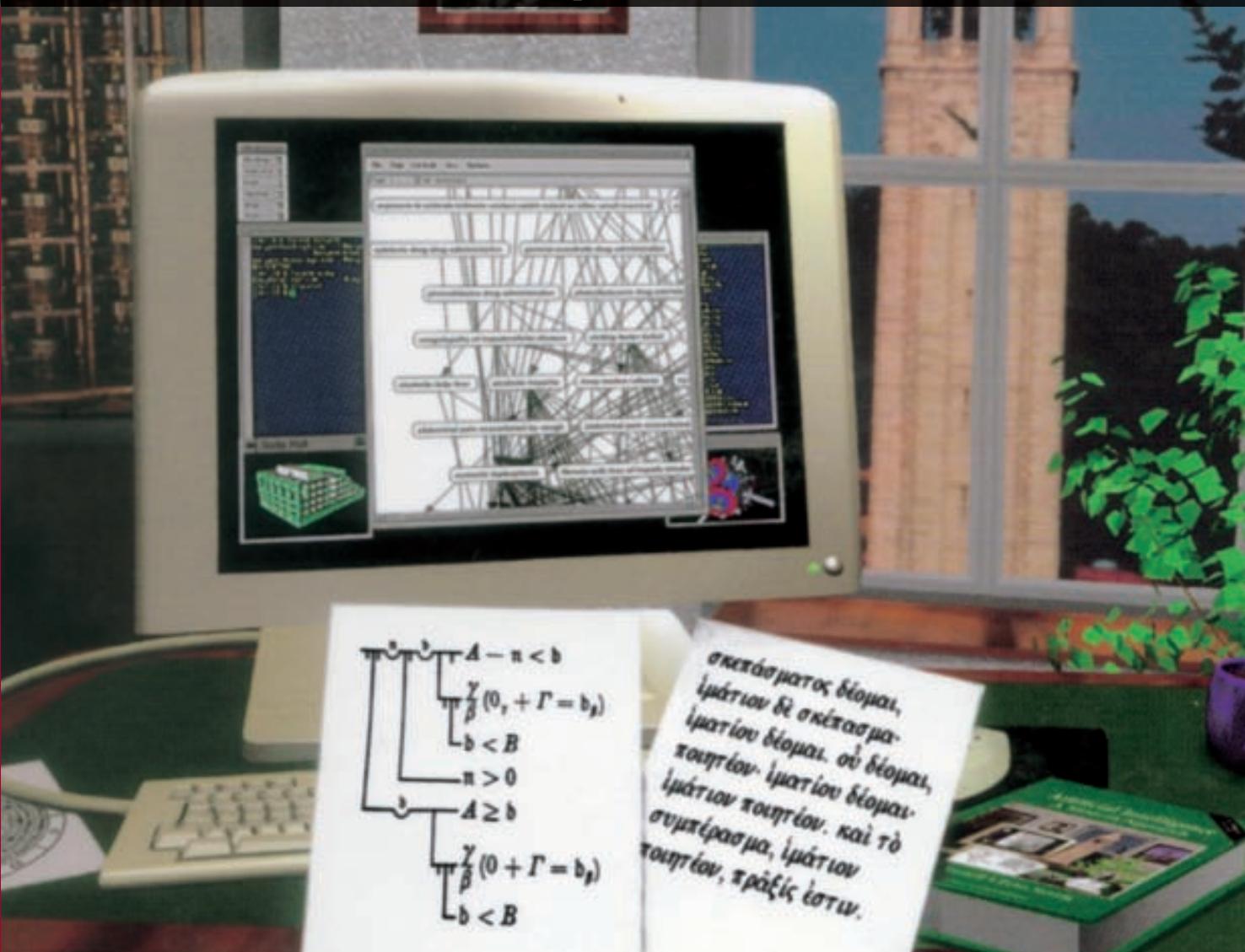


2^a Edición

Inteligencia Artificial

Un Enfoque Moderno



INTELIGENCIA ARTIFICIAL

UN ENFOQUE MODERNO

Segunda edición

INTELIGENCIA ARTIFICIAL

UN ENFOQUE MODERNO

Segunda edición

Stuart J. Russell y Peter Norvig

Traducción:

Juan Manuel Corchado Rodríguez

Facultad de Ciencias

Universidad de Salamanca

**Fernando Martín Rubio, José Manuel Cadenas Figueredo,
Luis Daniel Hernández Molinero y Enrique Paniagua Arís**

Facultad de Informática

Universidad de Murcia

Raquel Fuentetaja Pinzán y Mónica Robledo de los Santos

Universidad Pontificia de Salamanca, campus Madrid

Ramón Rizo Aldeguer

Escuela Politécnica Superior

Universidad de Alicante

Revisión técnica:

Juan Manuel Corchado Rodríguez

Facultad de Ciencias

Universidad de Salamanca

Fernando Martín Rubio

Facultad de Informática

Universidad de Murcia

Andrés Castillo Sanz y María Luisa Díez Plata

Facultad de Informática

Universidad Pontificia de Salamanca, campus Madrid

Coordinación general de la traducción y revisión técnica:

Luis Joyanes Aguilar

Facultad de Informática

Universidad Pontificia de Salamanca, campus Madrid



Madrid • México • Santafé de Bogotá • Buenos Aires • Caracas • Lima
Montevideo • San Juan • San José • Santiago • São Paulo • White Plains

Datos de catalogación bibliográfica	
RUSSELL, S. J.; NORVIG, P.	
INTELIGENCIA ARTIFICIAL. UN ENFOQUE MODERNO	
<i>Segunda edición</i>	
PEARSON EDUCACIÓN, S.A., Madrid, 2004	
ISBN: 978-84-205-4003-0	
Materia: Informática 681.3	
Formato 195 × 250	Páginas: 1240

Todos los derechos reservados.

Queda prohibida, salvo excepción prevista en la ley, cualquier forma de reproducción, distribución, comunicación pública y transformación de esta obra sin contar con autorización de los titulares de propiedad intelectual.

La infracción de los derechos mencionados puede ser constitutiva de delito contra la propiedad intelectual (*arts. 270 y sgts. Código Penal*).

DERECHOS RESERVADOS

© 2004 por PEARSON EDUCACIÓN, S.A.
Ribera del Loira, 28
28042 Madrid (España)

INTELIGENCIA ARTIFICIAL. UN ENFOQUE MODERNO. Segunda edición
RUSSELL, S. J.; NORVIG, P.

ISBN: 978-84-205-4003-0

Depósito Legal: M-14511-2008
Última reimpresión: 2008

PEARSON PRENTICE HALL es un sello editorial autorizado de PEARSON EDUCACIÓN, S.A.

Authorized translation from the English language edition, entitled *ARTIFICIAL INTELLIGENCE: A MODERN APPROACH*, 2nd edition by RUSSELL, STUART; NORVIG, PETER.

Published by Pearson Education, Inc, publishing as Prentice Hall.

© 2003. All rights reserved.

No part or 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 Pearson Education, Inc.

ISBN: 0-13-790395-2

Equipo editorial:

Editor: David Fayerman Aragón
Técnico editorial: Ana Isabel García Borro

Equipo de producción:

Director: José Antonio Clares
Técnico: José Antonio Hernán

Diseño de cubierta: Equipo de diseño de PEARSON EDUCACIÓN, S.A.

Composición: COPIBOOK, S.L.

Impreso por:

IMPRESO EN MÉXICO - PRINTED IN MEXICO



Contenido

Prólogo	XIX
Sobre los autores	XXV
1 Introducción	1
1.1 ¿Qué es la IA?	2
Comportamiento humano: el enfoque de la Prueba de Turing	3
Pensar como un humano: el enfoque del modelo cognitivo	3
Pensamiento racional: el enfoque de las «leyes del pensamiento»	4
Actuar de forma racional: el enfoque del agente racional	5
1.2 Los fundamentos de la inteligencia artificial	6
Filosofía (desde el año 428 a.C. hasta el presente)	6
Matemáticas (aproximadamente desde el año 800 al presente)	9
Economía (desde el año 1776 hasta el presente)	11
Neurociencia (desde el año 1861 hasta el presente)	12
Psicología (desde el año 1879 hasta el presente)	14
Ingeniería computacional (desde el año 1940 hasta el presente)	16
Teoría de control y cibernetica (desde el año 1948 hasta el presente)	17
Lingüística (desde el año 1957 hasta el presente)	18
1.3 Historia de la inteligencia artificial	19
Génesis de la inteligencia artificial (1943-1955)	19
Nacimiento de la inteligencia artificial (1956)	20
Entusiasmo inicial, grandes esperanzas (1952-1969)	21
Una dosis de realidad (1966-1973)	24
Sistemas basados en el conocimiento: ¿clave del poder? (1969-1979)	26
La IA se convierte en una industria (desde 1980 hasta el presente)	28
Regreso de las redes neuronales (desde 1986 hasta el presente)	29
IA se convierte en una ciencia (desde 1987 hasta el presente)	29
Emergencia de los sistemas inteligentes (desde 1995 hasta el presente)	31
1.4 El estado del arte	32
1.5 Resumen	33
Notas bibliográficas e históricas	34
Ejercicios	35
2 Agentes inteligentes	37
2.1 Agentes y su entorno	37
2.2 Buen comportamiento: el concepto de racionalidad	40
Medidas de rendimiento	40
Racionalidad	41

