заменены работающей в электронных калькуляторах программой из пяти строк, в которой применяется метод Ньютона. Вопрос заключается в том, может ли искусственный интеллект сделать для интеллектуального поведения в целом то, что Ньютон сделал для упрощения вычисления квадратных корней? Авторы данной книги полагают, что ответ на этот вопрос является положительным.

В остальной части этого раздела рассматриваются четыре основных вида программ агентов, которые воплощают принципы, лежащие в основе почти всех интеллектуальных систем:

- простые рефлексные агенты;
- рефлексные агенты, основанные на модели;
- агенты, действующие на основе цели;
- агенты, действующие на основе полезности.

Затем приведено описание в общих терминах того, как преобразовать агентов всех этих типов в обучающихся агентов.

Простые рефлексные агенты

Простейшим видом агента является **простой рефлексный агент**. Подобные агенты выбирают действия на основе текущего акта восприятия, игнорируя всю остальную историю актов восприятия. Например, агент-пылесос, для которого результаты табуляции функции агента приведены в табл. 2.1, представляет собой простой рефлексный агент, поскольку его решения основаны только на информации о текущем местонахождении и о том, содержит ли оно мусор. Программа для данного агента приведена в листинге 2.2.

Листинг 2.2. Программа простого рефлексного агента в среде пылесоса с двумя состояниями. Эта программа реализует функцию агента, которая табулирована в табл. 2.1

```
function Reflex-Vacuum-Agent([location,status]) returns действие  
action
```

```
    if status = Dirty then return Suck  
    else if location = A then return Right  
    else if location = B then return Left
```

Обратите внимание на то, что эта программа агента-пылесоса действительно очень мала по сравнению с соответствующей таблицей. Наиболее очевидное сокращение обусловлено тем, что в ней игнорируется история актов восприятия, в результате чего количество возможных вариантов сокращается от 4^T просто до 4. Дополнительное небольшое сокращение обусловлено тем фактом, что если в текущем квадрате имеется мусор, то выполняемое при этом действие не зависит от местонахождения пылесоса.

Представьте себя на месте водителя автоматизированного такси. Если движущийся впереди автомобиль тормозит и загораются его тормозные огни, то вы должны заметить это и тоже начать торможение. Иными словами, над визуальными входными данными выполняется определенная обработка для выявления условия,

которое обозначается как “*car-in-front-is-braking*” (движущийся впереди автомобиль тормозит). Затем это условие активизирует некоторую связь с действием “*initiate-braking*” (начать торможение), установленную в программе агента. Такая связь называется ~~э~~ **правилом условие—действие**⁷ и записывается следующим образом:

```
if car-in-front-is-braking then initiate-braking
```

Люди также используют большое количество таких связей, причем некоторые из них представляют собой сложные отклики, освоенные в результате обучения (как при вождении автомобиля), а другие являются врожденными рефлексами (такими как моргание, которое происходит при приближении к глазу постороннего предмета). В разных главах данной книги будет показано несколько различных способов, с помощью которых можно организовать обучение агента и реализацию таких связей.

Программа, приведенная в листинге 2.2, специализирована для одной конкретной среды пылесоса. Более общий и гибкий подход состоит в том, чтобы вначале создать интерпретатор общего назначения для правил условие—действие, а затем определить наборы правил для конкретной проблемной среды. На рис. 2.3 приведена структура такой общей программы в схематической форме и показано, каким образом правила условие—действие позволяют агенту создать связь от восприятия к действию. (Не следует беспокоиться, если такой способ покажется тривиальным; вскоре он обнаружит намного более интересные возможности.) В подобных схемах для обозначения текущего внутреннего состояния процесса принятия решения агентом используются прямоугольники, а для представления фоновой информации, применяемой в этом процессе, служат овалы. Программа этого агента, которая также является очень простой, приведена в листинге 2.3. Функция *Interpret-Input* вырабатывает абстрагированное описание текущего состояния по результатам восприятия, а функция *Rule-Match* возвращает первое правило во множестве правил, которое соответствует заданному описанию состояния. Следует отметить, что приведенное здесь изложение в терминах “правил” и “соответствия” является чисто концептуальным; фактические реализации могут быть настолько простыми, как совокупность логических элементов, реализующих логическую схему.