Omnisciencia, aprendizaje y autonomía	42
2.3 La naturaleza del entorno	44
Especificación del entorno de trabajo	44
Propiedades de los entornos de trabajo	47
2.4 Estructura de los agentes	51
Programas de los agentes	51
Agentes reactivos simples	53
Agentes reactivos basados en modelos	55
Agentes basados en objetivos	57
Agentes basados en utilidad	58
Agentes que aprenden	59
2.5 Resumen	62
Notas bibliográficas e históricas	63
Ejercicios	65
3 Resolver problemas mediante búsqueda	67
3.1 Agentes resolventes-problemas	67
Problemas y soluciones bien definidos	70
Formular los problemas	71
3.2 Ejemplos de problemas	72
Problemas de juguete	73
Problemas del mundo real	76
3.3 Búsqueda de soluciones	78
Medir el rendimiento de la resolución del problema	80
3.4 Estrategias de búsqueda no informada	82
Búsqueda primero en anchura	82
Búsqueda de costo uniforme	84
Búsqueda primero en profundidad	85
Búsqueda de profundidad limitada	87
Búsqueda primero en profundidad con profundidad iterativa	87
Búsqueda bidireccional	89
Comparación de las estrategias de búsqueda no informada	91
3.5 Evitar estados repetidos	91
3.6 Búsqueda con información parcial	94
Problemas sin sensores	95
Problemas de contingencia	96
3.7 Resumen	97
Notas bibliográficas e históricas	98
Ejercicios	100
4 Búsqueda informada y exploración	107
4.1 Estrategias de búsqueda informada (heurísticas)	107
Búsqueda voraz primero el mejor	108
Búsqueda A*: minimizar el costo estimado total de la solución	110
Búsqueda heurística con memoria acotada	115
Aprender a buscar mejor	118
4.2 Funciones heurísticas	119
El efecto de la precisión heurística en el rendimiento	120
Inventar funciones heurísticas admisibles	121
Aprendizaje de heurísticas desde la experiencia	124
4.3 Algoritmos de búsqueda local y problemas de optimización	125

Búsqueda de ascensión de colinas	126
Búsqueda de temple simulado	129
Búsqueda por haz local	131
Algoritmos genéticos	131
4.4 Búsqueda local en espacios continuos	136
4.5 Agentes de búsqueda <i>online</i> y ambientes desconocidos	138
Problemas de búsqueda en línea (<i>online</i>)	138
Agentes de búsqueda en línea (<i>online</i>)	141
Búsqueda local en línea (<i>online</i>)	142
Aprendizaje en la búsqueda en línea (<i>online</i>)	144
4.6 Resumen	145
Notas bibliográficas e históricas	146
Ejercicios	151
5 Problemas de satisfacción de restricciones	155
5.1 Problemas de satisfacción de restricciones	155
5.2 Búsqueda con vuelta atrás para PSR	159
Variable y ordenamiento de valor	162
Propagación de la información a través de las restricciones	163
Comprobación hacia delante	163
Propagación de restricciones	164
Manejo de restricciones especiales	166
Vuelta atrás inteligente: mirando hacia atrás	167
5.3 Búsqueda local para problemas de satisfacción de restricciones	169
5.4 La estructura de los problemas	171
5.5 Resumen	175
Notas bibliográficas e históricas	176
Ejercicios	178
6 Búsqueda entre adversarios	181
6.1 Juegos	181
6.2 Decisiones óptimas en juegos	183
Estrategias óptimas	183
El algoritmo minimax	185
Decisiones óptimas en juegos multi-jugador	186
6.3 Poda alfa-beta	188
6.4 Decisiones en tiempo real imperfectas	191
Funciones de evaluación	192
Corte de la búsqueda	194
6.5 Juegos que incluyen un elemento de posibilidad	196
Evaluación de la posición en juegos con nodos de posibilidad	198
Complejidad del minimaxesperado	199
Juegos de cartas	200
6.6 Programas de juegos	202
6.7 Discusión	205
6.8 Resumen	207
Notas bibliográficas e históricas	208
Ejercicios	212
7 Agentes lógicos	217
7.1 Agentes basados en conocimiento	219
7.2 El mundo de <i>wumpus</i>	221

7.3	Lógica	224
7.4	Lógica proposicional: una lógica muy sencilla	229
	Sintaxis	229
	Semántica	230
	Una base de conocimiento sencilla	233
	Inferencia	233
	Equivalencia, validez y <i>satisfacibilidad</i>	235
7.5	Patrones de razonamiento en lógica proposicional	236
	Resolución	239
	Forma normal conjuntiva	241
	Un algoritmo de resolución	242
	Complejidad de la resolución	243
	Encadenamiento hacia delante y hacia atrás	244
7.6	Inferencia proposicional efectiva	248
	Un algoritmo completo con <i>backtracking</i> («vuelta atrás»)	248
	Algoritmos de búsqueda local	249
	Problemas duros de <i>satisfacibilidad</i>	251
7.7	Agentes basados en lógica proposicional	253
	Encontrar hoyos y <i>wumpus</i> utilizando la inferencia lógica	253
	Guardar la pista acerca de la localización y la orientación del agente	255
	Agentes basados en circuitos	256
	Una comparación	260
7.8	Resumen	261
	Notas bibliográficas e históricas	262
	Ejercicios	266
8	Lógica de primer orden	271
8.1	Revisión de la representación	271
8.2	Sintaxis y semántica de la lógica de primer orden	277
	Modelos en lógica de primer orden	277
	Símbolos e interpretaciones	278
	Términos	280
	Sentencias atómicas	281
	Sentencias compuestas	281
	Cuantificadores	281
	Cuantificador universal (\forall)	282
	Cuantificación existencial (\exists)	283
	Cuantificadores anidados	284
	Conexiones entre \forall y \exists	285
	Igualdad	286
8.3	Utilizar la lógica de primer orden	287
	Aserciones y peticiones en lógica de primer orden	287
	El dominio del parentesco	288
	Números, conjuntos y listas	290
	El mundo de <i>wumpus</i>	292
8.4	Ingeniería del conocimiento con lógica de primer orden	295
	El proceso de ingeniería del conocimiento	296
	El dominio de los circuitos electrónicos	297
	Identificar la tarea	298
	Recopilar el conocimiento relevante	298
	Decidir el vocabulario	299

Codificar el conocimiento general del dominio	300
Codificar la instancia del problema específico	300
Plantear peticiones al procedimiento de inferencia	301
Depurar la base de conocimiento	301
8.5 Resumen	302
Notas bibliográficas e históricas	303
Ejercicios	304
9 Inferencia en lógica de primer orden 309	
9.1 Lógica proposicional vs. Lógica de primer orden	310
Reglas de inferencia para cuantificadores	310
Reducción a la inferencia proposicional	311
9.2 Unificación y sustitución	312
Una regla de inferencia de primer orden	313
Unificación	314
Almacenamiento y recuperación	315
9.3 Encadenamiento hacia delante	318
Cláusulas positivas de primer orden	318
Un algoritmo sencillo de encadenamiento hacia delante	320
Encadenamiento hacia delante eficiente	322
Emparejar reglas con los hechos conocidos	322
Encadenamiento hacia delante incremental	324
Hechos irrelevantes	326
9.4 Encadenamiento hacia atrás	326
Un algoritmo de encadenamiento hacia atrás	327
Programación lógica	328
Implementación eficiente de programas lógicos	330
Inferencia redundante y bucles infinitos	332
Programación lógica con restricciones	334
9.5 Resolución	335
Formas normales conjuntivas en lógica de primer orden	336
La regla de inferencia de resolución	338
Demostraciones de ejemplo	338
Completitud de la resolución	341
Manejar la igualdad	345
Estrategias de resolución	346
Resolución unitaria	346
Resolución mediante conjunto soporte	347
Resolución lineal	347
Subsunción	347
Demostradores de teoremas	348
Diseño de un demostrador de teoremas	348
Ampliar el Prolog	349
Demostradores de teoremas como asistentes	350
Usos prácticos de los demostradores de teoremas	351
9.6 Resumen	352
Notas bibliográficas e históricas	353
Ejercicios	359
10 Representación del conocimiento 363	
10.1 Ingeniería ontológica	363
10.2 Categoría y objetos	366

Objetos compuestos	368
Medidas	369
Sustancias y objetos	371
10.3 Acciones, situaciones y eventos	373
La ontología del cálculo de situaciones	373
Descripción de acciones en el cálculo de situaciones	375
Resolver el problema de la representación del marco	377
Resolver el problema de la inferencia del marco	379
El tiempo y el cálculo de eventos	380
Eventos generalizados	381
Procesos	383
Intervalos	384
Flujos y objetos	386
10.4 Eventos mentales y objetos mentales	387
Una teoría formal de creencias	387
Conocimiento y creencia	389
Conocimiento, tiempo y acción	390
10.5 El mundo de la compra por Internet	391
Comparación de ofertas	395
10.6 Sistemas de razonamiento para categorías	397
Redes semánticas	397
Lógica descriptiva	401
10.7 Razonamiento con información por defecto	402
Mundos abiertos y cerrados	403
Negación como fallo y semánticas de modelado estables	405
Circunscripción y lógica por defecto	406
10.8 Sistemas de mantenimiento de verdad	409
10.9 Resumen	411
Notas bibliográficas e históricas	412
Ejercicios	419
11 Planificación	427
11.1 El problema de planificación	428
El lenguaje de los problemas de planificación	429
Expresividad y extensiones	431
Ejemplo: transporte de carga aéreo	433
Ejemplo: el problema de la rueda de recambio	434
Ejemplo: el mundo de los bloques	434
11.2 Planificación con búsquedas en espacios de estado	436
Búsquedas hacia-delante en el espacio de estados	436
Búsquedas hacia-atrás en el espacio de estados	438
Heurísticas para la búsqueda en el espacio de estados	439
11.3 Planificación ordenada parcialmente	441
Ejemplo de planificación de orden parcial	445
Planificación de orden parcial con variables independientes	448
Heurísticas para planificación de orden parcial	449
11.4 Grafos de planificación	450
Grafos de planificación para estimación de heurísticas	453
El algoritmo GRAPHPLAN	454
Interrupción de GRAPHPLAN	457
11.5 Planificación con lógica proposicional	458

Descripción de problemas de planificación en lógica proposicional	458
Complejidad de codificaciones proposicionales	462
11.6 Análisis de los enfoques de planificación	463
11.7 Resumen	465
Notas bibliográficas e históricas	466
Ejercicios	469
12 Planificación y acción en el mundo real	475
12.1 Tiempo, planificación y recursos	475
Programación con restricción de recursos	478
12.2 Redes de planificación jerárquica de tareas	481
Representación de descomposición de acciones	482
Modificación de planificadores para su descomposición	484
Discusión	487
12.3 Planificación y acción en dominios no deterministas	490
12.4 Planificación condicional	493
Planificación condicional en entornos completamente observables	493
Planificación condicional en entornos parcialmente observables	498
12.5 Vigilancia de ejecución y replanificación	502
12.6 Planificación continua	507
12.7 Planificación multiagente	512
Cooperación: planes y objetivos conjuntos	512
Planificación condicional en entornos parcialmente observables	514
Mecanismos de coordinación	515
Mecanismos de coordinación	517
12.8 Resumen	517
Notas bibliográficas e históricas	518
Ejercicios	522
13 Incertidumbre	527
13.1 Comportamiento bajo incertidumbre	527
Manipulación del conocimiento incierto	528
Incertidumbre y decisiones racionales	530
Diseño de un agente de decisión teórico	531
13.2 Notación básica con probabilidades	532
Proposiciones	532
Sucesos atómicos	534
Probabilidad priori	534
Probabilidad condicional	536
13.3 Los axiomas de la probabilidad	537
Utilización de los axiomas de probabilidad	539
Por qué los axiomas de la probabilidad son razonables	540
13.4 Inferencia usando las distribuciones conjuntas totales	541
13.5 Independencia	544
13.6 La Regla de Bayes y su uso	546
Aplicación de la regla de Bayes: el caso sencillo	547
Utilización de la regla de Bayes: combinación de evidencia	548
13.7 El mundo <i>wumpus</i> revisado	550
13.8 Resumen	554
Notas bibliográficas e históricas	555
Ejercicios	557

14 Razonamiento probabilista	561
14.1 La representación del conocimiento en un dominio incierto	561
14.2 La semántica de las redes bayesianas	564
La representación de la distribución conjunta completa	564
Un método para la construcción de redes bayesianas	565
Compactación y ordenación de nodos	566
Relaciones de independencia condicional en redes bayesianas	568
14.3 Representación eficiente de las distribuciones condicionales	569
Redes bayesianas con variables continuas	571
14.4 Inferencia exacta en redes bayesianas	574
Inferencia por enumeración	575
El algoritmo de eliminación de variables	577
La complejidad de la inferencia exacta	580
Algoritmos basados en grupos	580
14.5 Inferencia aproximada en redes bayesianas	581
Métodos de muestreo directo	582
Muestreo por rechazo en redes bayesianas	583
Ponderación de la verosimilitud	585
Inferencia por simulación en cadenas de Markov	587
14.6 Extensión de la probabilidad a representaciones de primer orden	590
14.7 Otros enfoques al razonamiento con incertidumbre	595
Métodos basados en reglas para razonamiento con incertidumbre	596
Representación de la ignorancia: teoría de Dempster-Shafer	598
Representación de la vaguedad: conjuntos difusos y lógica difusa	599
14.8 Resumen	601
Notas bibliográficas e históricas	601
Ejercicios	606
15 Razonamiento probabilista en el tiempo	611
15.1 El tiempo y la incertidumbre	611
Estados y observaciones	612
Procesos estacionarios e hipótesis de Markov	613
15.2 Inferencia en modelos temporales	616
Filtrado y predicción	617
Suavizado	619
Encontrar la secuencia más probable	622
15.3 Modelos ocultos de Markov	624
Algoritmos matriciales simplificados	624
15.4 Filtros de Kalman	627
Actualización de distribuciones gaussianas	628
Un ejemplo unidimensional sencillo	629
El caso general	632
Aplicabilidad del filtrado de Kalman	633
15.5 Redes bayesianas dinámicas	635
Construcción de RBDs	636
Inferencia exacta en RBDs	640
Inferencia aproximada en RBDs	641
15.6 Reconocimiento del habla	645
Sonidos del habla	647
Palabras	649
Oraciones	651

Construcción de un reconocedor del habla	654
15.7 Resumen	656
Notas bibliográficas e históricas	656
Ejercicios	659
16 Toma de decisiones sencillas 663	
16.1 Combinación de creencias y deseos bajo condiciones de incertidumbre	664
16.2 Los fundamentos de la teoría de la utilidad	665
Restricciones sobre preferencias racionales	666
... y entonces apareció la utilidad	668
16.3 Funciones de utilidad	669
La utilidad del dinero	669
Escalas de utilidad y evaluación de la utilidad	671
16.4 Funciones de utilidad multiatributo	674
Predominio	674
Estructura de preferencia y utilidad multiatributo	677
Preferencias sin incertidumbre	677
Preferencias con incertidumbre	678
16.5 Redes de decisión	679
Representación de un problema de decisión mediante una red de decisión	679
Evaluación en redes de decisión	681
16.6 El valor de la información	682
Un ejemplo sencillo	682
Una fórmula general	683
Propiedades del valor de la información	685
Implementación de un agente recopilador de información	685
16.7 Sistemas expertos basados en la teoría de la decisión	686
16.8 Resumen	690
Notas bibliográficas e históricas	690
Ejercicios	692
17 Toma de decisiones complejas 697	
17.1 Problemas de decisión secuenciales	698
Un ejemplo	698
Optimalidad en problemas de decisión secuenciales	701
17.2 Iteración de valores	704
Utilidades de los estados	704
El algoritmo de iteración de valores	705
Convergencia de la iteración de valores	707
17.3 Iteración de políticas	710
17.4 Procesos de decisión de Markov parcialmente observables	712
17.5 Agentes basados en la teoría de la decisión	716
17.6 Decisiones con varios agentes: teoría de juegos	719
17.7 Diseño de mecanismos	729
17.8 Resumen	732
Notas bibliográficas e históricas	733
Ejercicios	736
18 Aprendizaje de observaciones 739	
18.1 Formas de aprendizaje	739
18.2 Aprendizaje inductivo	742

18.3	Aprender árboles de decisión	744
	Árboles de decisión como herramienta de desarrollo	744
	Expresividad de los árboles de decisión	745
	Inducir árboles de decisión a partir de ejemplos	746
	Elección de los atributos de test	750
	Valoración de la calidad del algoritmo de aprendizaje	752
	Ruido y sobreajuste	753
	Extensión de la aplicabilidad de los árboles de decisión	755
18.4	Aprendizaje de conjuntos de hipótesis	756
18.5	¿Por qué funciona el aprendizaje?: teoría computacional del aprendizaje	760
	¿Cuántos ejemplos se necesitan?	761
	Aprendizaje de listas de decisión	763
	Discusión	765
18.6	Resumen	766
	Notas bibliográficas e históricas	767
	Ejercicios	769
19	Conocimiento en el aprendizaje	773
19.1	Una formulación lógica del aprendizaje	773
	Ejemplos e hipótesis	774
	Búsqueda mejor-hipótesis-actual	776
	Búsqueda de mínimo compromiso	778
19.2	Conocimiento en el aprendizaje	782
	Algunos ejemplos sencillos	784
	Algunos esquemas generales	784
19.3	Aprendizaje basado en explicaciones	786
	Extraer reglas generales a partir de ejemplos	787
	Mejorar la eficiencia	789
19.4	Aprendizaje basado en información relevante	791
	Determinar el espacio de hipótesis	792
	Aprender y utilizar información relevante	792
19.5	Programación lógica inductiva	795
	Un ejemplo	795
	Métodos de aprendizaje inductivo de arriba a abajo (<i>Top-down</i>)	798
	Aprendizaje inductivo con deducción inversa	801
	Hacer descubrimientos con la programación lógica inductiva	803
18.6	Resumen	805
	Notas bibliográficas e históricas	806
	Ejercicios	809
20	Métodos estadísticos de aprendizaje	811
20.1	Aprendizaje estadístico	811
20.2	Aprendizaje con datos completos	815
	Aprendizaje del parámetro de máxima verosimilitud: modelos discretos	815
	Modelos de Bayes simples (Naive Bayes)	818
	Aprendizaje de parámetros de máxima verosimilitud: modelos continuos	819
	Aprendizaje de parámetros Bayesianos	821
	Aprendizaje de la estructura de las redes bayesianas	823
20.3	Aprendizaje con variables ocultas: el algoritmo EM	825
	Agrupamiento no supervisado: aprendizaje de mezclas de gaussianas	826
	Aprendizaje de redes bayesianas con variables ocultas	829