Листинг 2.3. Программа простого рефлексного агента, который действует согласно правилу, условие которого соответствует текущему состоянию, определяемому результатом восприятия

```
function Simple-Reflex-Agent(percept) returns действие action
    static: rules, множество правил условие—действие

    state ← Interpret-Input(percept)
    rule ← Rule-Match(state, rules)
    action ← Rule-Action[rule]
    return action
```

Простые рефлексные агенты характеризуются той замечательной особенностью, что они чрезвычайно просты, но зато обладают весьма ограниченным интеллектом. Агент, программа которого приведена в листинге 2.3, работает, ~~с~~ **только если правильное решение может быть принято на основе исключительно текущего восприятия,**

⁷ Эти связи называются также **правилами ситуация—действие**, **продукциями** или **правилами if-then**.

иначе говоря, только если среда является полностью наблюдаемой. Внесение даже небольшой доли ненаблюдаемости может вызвать серьезное нарушение его работы. Например, в приведенном выше правиле торможения принято предположение, что условие *car-in-front-is-braking* может быть определено из текущего восприятия (текущего видеоизображения), если движущийся автомобиль имеет тормозной сигнал, расположенный на центральном месте среди других сигналов. К сожалению, некоторые более старые модели имеют другие конфигурации задних фар, тормозных сигналов, габаритных огней, сигналов торможения и сигналов поворота, поэтому не всегда возможно определить из единственного изображения, тормозит ли этот автомобиль или нет. Простой рефлексный агент, ведущий свой автомобиль вслед за таким автомобилем, либо будет постоянно тормозить без всякой необходимости, либо, что еще хуже, вообще не станет тормозить.



Рис. 2.3. Схематическое изображение структуры простого рефлексного агента

Возникновение аналогичной проблемы можно обнаружить и в мире пылесоса. Предположим, что в простом рефлексном агенте-пылесосе испортился датчик местонахождения и работает только датчик мусора. Такой агент получает только два возможных восприятия: *[Dirty]* и *[Clean]*. Он может выполнить действие *Suck* в ответ на восприятие *[Dirty]*, а что он должен делать в ответ на восприятие *[Clean]*? Выполнение движения *Left* завершится отказом (на неопределенно долгое время), если окажется, что он начинает это движение с квадрата *A*, а если он начинает движение с квадрата *B*, то завершится отказом на неопределенно долгое время движение *Right*. Для простых рефлексных агентов, действующих в частично наблюдаемых вариантах среды, часто бывают неизбежными бесконечные циклы.

Выход из бесконечных циклов становится возможным, если агент обладает способностью **рандомизировать** свои действия (вводить в них элемент случайности). Например, если агент-пылесос получает результат восприятия *[Clean]*, то может подбросить монету, чтобы выбрать между движениями *Left* и *Right*. Легко показать, что агент достигнет другого квадрата в среднем за два этапа. Затем, если в этом квадрате имеется мусор, то пылесос его уберет и задача очистки будет выполнена.

Поэтому рандомизированный простой рефлексный агент может превзойти по своей производительности детерминированного простого рефлексного агента.

В разделе 2.3 уже упоминалось, что в некоторых мультиагентных вариантах среды может оказаться рациональным рандомизированное поведение правильного типа. А в одноагентных вариантах среды рандомизация обычно не является рациональной. Это лишь полезный трюк, который помогает простому рефлексному агенту в некоторых ситуациях, но в большинстве случаев можно добиться гораздо большего с помощью более сложных детерминированных агентов.

Рефлексные агенты, основанные на модели

Наиболее эффективный способ организации работы в условиях частичной наблюдаемости состоит в том, чтобы агент отслеживал ту часть мира, которая воспринимается им в текущий момент. Это означает, что агент должен поддерживать своего рода **внутреннее состояние**, которое зависит от истории актов восприятия и поэтому отражает по крайней мере некоторые из ненаблюдаемых аспектов текущего состояния. Для решения задачи торможения поддержка внутреннего состояния не требует слишком больших затрат — для этого достаточно сохранить предыдущий кадр, снятый видеокамерой, чтобы агент мог определить тот момент, когда два красных световых сигнала с обеих сторон задней части идущего впереди автомобиля загораются или гаснут одновременно. Для решения других задач вождения, таких как переход с одной полосы движения на другую, агент должен следить за тем, где находятся другие автомобили, если он не может видеть все эти автомобили одновременно.