20.4	Aprendizaje basado en instancias	834
	Modelos de vecinos más cercanos	835
	Modelos núcleo	837
20.5	Redes neuronales	838
	Unidades en redes neuronales	839
	Estructuras de las redes	840
	Redes neuronales de una sola capa con alimentación-hacia-delante (perceptrones) ..	842
	Redes neuronales multicapa con alimentación hacia delante	846
	Aprendizaje de la estructura de las redes neuronales	851
20.6	Máquinas núcleo	851
20.7	Caso de estudio: reconocedor de dígitos escritos a mano	855
20.8	Resumen	857
	Notas bibliográficas e históricas	859
	Ejercicios	863
21	Aprendizaje por refuerzo	867
21.1	Introducción	867
21.2	Aprendizaje por refuerzo pasivo	869
	Estimación directa de la utilidad	870
	Programación dinámica adaptativa	871
	Aprendizaje de diferencia temporal	872
21.3	Aprendizaje por refuerzo activo	876
	Exploración	876
	Aprendizaje de una Función Acción-Valor	880
21.4	Generalización en aprendizaje por refuerzo	882
	Aplicaciones a juegos	885
	Aplicación a control de robots	886
21.5	Búsqueda de la política	887
21.6	Resumen	890
	Notas bibliográficas e históricas	891
	Ejercicios	894
22	La comunicación	897
22.1	La comunicación como acción	898
	Fundamentos del lenguaje	899
	Etapas de la comunicación	900
22.2	Una gramática formal para un fragmento del español	903
	El léxico de e0	904
	La Gramática de e0	904
22.3	Análisis sintáctico	905
	Análisis sintáctico eficiente	908
22.4	Gramáticas aumentadas	914
	Subcategorización del verbo	916
	Capacidad generativa de las gramáticas aumentadas	918
22.5	Interpretación semántica	919
	La semántica de un fragmento en español	920
	Tiempo y forma verbal	921
	Cuantificación	922

Interpretación pragmática	925
Generación de lenguajes con DCGs	926
22.6 Ambigüedad y desambigüedad	927
Desambiguación	929
22.7 Comprensión del discurso	930
Resolución por referencia	931
La estructura de un discurso coherente	932
22.8 Inducción gramatical	934
22.9 Resumen	936
Notas bibliográficas e históricas	937
Ejercicios	941
23 Procesamiento probabilístico del lenguaje	945
23.1 Modelos probabilísticos del lenguaje	945
Gramáticas probabilísticas independientes del contexto	949
Aprendizaje de probabilidades para PCFGs	950
Aprendizaje de la estructura de las reglas para PCFGs	951
23.2 Recuperación de datos	952
Evaluación de los Sistemas de RD	955
Refinamientos RD	956
Presentación de los conjuntos de resultados	957
Implementar sistemas RD	959
23.3 Extracción de la información	961
23.4 Traducción automática	964
Sistemas de traducción automáticos	966
Traducción automática estadística	967
Probabilidades de aprendizaje para la traducción automática	970
23.5 Resumen	972
Notas bibliográficas e históricas	972
Ejercicios	975
24 Percepción	979
24.1 Introducción	979
24.2 Formación de la imagen	981
Imágenes sin lentes: la cámara de orificio o pinhole	982
Sistemas de lentes	983
Luz: la fotometría de la formación de imágenes	983
Color: la espectrofotometría de la formación de imágenes	985
24.3 Operaciones de procesamiento de imagen a bajo nivel	986
Detección de aristas	987
Segmentación de la imagen	990
24.4 Extracción de información tridimensional	991
Movimiento	993
Estereoscopía binocular	996
Gradientes de textura	997
Sombreado	999
Contorno	1000
24.5 Reconocimiento de objetos	1004
Reconocimiento basado en la intensidad	1007
Reconocimiento basado en las características	1008
Estimación de postura	1010

24.6	Empleo de la visión para la manipulación y navegación	1012
24.7	Resumen	1014
	Notas bibliográficas e históricas	1015
	Ejercicios	1018
25	Robótica	1023
25.1	Introducción	1023
25.2	<i>Hardware</i> robótico	1025
	Sensores	1025
	Efectores	1027
25.3	Percepción robótica	1029
	Localización	1031
	Generación de mapas	1036
	Otros tipos de percepción	1039
25.4	Planear el movimiento	1039
	Espacio de configuración	1040
	Métodos de descomposición en celdas	1043
	Métodos de esqueletización	1046
25.5	Planificar movimientos inciertos	1047
	Métodos robustos	1048
25.6	Movimiento	1051
	Dinámica y control	1051
	Control del campo de potencial	1054
	Control reactivo	1055
25.7	Arquitecturas <i>software</i> robóticas	1057
	Arquitectura de subsumpción	1058
	Arquitectura de tres capas	1059
	Lenguajes de programación robóticos	1060
25.8	Dominios de aplicación	1061
25.9	Resumen	1064
	Notas bibliográficas e históricas	1065
	Ejercicios	1069
26	Fundamentos filosóficos	1075
26.1	IA débil: ¿pueden las máquinas actuar con inteligencia?	1075
	El argumento de incapacidad	1077
	La objeción matemática	1078
	El argumento de la informalidad	1079
26.2	IA fuerte: ¿pueden las máquinas pensar de verdad?	1081
	El problema de mente-cuerpo	1084
	El experimento del «cerebro en una cubeta»	1085
	El experimento de la prótesis cerebral	1086
	La habitación china	1088
26.3	La ética y los riesgos de desarrollar la Inteligencia Artificial	1090
26.4	Resumen	1095
	Notas bibliográficas e históricas	1095
	Ejercicios	1098
27	IA: presente y futuro	1099
27.1	Componentes de los agentes	1099
27.2	Arquitecturas de agentes	1102

XVIII CONTENIDO

27.3	¿Estamos llevando la dirección adecuada?	1104
27.4	¿Qué ocurriría si la IA tuviera éxito?	1106
A	Fundamentos matemáticos	1109
A.1	Análisis de la complejidad y la notación O()	1109
	Análisis asintótico	1109
	Los problemas inherentemente difíciles y NP	1110
A.2	Vectores, matrices y álgebra lineal	1112
A.3	Distribuciones de probabilidades	1114
	Notas bibliográficas e históricas	1115
B	Notas sobre lenguajes y algoritmos	1117
B.1	Definición de lenguajes con Backus-Naur Form (BNF)	1117
B.2	Algoritmos de descripción en pseudocódigo	1118
B.3	Ayuda en línea	1119
	Bibliografía	1121
	Índice alfabético	1179

Prólogo

La Inteligencia Artificial (IA) es un campo grande (enorme), y este libro también. He intentado explorarlo con plena profundidad acompañándolo constantemente de lógica, probabilidad y matemáticas; de percepción, razonamiento, aprendizaje y acción, es decir, de todo lo que procede de los dispositivos microelectrónicos hasta los exploradores del planetario de la robótica. Otra razón para que este libro se pueda considerar espléndido es la profundidad en la presentación de los resultados, aunque nos hayamos esforzado por abarcar sólo las ideas más centrales en la parte principal de cada capítulo. En las notas bibliográficas al final de cada capítulo se proporcionan consejos para promover resultados.

El subtítulo de este libro es «Un Enfoque Moderno». La intención de esta frase bastante vacía es el hecho de que hemos intentado sintetizar lo que se conoce ahora dentro de un marco de trabajo común, en vez de intentar explicar cada uno de los subcampos de la IA dentro de su propio contexto histórico. Nos disculpamos ante aquellos cuyos subcampos son, como resultado, menos reconocibles de lo que podrían haber sido de cualquier otra forma.

El principal tema unificador es la idea del **agente inteligente**. Definimos la IA como el estudio de los agentes que reciben percepciones del entorno y llevan a cabo las acciones. Cada agente implementa una función la cual estructura las secuencias de las percepciones en acciones; también tratamos las diferentes formas de representar estas funciones, tales como sistemas de producción, agentes reactivos, planificadores condicionales en tiempo real, redes neurales y sistemas teóricos para las decisiones. Explicaremos el papel del aprendizaje cuando alcanza al diseñador y cómo se introduce en entornos desconocidos, mostrando también cómo ese papel limita el diseño del agente, favoreciendo así la representación y el razonamiento explícitos del conocimiento. Trataremos la robótica y su visión no como problemas con una definición independiente, sino como algo que ocurre para lograr los objetivos. Daremos importancia al entorno de las tareas al determinar el diseño apropiado de los agentes.

Nuestro objetivo principal es el de transmitir las *ideas* que han surgido durante los últimos 50 años de investigación en IA y trabajos afines durante los dos últimos milenios. Hemos intentado evitar una excesiva formalidad en la presentación de estas ideas a la vez que hemos intentado cuidar la precisión. Siempre que es necesario y adecuado, incluimos algoritmos en pseudo código para concretar las ideas, lo que se describe en el Apéndice B. Las implementaciones en varios lenguajes de programación están disponibles en el sitio Web de este libro, en la dirección de Internet aima.cs.berkeley.edu.

Este libro se ha pensado principalmente para utilizarse en un curso de diplomatura o en una serie de varios cursos. También se puede utilizar en un curso de licenciatura (quizás acompañado de algunas de las fuentes primarias, como las que se sugieren en las notas bibliográficas). Debida a la extensa cobertura y a la gran cantidad de algoritmos detallados, también es una herramienta útil como manual de referencia primario para alumnos de licenciatura superior y profesionales que deseen abarcar más allá de su propio subcampo. El único prerequisito es familiarizarse con los conceptos básicos de la ciencia de la informática (algoritmos, estructuras de datos, complejidad) a un nivel de aprendizaje de segundo año. El cálculo de Freshman es útil para entender las redes neuronales y el aprendizaje estadístico en detalle. En el Apéndice A se ofrecen los fundamentos matemáticos necesarios.

Visión general del libro

El libro se divide en ocho partes. La Parte I, **Inteligencia Artificial**, ofrece una visión de la empresa de IA basado en la idea de los agentes inteligentes, sistemas que pueden decidir qué hacer y entonces actuar. La Parte II, **Resolución de Problemas**, se concentra en los métodos para decidir qué hacer cuando se planean varios pasos con antelación, por ejemplo, navegar para cruzar un país o jugar al ajedrez. La Parte III, **Conocimiento y Razonamiento**, abarca las formas de representar el conocimiento sobre el mundo, es decir, cómo funciona, qué aspecto tiene actualmente y cómo podría actuar, y estudia también cómo razonar de forma lógica con ese conocimiento. La Parte IV, **Planificación**, abarca cómo utilizar esos métodos de razonamiento para decidir qué hacer, particularmente construyendo *planes*. La Parte V, **Conocimiento y Razonamiento Inciertos**, se asemeja a las Partes III y IV, pero se concentra en el razonamiento y en la toma de decisiones en presencia de la *incertidumbre* del mundo, la forma en que se podría enfrentar, por ejemplo, a un sistema de tratamiento y diagnosis médicos.

Las Partes II y V describen esa parte del agente inteligente responsable de alcanzar las decisiones. La Parte IV, **Aprendizaje**, describe los métodos para generar el conocimiento requerido por los componentes de la toma de decisiones. La Parte VII, **Comunicación, Percepción y Actuación**, describe las formas en que un agente inteligente puede percibir su entorno para saber lo que está ocurriendo, tanto si es mediante la visión, el tacto, el oído o el entendimiento del idioma; también describe cómo se pueden transformar los planes en acciones reales, o bien en movimientos de un robot, o bien en órdenes del lenguaje natural. Finalmente, la Parte VIII, **Conclusiones**, analiza el pasado y el futuro de la IA y sus repercusiones filosóficas y éticas.

Cambios de la primera edición

Desde que en 1995 se publicó la primera edición, la IA ha sufrido muchos cambios, por lo tanto esta edición también ha sufrido muchos cambios. Se han vuelto a escribir todos los capítulos elocuentemente para reflejar los últimos trabajos de este campo, reinterpretar los trabajos anteriores para que haya mayor coherencia con los actuales y mejorar la dirección pedagógica de las ideas. Los seguidores de la IA deberían estar alentados, ya que las técnicas actuales son mucho más prácticas que las del año 1995; por ejemplo, los algoritmos de planificación de la primera edición podrían generar planes sólo de unos cuantos pasos, mientras que los algoritmos de esta edición superan los miles de pasos. En la deducción probalística se han visto mejoras similares de órdenes de magnitud, procesamiento de lenguajes y otros subcampos. A continuación se muestran los cambios más destacables de esta edición:

- En la Parte I hacemos un reconocimiento de las contribuciones históricas a la teoría de controles, teoría de juegos, economía y neurociencia. Esto ayudará a sintonizar con una cobertura más integrada de estas ideas en los siguientes capítulos.
- En la Parte II se estudian los algoritmos de búsqueda en línea, y se ha añadido un capítulo nuevo sobre la satisfacción de las limitaciones. Este último proporciona una conexión natural con el material sobre la lógica.
- En la Parte III la lógica proposicional, que en la primera edición, se presentó como un peldaño a la lógica de primer orden, ahora se presenta como un lenguaje de representación útil por propio derecho, con rápidos algoritmos de deducción y diseños de agentes basados en circuitos. Los capítulos sobre la lógica de primer orden se han reorganizado para presentar el material de forma más clara, añadiendo el ejemplo del dominio de compras en Internet.
- En la Parte IV incluimos los métodos de planificación más actuales, tales como GRAPHPLAN y la planificación basada en la *satisfabilidad*, e incrementamos la cobertura de la planificación temporal, planificación condicional, planificación jerárquica y planificación multiagente.
- En la Parte V hemos aumentado el material de redes Bayesianas con algoritmos nuevos, tales como la eliminación de variables y el algoritmo Monte Carlo de cadena Markov; también hemos creado un capítulo nuevo sobre el razonamiento temporal incierto, abarcando los modelos ocultos de Markov, los filtros Kalman y las redes dinámicas Bayesianas. Los procesos de decisión de Markov se estudian en profundidad, añadiendo también secciones de teoría de juegos y diseños de mecanismos.
- En la Parte VI combinamos trabajos de aprendizaje estadístico, simbólico y neutral, y añadimos secciones para impulsar los algoritmos, el algoritmo EM, aprendizaje basado en instancias, y métodos kernel «de núcleo» (soporte a máquinas vectoriales).
- En la Parte VII el estudio del procesamiento de lenguajes añade secciones sobre el procesamiento del discurso y la inducción gramatical, así como un capítulo so-

bre los modelos de lenguaje probalístico, con aplicaciones en la recuperación de información y en la traducción automática. El estudio de la robótica enfatiza la integración de datos inciertos de sensores, y el capítulo sobre la visión actualiza el material sobre el reconocimiento de objetos.

- En la Parte VIII introducimos una sección sobre las repercusiones éticas de la IA.

Utilización del libro

27 capítulos componen la totalidad del libro, que a su vez están compuestos por sesiones para clases que pueden durar una semana, por tanto para completar el libro se pueden necesitar hasta dos semestres. Alternativamente, se puede adaptar un curso de forma que se aadecue a los intereses del instructor o del alumno. Si se utiliza el libro completo, servirá como soporte para cursos tanto si son reducidos, de introducción, para diplomaturas o para licenciados e ingenieros, de especialización, en temas avanzados. En la página Web aima.cs.berkeley.edu, se ofrecen programas de muestra recopilados de más de 600 escuelas y facultades, además de sugerencias que ayudarán a encontrar la secuencia más adecuada para sus necesidades.

Se han incluido 385 ejercicios que requieren una buena programación, y se han marcado con el icono de un teclado. Estos ejercicios se pueden resolver utilizando el repositorio de códigos que se encuentra en la página Web, aima.cs.berkeley.edu. Algunos de ellos son tan grandes que se pueden considerar proyectos de fin de trimestre. Algunos ejercicios requieren consultar otros libros de texto para investigar y se han marcado con el icono de un libro.

Los puntos importantes también se han marcado con un ícono que indica puntos importantes. Hemos incluido un índice extenso de 10.000 elementos que facilitan la utilización del libro. Además, siempre que aparece un **TÉRMINO NUEVO**, también se destaca en el margen.



TÉRMINO NUEVO

Utilización de la página Web

En la página Web se puede encontrar:

- implementaciones de los algoritmos del libro en diferentes lenguajes de programación,
- una relación de aproximadamente 600 universidades y centros, que han utilizado este libro, muchos de ellos con enlaces a los materiales de cursos en la red,
- una relación de unos 800 enlaces a sitios Web con contenido útil sobre la IA,
- una lista de todos los capítulos, uno a uno, material y enlaces complementarios,
- instrucciones sobre cómo unirse a un grupo de debate sobre el libro,
- instrucciones sobre cómo contactar con los autores mediante preguntas y comentarios,

- instrucciones sobre cómo informar de los errores del libro, en el caso probable de que existieran, y
- copias de las figuras del libro, junto con diapositivas y otro tipo de material para profesores.

Agradecimientos

Jitendra Malik ha escrito la mayor parte del Capítulo 24 (sobre la visión). Gran parte del Capítulo 25 (sobre robótica) de esta edición ha sido escrita por Sebastián Thrun, mientras que la primera edición fue escrita por John Canny. Doug Edwards investigó las notas históricas de la primera edición. Tim Huang, Mark Paskin y Cinthia Bruyns ayudó a realizar el formato de los diagramas y algoritmos. Alan Apt, Sondra Chavez, Toni Holm, Jake Warde, Irwin Zucker y Camille Trentacoste de Prentice Hall hicieron todo lo que pudieron para cumplir con el tiempo de desarrollo y aportaron muchas sugerencias sobre el diseño y el contenido del libro.

Stuart quiere agradecer a sus padres el continuo apoyo y aliento demostrado, y a su esposa, Loy Sheflott su gran paciencia e infinita sabiduría. También desea que Gordon y Lucy puedan leer esta obra muy pronto. Y a RUGS (Russell's Unusual Group of Students) por haber sido de verdad de gran ayuda.

Peter quiere agradecer a sus padres (Torsten y Gerda) el ánimo que le aportaron para impulsarle a empezar el proyecto, y a su mujer Kris, hijos y amigos por animarle y tolerar todas las horas que ha dedicado en escribir y en volver a escribir su trabajo.

Estamos en deuda con los bibliotecarios de las universidades de Berkeley, Stanford, y con el MIT y la NASA; y con los desarrolladores de CiteSeer y Google por haber revolucionado la forma en que hemos realizado gran parte de la investigación.

Obviamente, nos es imposible agradecer a todas las personas que han utilizado el libro y que han realizado sugerencias, sin embargo nos gustaría mostrar especial agradecimiento a los comentarios que han realizado personas como Eyal Amir, Krzysztof Apt, Ellery Aziel, Jeff Van Baalen, Brian Baker, Don Barker, Tony Barrett, James Newton Bass, Don Beal, Howard Beck, Wolfgang Bibel, John Binder, Larry Bookman, David R. Boxall, Gerhard Brewka, Selmer Bringsjord, Carla Brodley, Chris Brown, Wilhelm Burger, Lauren Burka, Joao Cachopo, Murray Campbell, Norman Carver, Emmanuel Castro, Anil Chakravarthy, Dan Chisarick, Roberto Cipolla, David Cohen, James Coleman, Julie Ann Comparini, Gary Cottrell, Ernest Davis, Rina Dechter, Tom Dietterich, Chuck Dyer, Barbara Engelhardt, Doug Edwards, Kutluhan Erol, Oren Etzioni, Hana Filip, Douglas Fisher, Jeffrey Forbes, Ken Ford, John Fosler, Alex Franz, Bob Futrelle, Marek Galecki, Stefan Gerberding, Stuart Gill, Sabine Glesner, Seth Golub, Gosta Grahe, Russ Greiner, Eric Grimson, Barbara Grosz, Larry Hall, Steve Hanks, Othar Hansson, Ernst Heinz, Jim Hendler, Christoph Herrmann, Vasant Honavar, Tim Huang, Seth Hutchinson, Joost Jacob, Magnus Johansson, Dan Jurafsky, Leslie Kaelbling, Keiji Kanazawa, Surekha Kasibhatla, Simon Kasif, Henry Kautz, Gernot Kerschbaumer, Richard Kirby, Kevin Knight, Sven Koenig, Daphne Koller, Rich Korf, James Kurien, John Lafferty, Gus Larsson, John Lazzaro, Jon LeBlanc, Jason Leatherman, Frank Lee,