Для обеспечения возможности обновления этой внутренней информации о состоянии в течение времени необходимо, чтобы в программе агента были закодированы знания двух видов. Во-первых, нужна определенная информация о том, как мир изменяется независимо от агента, например, о том, что автомобиль, идущий на обгон, обычно становится ближе, чем в какой-то предыдущий момент. Во-вторых, требуется определенная информация о том, как влияют на мир собственные действия агента, например, что при повороте агентом рулевого колеса по часовой стрелке автомобиль поворачивает вправо или что после проезда по автомагистрали в течение пяти минут на север автомобиль находится на пять миль севернее от того места, где он был пять минут назад. Эти знания о том, “как работает мир” (которые могут быть воплощены в простых логических схемах или в сложных научных теориях) называются **моделью мира**. Агент, в котором используется такая модель, называется **агентом, основанным на модели**.

На рис. 2.4 приведена структура рефлексного агента, действующего с учетом внутреннего состояния, и показано, как текущее восприятие комбинируется с прежним внутренним состоянием для выработки обновленного описания текущего состояния. Программа такого агента приведена в листинге 2.4. В этом листинге интерес представляет функция `Update-State`, которая отвечает за создание нового описания внутреннего состояния. Эта функция не только интерпретирует результаты нового восприятия в свете существующих знаний о состоянии, но и использует информацию о том, как изменяется мир, для слежения за невидимыми частями мира, поэтому должна учитывать информацию о том, как действия агента влияют на состояние мира. Более подробные примеры приведены в главах 10 и 17.

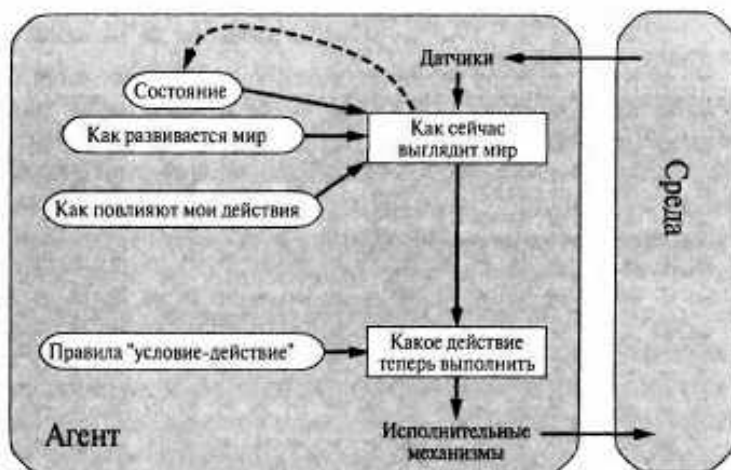


Рис. 2.4. Рефлективный агент, основанный на модели

Листинг 2.4. Рефлективный агент, основанный на модели, который следит за текущим состоянием мира с использованием внутренней модели, затем выбирает действие таким же образом, как и простой рефлективный агент

```
function Reflex-Agent-With-State(percept) returns действие action
  static: state, описание текущего состояния мира
         rules, множество правил условие-действие
         action, последнее по времени действие;
              первоначально не определено

  state ← Update-State(state, action, percept)
  rule ← Rule-Match(state, rules)
  action ← Rule-Action[rule]
  return action
```

Агенты, основанные на цели

Знаний о текущем состоянии среды не всегда достаточно для принятия решения о том, что делать. Например, на перекрестке дорог такси может повернуть налево, повернуть направо или ехать прямо. Правильное решение зависит от того, куда должно попасть это такси. Иными словами, агенту требуется не только описание текущего состояния, но и своего рода информация о **цели**, которая описывает желаемые ситуации, такие как доставка пассажира в место назначения. Программа агента может комбинировать эту информацию с информацией о результатах возможных действий (с такой же информацией, как и та, что использовалась при обновлении внутреннего состояния рефлективного агента) для выбора действий, позволяющих достичь этой цели. Структура агента, действующего на основе цели, показана на рис. 2.5.

Иногда задача выбора действия на основе цели решается просто, когда достижение цели немедленно становится результатом единственного действия, а иногда эта задача становится более сложной, и агенту требуется рассмотреть длинные последовательности движений и поворотов, чтобы найти способ достижения цели. Подобными искусствами искусственного интеллекта, посвященными выработке последовательно-

стей действий, позволяющих агенту достичь его целей, являются **поиск** (главы 3–6) и **планирование** (главы 11 и 12).