Edward Lim, Pierre Louveaux, Don Loveland, Sridhar Mahadevan, Jim Martin, Andy Mayer, David McGrane, Jay Mendelsohn, Brian Milch, Steve Minton, Vibhu Mittal, Leora Morgenstern, Stephen Muggleton, Kevin Murphy, Ron Musick, Sung Myaeng, Lee Naish, Pandu Nayak, Bernhard Nebel, Stuart Nelson, XuanLong Nguyen, Illah Nourbakhsh, Steve Omohundro, David Page, David Palmer, David Parkes, Ron Parr, Mark Paskin, Tony Passera, Michael Pazzani, Wim Pijls, Ira Pohl, Martha Pollack, David Poole, Bruce Porter, Malcolm Pradhan, Bill Pringle, Lorraine Prior, Greg Provan, William Rapaport, Philip Resnik, Francesca Rossi, Jonathan Schaeffer, Richard Scherl, Lars Schuster, Soheil Shams, Stuart Shapiro, Jude Shavlik, Satinder Singh, Daniel Sleator, David Smith, Bryan So, Robert Sproull, Lynn Stein, Larry Stephens, Andreas Stolcke, Paul Stradling, Devika Subramanian, Rich Sutton, Jonathan Tash, Austin Tate, Michael Thielscher, William Thompson, Sebastian Thrun, Eric Tiedemann, Mark Torrance, Randall Upham, Paul Utgoff, Peter van Beek, Hal Varian, Sunil Vemuri, Jim Waldo, Bonnie Webber, Dan Weld, Michael Wellman, Michael Dean White, Kamin Whitehouse, Brian Williams, David Wolfe, Bill Woods, Alden Wright, Richard Yen, Weixiong Zhang, Shlomo Zilberstein, y los revisores anónimos proporcionados por Prentice Hall.

Acerca de la portada

La ilustración de la portada ha sido diseñada por los autores y ejecutada por Lisa Marie Sardegna y Maryann Simmons utilizando SGI Inventor™ y Photshop™ de Adobe, y representa los siguientes elementos de la historia de la IA:

1. El algoritmo de planificación del *De Motu Animalium*, Aristóteles (400 a.C.).
2. El generador de conceptos del *Ars Magna* de Ramón Lull (1300 d.C.).
3. El motor de diferencias de Charles Babbage, un prototipo del primer computador universal (1848).
4. La notación de lógica de primer orden de Gottlob Frege (1789).
5. Los diagramas del razonamiento lógico de Lewis Carroll (1886).
6. La notación de redes probalísticas de Sewall Wright (1921).
7. Alan Turing (1912-1954).
8. El Robot Shakey (1969-1973).
9. Un sistema experto de diagnóstico actual (1993).

Sobre los autores

Stuart Russell nació en 1962 en Portsmouth, Inglaterra. En 1982, obtuvo su B.A. en Física formando parte de los primeros en el cuadro de honor de la Universidad de Oxford. En 1986 obtuvo el Ph. D. en Informática por la Universidad de Stanford. Más tarde empezó a trabajar en la Universidad de California, Berkeley, en donde es profesor de Informática, director del centro de Sistemas Inteligentes y colaborador del Smith-Zadeh en Ingeniería. En 1990, recibió el premio Presidential Young Investigator Award de la National Science Foundation (Fundación Nacional de las Ciencias), y en 1995 obtuvo el premio compartido en el Computers y Thought Award. En 1996 se convirtió en Millar Profesor de la University of California, y en el año 2000 fue nombrado rector de la universidad. En 1998, dio la lectura inaugural de la Stanford University, Forsythe Memorial Lectures. Forma parte como «Fellow», y es uno de los primeros miembros del Executive Council de la American Association for Artificial Intelligence. Ha publicado un gran número de trabajos, más de 100 sobre una gran variedad de temas en inteligencia artificial. Y es autor de otros dos libros: *The Use of Knowledge in Analogy and Induction*, y (con Erik Wefald) *Do the Right Thing: Studies in Limited Rationality*.

Peter Norvig es director de Search Quality de Google, Inc. y forma parte, como «Fellow» y miembro del Executive Council en la American Association for Artificial Intelligence. Anteriormente, fue director de la División de Ciencias Computacionales del Ames Research Center de la NASA, en donde supervisaba la investigación y el desarrollo en inteligencia artificial. Antes de eso, trabajó como director de ciencia en Junglee, donde ayudó a desarrollar uno de los primeros servicios de extracción de información en Internet, y trabajó también como científico senior en Sun Microsystems Laboratories, cuyo trabajo consistía en recuperar información inteligente. Recibió un B.S. en Matemáticas Aplicadas por la Brown University, y un Ph. D. en informática por la Universidad de California en Berkeley. Es profesor de la University of Southern California, y miembro de la facultad de investigación en Berkeley. Tiene en su haber más de 50 publicaciones en Informática entre las que se incluyen los libros: *Paradigms of AI Programming: Case Studies in Common Lisp*, *Verbmobil: A Translation System for Face-to-Face Dialog*, e *Intelligent Help Systems for UNIX*.



Introducción

Donde se intentará explicar por qué se considera a la inteligencia artificial un tema digno de estudio y donde se intentará definirla con exactitud; es esta tarea muy recomendable antes de emprender de lleno su estudio.

INTELIGENCIA
ARTIFICIAL

Los hombres se han denominado a sí mismos como *Homo sapiens* (hombre sabio) porque nuestras capacidades mentales son muy importantes para nosotros. Durante miles de años, hemos tratado de entender *cómo pensamos*; es decir, entender cómo un simple puñado de materia puede percibir, entender, predecir y manipular un mundo mucho más grande y complicado que ella misma. El campo de la **inteligencia artificial**, o IA, va más allá: no sólo intenta comprender, sino que también se esfuerza en construir entidades inteligentes.

La IA es una de las ciencias más recientes. El trabajo comenzó poco después de la Segunda Guerra Mundial, y el nombre se acuñó en 1956. La IA se cita, junto a la biología molecular, como un campo en el que a la mayoría de científicos de otras disciplinas «les gustaría trabajar». Un estudiante de ciencias físicas puede pensar razonablemente que todas las buenas ideas han sido ya propuestas por Galileo, Newton, Einstein y otros. Por el contrario, la IA aún tiene flecos sin cerrar en los que podrían trabajar varios Einstein a tiempo completo.

La IA abarca en la actualidad una gran variedad de subcampos, que van desde áreas de propósito general, como el aprendizaje y la percepción, a otras más específicas como el ajedrez, la demostración de teoremas matemáticos, la escritura de poesía y el diagnóstico de enfermedades. La IA sintetiza y automatiza tareas intelectuales y es, por lo tanto, potencialmente relevante para cualquier ámbito de la actividad intelectual humana. En este sentido, es un campo genuinamente universal.

1.1 ¿Qué es la IA?

RACIONALIDAD

Hemos proclamado que la IA es excitante, pero no hemos dicho qué *es*. La Figura 1.1 presenta definiciones de inteligencia artificial extraídas de ocho libros de texto. Las que aparecen en la parte superior se refieren a *procesos mentales* y al *razonamiento*, mientras que las de la parte inferior aluden a la *conducta*. Las definiciones de la izquierda miden el éxito en términos de la fidelidad en la forma de actuar de los *humanos*, mientras que las de la derecha toman como referencia un concepto ideal de inteligencia, que llamaremos **racionalidad**. Un sistema es racional si hace «lo correcto», en función de su conocimiento.

A lo largo de la historia se han seguido los cuatro enfoques mencionados. Como es de esperar, existe un enfrentamiento entre los enfoques centrados en los humanos y los centrados en torno a la *racionalidad*¹. El enfoque centrado en el comportamiento humano debe ser una ciencia empírica, que incluya hipótesis y confirmaciones mediante experimentos. El enfoque racional implica una combinación de matemáticas e ingeniería. Cada grupo al mismo tiempo ha ignorado y ha ayudado al otro. A continuación revisaremos cada uno de los cuatro enfoques con más detalle.

Sistemas que piensan como humanos	Sistemas que piensan racionalmente
«El nuevo y excitante esfuerzo de hacer que los computadores piensen... máquinas con mentes, en el más amplio sentido literal». (Haugeland, 1985)	«El estudio de las facultades mentales mediante el uso de modelos computacionales». (Charniak y McDermott, 1985)
«[La automatización de] actividades que vinculamos con procesos de pensamiento humano, actividades como la toma de decisiones, resolución de problemas, aprendizaje....» (Bellman, 1978)	«El estudio de los cálculos que hacen posible percibir, razonar y actuar». (Winston, 1992)
Sistemas que actúan como humanos	Sistemas que actúan racionalmente
«El arte de desarrollar máquinas con capacidad para realizar funciones que cuando son realizadas por personas requieren de inteligencia». (Kurzweil, 1990)	«La Inteligencia Computacional es el estudio del diseño de agentes inteligentes». (Poole <i>et al.</i> , 1998)
«El estudio de cómo lograr que los computadores realicen tareas que, por el momento, los humanos hacen mejor». (Rich y Knight, 1991)	«IA... está relacionada con conductas inteligentes en artefactos». (Nilsson, 1998)
Figura 1.1 Algunas definiciones de inteligencia artificial, organizadas en cuatro categorías.	