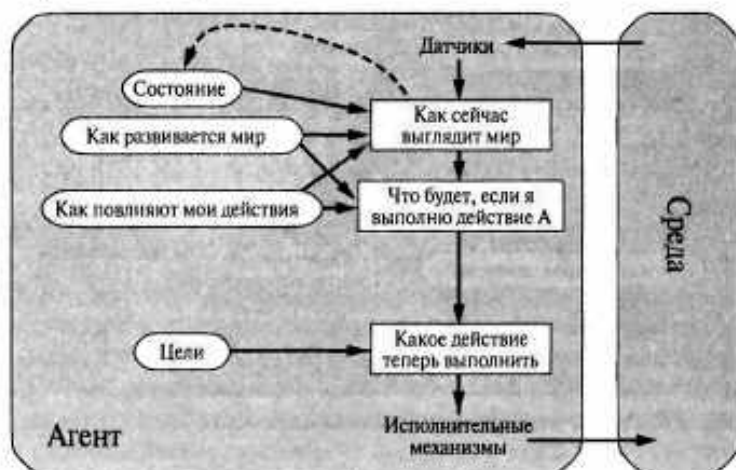


Рис. 2.5. Агент, основанный на модели и на цели. Он следит за состоянием мира, а также за множеством целей, которых он пытается достичь, и выбирает действие, позволяющее (в конечном итоге) добиться достижения этих целей

Следует учитывать, что процедура принятия решений такого рода имеет фундаментальные отличия от описанной выше процедуры применения правил условие–действие, поскольку в ней приходится размышлять о будущем, отвечая на два вопроса: “Что произойдет, если я сделаю то-то и то-то?” и “Позволит ли это мне достичь удовлетворения?” В проектах рефлексных агентов такая информация не представлена явно, поскольку встроенные правила устанавливают непосредственное соответствие между восприятиями и действиями. Рефлексный агент тормозит, увидев сигналы торможения движущегося впереди автомобиля, а агент, основанный на цели, может рассудить, что если на движущемся впереди автомобиле загорелись тормозные огни, то он замедляет свое движение. Учитывая принцип, по которому обычно изменяется этот мир, для него единственным действием, позволяющим достичь такой цели, как предотвращение столкновения с другими автомобилями, является торможение.

Хотя на первый взгляд кажется, что агент, основанный на цели, менее эффективен, он является более гибким, поскольку знания, на которые опираются его решения, представлены явно и могут быть модифицированы. Если начинается дождь, агент может обновить свои знания о том, насколько эффективно теперь будут работать его тормоза; это автоматически вызывает изменение всех соответствующих правил поведения с учетом новых условий. Для рефлексного агента, с другой стороны, в таком случае пришлось бы переписать целый ряд правил условие–действие. Поведение агента, основанного на цели, можно легко изменить, чтобы направить его в другое место, а правила рефлексного агента, которые указывают, где поворачивать и где ехать прямо, окажутся применимыми только для единственного места назначения; для того чтобы этого агента можно было направить в другое место, все эти правила должны быть заменены.

Агенты, основанные на полезности

В действительности в большинстве вариантов среды для выработки высококачественного поведения одного лишь учета целей недостаточно. Например, обычно существует много последовательностей действий, позволяющих такси добраться до места назначения (и тем самым достичь поставленной цели), но некоторые из этих последовательностей обеспечивают более быструю, безопасную, надежную или недорогую поездку, чем другие. Цели позволяют провести лишь жесткое бинарное различие между состояниями “удовлетворенности” и “неудовлетворенности”, тогда как более общие показатели производительности должны обеспечивать сравнение различных состояний мира в точном соответствии с тем, насколько удовлетворенным станет агент, если их удастся достичь. Поскольку понятие “удовлетворенности” представляется не совсем научным, чаще применяется терминология, согласно которой состояние мира, более предпочтительное по сравнению с другим, рассматривается как имеющее более высокую \propto полезность для агента⁸.

\propto **Функция полезности** отображает состояние (или последовательность состояний) на вещественное число, которое обозначает соответствующую степень удовлетворенности агента. Полная спецификация функции полезности обеспечивает возможность принимать рациональные решения в описанных ниже двух случаях, когда этого не позволяют сделать цели. Во-первых, если имеются конфликтующие цели, такие, что могут быть достигнуты только некоторые из них (например, или скорость, или безопасность), то функция полезности позволяет найти приемлемый компромисс. Во-вторых, если имеется несколько целей, к которым может стремиться агент, но ни одна из них не может быть достигнута со всей определенностью, то функция полезности предоставляет удобный способ взвешенной оценки вероятности успеха с учетом важности целей.