¹ Conviene aclarar, que al distinguir entre comportamiento *humano* y *racional* no se está sugiriendo que los humanos son necesariamente «irracionales» en el sentido de «inestabilidad emocional» o «desequilibrio mental». Basta con darnos cuenta de que no somos perfectos: no todos somos maestros de ajedrez, incluso aquellos que conocemos todas las reglas del ajedrez; y desafortunadamente, no todos obtenemos un sobresaliente en un examen. Kahneman *et al.* (1982) ha elaborado un catálogo con algunos de los errores que sistemáticamente cometan los humanos cuando razonan.

Comportamiento humano: el enfoque de la Prueba de Turing

PRUEBA DE TURING

La **Prueba de Turing**, propuesta por Alan Turing (1950), se diseñó para proporcionar una definición operacional y satisfactoria de inteligencia. En vez de proporcionar una lista larga y quizás controvertida de cualidades necesarias para obtener inteligencia artificialmente, él sugirió una prueba basada en la incapacidad de diferenciar entre entidades inteligentes indiscutibles y seres humanos. El computador supera la prueba si un evaluador humano no es capaz de distinguir si las respuestas, a una serie de preguntas planteadas, son de una persona o no. En el Capítulo 26 se comentan detalles de esta prueba y se discute si un computador que supera la prueba es realmente inteligente. Hoy por hoy, podemos decir que programar un computador para que supere la prueba requiere un trabajo considerable. El computador debería poseer las siguientes capacidades:

PROCESAMIENTO DE LENGUAJE NATURAL

REPRESENTACIÓN DEL CONOCIMIENTO

RAZONAMIENTO AUTOMÁTICO

APRENDIZAJE MÁQUINA

PRUEBA DE TURING GLOBAL

VISTA COMPUTACIONAL

ROBÓTICA

- **Procesamiento de lenguaje natural** que le permita comunicarse satisfactoriamente en inglés.
- **Representación del conocimiento** para almacenar lo que se conoce o siente.
- **Razonamiento automático** para utilizar la información almacenada para responder a preguntas y extraer nuevas conclusiones.
- **Aprendizaje automático** para adaptarse a nuevas circunstancias y para detectar y extraer patrones.

La Prueba de Turing evitó deliberadamente la interacción *física* directa entre el evaluador y el computador, dado que para medir la inteligencia es innecesario simular físicamente a una persona. Sin embargo, la llamada Prueba Global de Turing incluye una señal de vídeo que permite al evaluador valorar la capacidad de percepción del evaluado, y también le da la oportunidad al evaluador de pasar objetos físicos «a través de una ventanita». Para superar la Prueba Global de Turing el computador debe estar dotado de

- **Visión computacional** para percibir objetos.
- **Robótica** para manipular y mover objetos.

Estas seis disciplinas abarcan la mayor parte de la IA, y Turing merece ser reconocido por diseñar una prueba que se conserva vigente después de 50 años. Los investigadores del campo de la IA han dedicado poco esfuerzo a la evaluación de sus sistemas con la Prueba de Turing, por creer que es más importante el estudio de los principios en los que se basa la inteligencia que duplicar un ejemplar. La búsqueda de un ingenio que «volara artificialmente» tuvo éxito cuando los hermanos Wright, entre otros, dejaron de imitar a los pájaros y comprendieron los principios de la aerodinámica. Los textos de ingeniería aerodinámica no definen el objetivo de su campo como la construcción de «máquinas que vuelen como palomas de forma que puedan incluso confundir a otras palomas».

Pensar como un humano: el enfoque del modelo cognitivo

Para poder decir que un programa dado piensa como un humano, es necesario contar con un mecanismo para determinar cómo piensan los humanos. Es necesario *penetrar* en el

funcionamiento de las mentes humanas. Hay dos formas de hacerlo: mediante introspección (intentando atrapar nuestros propios pensamientos conforme éstos van apareciendo) y mediante experimentos psicológicos. Una vez se cuente con una teoría lo suficientemente precisa sobre cómo trabaja la mente, se podrá expresar esa teoría en la forma de un programa de computador. Si los datos de entrada/salida del programa y los tiempos de reacción son similares a los de un humano, existe la evidencia de que algunos de los mecanismos del programa se pueden comparar con los que utilizan los seres humanos. Por ejemplo, a Allen Newell y Herbert Simon, que desarrollaron el «Sistema de Resolución General de Problemas» (SRGP) (Newell y Simon, 1961), no les bastó con que su programa resolviera correctamente los problemas propuestos. Lo que les interesaba era seguir la pista de las etapas del proceso de razonamiento y compararlas con las seguidas por humanos a los que se les enfrentó a los mismos problemas. En el campo interdisciplinario de la **ciencia cognitiva** convergen modelos computacionales de IA y técnicas experimentales de psicología intentando elaborar teorías precisas y verificables sobre el funcionamiento de la mente humana.

La ciencia cognitiva es un campo fascinante, merecedora de una enciclopedia dedicada a ella (Wilson y Keil, 1999). En este libro no se intenta describir qué se conoce de la cognición humana. Ocasionalmente se hacen comentarios acerca de similitudes o diferencias entre técnicas de IA y cognición humana. La auténtica ciencia cognitiva se fundamenta necesariamente en la investigación experimental en humanos y animales, y en esta obra se asume que el lector sólo tiene acceso a un computador para experimentar.

En los comienzos de la IA había confusión entre las distintas aproximaciones: un autor podría argumentar que un algoritmo resolvía adecuadamente una tarea y que *por tanto* era un buen modelo de representación humana, o viceversa. Los autores actuales hacen diferencia entre las dos reivindicaciones; esta distinción ha permitido que ambas disciplinas, IA y ciencia cognitiva, se desarrollen más rápidamente. Los dos campos continúan alimentándose entre sí, especialmente en las áreas de la visión y el lenguaje natural. En particular, el campo de la visión ha avanzado recientemente con la ayuda de una propuesta integrada que tiene en cuenta la evidencia neurofisiológica y los modelos computacionales.

Pensamiento racional: el enfoque de las «leyes del pensamiento»

El filósofo griego Aristóteles fue uno de los primeros en intentar codificar la «manera correcta de pensar», es decir, un proceso de razonamiento irrefutable. Sus **silogismos** son esquemas de estructuras de argumentación mediante las que siempre se llega a conclusiones correctas si se parte de premisas correctas (por ejemplo: «Sócrates es un hombre; todos los hombres son mortales; por lo tanto Sócrates es mortal»). Estas leyes de pensamiento supuestamente gobiernan la manera de operar de la mente; su estudio fue el inicio de un campo llamado **lógica**.

Estudiosos de la lógica desarrollaron, en el siglo XIX, una notación precisa para definir sentencias sobre todo tipo de elementos del mundo y especificar relaciones entre

LOGISTA

ellos (compárese esto con la notación aritmética común, que prácticamente sólo sirve para representar afirmaciones acerca de la igualdad y desigualdad entre números). Ya en 1965 existían programas que, en principio, resolvían *cualquier* problema resoluble descrito en notación lógica². La llamada tradición **logista** dentro del campo de la inteligencia artificial trata de construir sistemas inteligentes a partir de estos programas.

Este enfoque presenta dos obstáculos. No es fácil transformar conocimiento informal y expresarlo en los términos formales que requieren de notación lógica, particularmente cuando el conocimiento que se tiene es inferior al 100 por 100. En segundo lugar, hay una gran diferencia entre poder resolver un problema «en principio» y hacerlo en la práctica. Incluso problemas con apenas una docena de datos pueden agotar los recursos computacionales de cualquier computador a menos que cuente con alguna directiva sobre los pasos de razonamiento que hay que llevar a cabo primero. Aunque los dos obstáculos anteriores están presentes en *todo* intento de construir sistemas de razonamiento computacional, surgieron por primera vez en la tradición lógica.

Actuar de forma racional: el enfoque del agente racional

AGENTE

AGENTE RACIONAL

Un **agente** es algo que razona (*agente* viene del latín *agere*, hacer). Pero de los agentes informáticos se espera que tengan otros atributos que los distingan de los «programas» convencionales, como que estén dotados de controles autónomos, que perciban su entorno, que persistan durante un período de tiempo prolongado, que se adapten a los cambios, y que sean capaces de alcanzar objetivos diferentes. Un **agente racional** es aquel que actúa con la intención de alcanzar el mejor resultado o, cuando hay incertidumbre, el mejor resultado esperado.

En el caso del enfoque de la IA según las «leyes del pensamiento», todo el énfasis se pone en hacer inferencias correctas. La obtención de estas inferencias correctas puede, a veces, formar *parte* de lo que se considera un agente racional, ya que una manera racional de actuar es llegar a la conclusión lógica de que si una acción dada permite alcanzar un objetivo, hay que llevar a cabo dicha acción. Sin embargo, el efectuar una inferencia correcta no depende siempre de la *racionalidad*, ya que existen situaciones para las que no hay nada correcto que hacer y en las que hay que tomar una decisión. Existen también formas de actuar racionalmente que no implican realizar inferencias. Por ejemplo, el retirar la mano de una estufa caliente es un acto reflejo mucho más eficiente que una respuesta lenta llevada a cabo tras una deliberación cuidadosa.

Todas las habilidades que se necesitan en la Prueba de Turing deben permitir emprender acciones racionales. Por lo tanto, es necesario contar con la capacidad para representar el conocimiento y razonar basándonos en él, porque ello permitirá alcanzar decisiones correctas en una amplia gama de situaciones. Es necesario ser capaz de generar sentencias comprensibles en lenguaje natural, ya que el enunciado de tales oraciones permite a los agentes desenvolverse en una sociedad compleja. El aprendizaje no se lleva a cabo por erudición exclusivamente, sino que profundizar en el conocimiento de cómo funciona el mundo facilita la concepción de estrategias mejores para manejarse en él.

² Si no se encuentra una solución, el programa nunca debe parar de buscarla.