В главе 16 будет показано, что любой рациональный агент должен вести себя так, как если бы он обладал функцией полезности, ожидаемое значение которой он пытается максимизировать. Поэтому агент, обладающий явно заданной функцией полезности, имеет возможность принимать рациональные решения и способен делать это с помощью алгоритма общего назначения, не зависящего от конкретной максимизируемой функции полезности. Благодаря этому “глобальное” определение рациональности (согласно которому рациональными считаются функции агента, имеющие наивысшую производительность) преобразуется в “локальное” ограничение на проекты рациональных агентов, которое может быть выражено в виде простой программы.

Структура агента, действующего с учетом полезности, показана на рис. 2.6. Программы агентов, действующих с учетом полезности, приведены в части V, которая посвящена проектированию агентов, принимающих решения, способных учитывать неопределенность, свойственную частично наблюдаемым вариантам среды.

⁸ Термин “полезность” имеет англоязычный эквивалент “utility”, который в данном контексте обозначает “свойство быть полезным”, а не электростанцию или предприятие, предоставляющее коммунальные услуги.

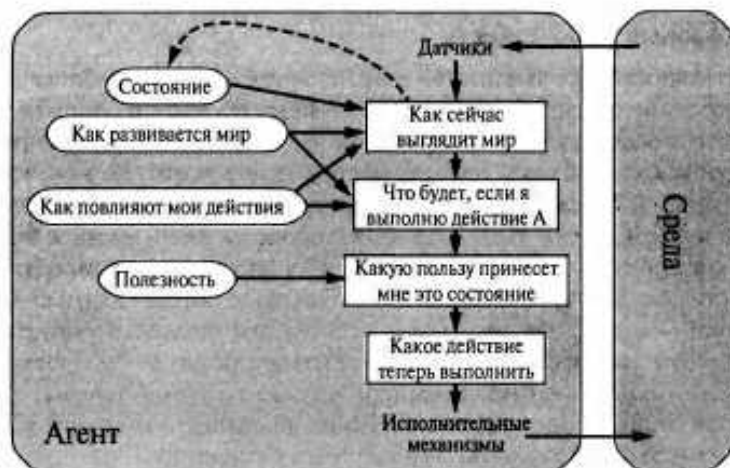


Рис. 2.6. Агент, основанный на модели и на полезности. В нем модель мира используется наряду с функцией полезности, которая измеряет предпочтения агента применительно к состояниям мира. Затем агент выбирает действие, которое ведет к наилучшей ожидаемой полезности. Для вычисления ожидаемой полезности выполняется усреднение по всем возможным результирующим состояниям с учетом коэффициента, определяющего вероятность каждого результата

Обучающиеся агенты

Выше были описаны программы агентов, в которых применяются различные методы выбора действий. Но до сих пор еще не были приведены сведения о том, как создаются программы агентов. В своей знаменитой ранней статье Тьюринг [1520] проанализировал идею о том, как фактически должно осуществляться программирование предложенных им интеллектуальных машин вручную. Он оценил объем работы, который для этого потребуются, и пришел к такому выводу: “Желательно было бы иметь какой-то более продуктивный метод”. Предложенный им метод заключался в том, что необходимо создавать обучающиеся машины, а затем проводить их обучение. Теперь этот метод стал доминирующим методом создания наиболее современных систем во многих областях искусственного интеллекта. Как отмечалось выше, обучение имеет еще одно преимущество: оно позволяет агенту функционировать в первоначально неизвестных ему вариантах среды и становится более компетентным по сравнению с тем, что могли бы позволить только его начальные знания. В данном разделе кратко представлены основные сведения об обучающихся агентах. Существующие возможности и методы обучения агентов конкретных типов рассматриваются почти в каждой главе данной книги, а в части VI более подробно описываются сами алгоритмы обучения.

Как показано на рис. 2.7, структура обучающегося агента может подразделяться на четыре концептуальных компонента. Наиболее важное различие наблюдается между **обучающим компонентом**, который отвечает за внесение усовершенствований, и **производительным компонентом**, который обеспечивает выбор внешних действий. Производительным компонентом является то, что до сих пор в данной книге рассматривалось в качестве всего агента: он получает воспринимаемую ин-