La percepción visual es necesaria no sólo porque ver es divertido, sino porque es necesaria para poder tener una idea mejor de lo que una acción puede llegar a representar, por ejemplo, el ver un delicioso bocadillo contribuirá a que nos acerquemos a él.

Por esta razón, el estudiar la IA desde el enfoque del diseño de un agente racional ofrece al menos dos ventajas. La primera es más general que el enfoque que proporcionan las «leyes del pensamiento», dado que el efectuar inferencias correctas es sólo uno de los mecanismos existentes para garantizar la racionalidad. La segunda es más afín a la forma en la que se ha producido el avance científico que los enfoques basados en la conducta o pensamiento humano, porque la norma de la racionalidad está claramente definida y es de aplicación general. Por el contrario, la conducta humana se adapta bien a un entorno específico, y en parte, es producto de un proceso evolutivo complejo, en gran medida desconocido, que aún está lejos de llevarnos a la perfección. *Por tanto, esta obra se centrará en los principios generales que rigen a los agentes racionales y en los elementos necesarios para construirlos.* Más adelante quedará patente que a pesar de la aparente facilidad con la que se puede describir un problema, cuando se intenta resolver surgen una enorme variedad de cuestiones. El Capítulo 2 revisa algunos de estos aspectos con más detalle.



Un elemento importante a tener en cuenta es el siguiente: más bien pronto que tarde se observará cómo obtener una racionalidad perfecta (hacer siempre lo correcto) no es posible en entornos complejos. La demanda computacional que esto implica es demasiado grande. En la mayor parte de esta obra se adoptará la hipótesis de trabajo de que la racionalidad perfecta es un buen punto de partida para el análisis. Lo cual simplifica el problema y proporciona el escenario base adecuado sobre el que se asientan los cimientos de este campo. Los Capítulos 6 y 17 se centran explícitamente en el tema de la **racionalidad limitada** (actuar adecuadamente cuando no se cuenta con el tiempo suficiente para efectuar todos los cálculos que serían deseables).

RACIONALIDAD
LIMITADA

1.2 Los fundamentos de la inteligencia artificial

Esta sección presenta una breve historia de las disciplinas que han contribuido con ideas, puntos de vista y técnicas al desarrollo de la IA. Como toda revisión histórica, en este caso se centra en un pequeño número de personas, eventos e ideas e ignora otras que también fueron importantes. La historia se organiza en torno a una serie de cuestiones, dejando claro que no se quiere dar la impresión de que estas cuestiones son las únicas por las que las disciplinas se han preocupado y que el objetivo último de todas estas disciplinas era hacer avanzar la IA.

Filosofía (desde el año 428 a.C. hasta el presente)

- ¿Se pueden utilizar reglas formales para extraer conclusiones válidas?
- ¿Cómo se genera la inteligencia mental a partir de un cerebro físico?
- ¿De dónde viene el conocimiento?
- ¿Cómo se pasa del conocimiento a la acción?

Aristóteles (384-322 a.C.) fue el primero en formular un conjunto preciso de leyes que gobernaban la parte racional de la inteligencia. Él desarrolló un sistema informal para razonar adecuadamente con silogismos, que en principio permitía extraer conclusiones mecánicamente, a partir de premisas iniciales. Mucho después, Ramón Lull (d. 1315) tuvo la idea de que el razonamiento útil se podría obtener por medios artificiales. Sus «ideas» aparecen representadas en la portada de este manuscrito. Thomas Hobbes (1588-1679) propuso que el razonamiento era como la computación numérica, de forma que «nosotros sumamos y restamos silenciosamente en nuestros pensamientos». La automatización de la computación en sí misma estaba en marcha; alrededor de 1500, Leonardo da Vinci (1452-1519) diseñó, aunque no construyó, una calculadora mecánica; construcciones recientes han mostrado que su diseño era funcional. La primera máquina calculadora conocida se construyó alrededor de 1623 por el científico alemán Wilhelm Schickard (1592-1635), aunque la Pascalina, construida en 1642 por Blaise Pascal (1623-1662), sea más famosa. Pascal escribió que «la máquina aritmética produce efectos que parecen más similares a los pensamientos que a las acciones animales». Gottfried Wilhelm Leibniz (1646-1716) construyó un dispositivo mecánico con el objetivo de llevar a cabo operaciones sobre conceptos en lugar de sobre números, pero su campo de acción era muy limitado.

Ahora que sabemos que un conjunto de reglas pueden describir la parte racional y formal de la mente, el siguiente paso es considerar la mente como un sistema físico. René Descartes (1596-1650) proporciona la primera discusión clara sobre la distinción entre la mente y la materia y los problemas que surgen. Uno de los problemas de una concepción puramente física de la mente es que parece dejar poco margen de maniobra al libre albedrío: si el pensamiento está totalmente gobernado por leyes físicas, entonces una piedra podría «decidir» caer en dirección al centro de la Tierra gracias a su libre albedrío. A pesar de ser denodado defensor de la capacidad de razonamiento, Descartes fue un defensor del **dualismo**. Sostenía que existe una parte de la mente (o del alma o del espíritu) que está al margen de la naturaleza, exenta de la influencia de las leyes físicas. Los animales, por el contrario, no poseen esta cualidad dual; a ellos se le podría concebir como si se tratases de máquinas. Una alternativa al dualismo es el **materiaлизmo**, que considera que las operaciones del cerebro realizadas de acuerdo a las leyes de la física *constituyen* la mente. El libre albedrío es simplemente la forma en la que la percepción de las opciones disponibles aparecen en el proceso de selección.

Dada una mente física que gestiona conocimiento, el siguiente problema es establecer las fuentes de este conocimiento. El movimiento **empírico**, iniciado con el *Novum Organum*³, de Francis Bacon (1561-1626), se caracteriza por el aforismo de John Locke (1632-1704): «Nada existe en la mente que no haya pasado antes por los sentidos». David Hume (1711-1776) propuso en *A Treatise of Human Nature* (Hume, 1739) lo que actualmente se conoce como principio de **inducción**: las reglas generales se obtienen mediante la exposición a asociaciones repetidas entre sus elementos. Sobre la base de las propuestas de Ludwig Wittgenstein (1889-1951) y Bertrand Russell (1872-1970), el famoso Círculo de Viena, liderado por Rudolf Carnap (1891-1970), desarrolló la doctrina del **positivismo lógico**. Esa doctrina sostiene que todo el conocimiento se puede

DUALISMO

MATERIALISMO

EMPÍRICO

INDUCCIÓN

³ Una actualización del *Organon*, o instrumento de pensamiento, de Aristóteles.



caracterizar mediante teorías lógicas relacionadas, en última instancia, con **sentencias de observación** que corresponden a estímulos sensoriales⁴. La **teoría de la confirmación** de Carnap y Carl Hempel (1905-1997) intenta explicar cómo el conocimiento se obtiene a partir de la experiencia. El libro de Carnap, *The Logical Structure of the World* (1928), define un procedimiento computacional explícito para la extracción de conocimiento a partir de experiencias primarias. Fue posiblemente la primera teoría en mostrar la mente como un proceso computacional.

El último elemento en esta discusión filosófica sobre la mente es la relación que existe entre conocimiento y acción. Este asunto es vital para la IA, ya que la inteligencia requiere tanto acción como razonamiento. Más aún, simplemente con comprender cómo se justifican determinadas acciones se puede llegar a saber cómo construir un agente cuyas acciones sean justificables (o racionales). Aristóteles argumenta que las acciones se pueden justificar por la conexión lógica entre los objetivos y el conocimiento de los efectos de las acciones (la última parte de este extracto también aparece en la portada de este libro):

¿Cómo es que el pensamiento viene acompañado en algunos casos de acciones y en otros no?, ¿en algunos casos por movimiento y en otros no? Parece como si la misma cosa sucediera tanto si razonáramos o hicieramos inferencias sobre objetos que no cambian; pero en este caso el fin es una proposición especulativa... mientras la conclusión resultante de las dos premisas es una acción... Yo necesito abrigarme; una manta abriga. Yo necesito una manta. Qué necesito, qué debo hacer; necesito una manta. Necesito hacer una manta. Y la conclusión, «Yo tengo que hacer una manta», es una acción. (Nussbaum, 1978, p. 40)

En *Nicomachean Ethics* (Libro III. 3, 1112b), Aristóteles continúa trabajando en este tema, sugiriendo un algoritmo:

Nosotros no reflexionamos sobre los fines, sino sobre los medios. Un médico no reflexiona sobre si debe curar, ni un orador sobre si debe persuadir... Ellos asumen el fin y consideran cómo y con qué medios se obtienen, y si resulta fácil y es por tanto productivo; mientras que si sólo se puede alcanzar por un medio se tiene en consideración *cómo* se alcanzará por este y por qué medios se obtendrá *éste*, hasta que se llegue a la causa primera..., y lo último en el orden del análisis parece ser lo primero en el orden de los acontecimientos. Y si se llega a un estado imposible, se abandona la búsqueda, como por ejemplo si se necesita dinero y no se puede conseguir; pero si hay una posibilidad se intentará.

El algoritmo de Aristóteles se implementó 2.300 años más tarde por Newell y Simon con la ayuda de su programa SRGP. El cual se conoce como sistema de planificación regresivo (véase el Capítulo 11).

El análisis basado en objetivos es útil, pero no indica qué hacer cuando varias acciones nos llevan a la consecución del objetivo, o cuando ninguna acción facilita su completa consecución. Antoine Arnauld (1612-1694) describió correctamente una forma cuantitativa para decidir qué acción llevar a cabo en un caso como este (véase el Capítulo 16). El libro *Utilitarianism* (Mill, 1863) de John Stuart Mill (1806-1873) propone

⁴ En este contexto, es posible comprobar o rechazar toda aseveración significativa mediante el análisis del significado de las palabras o mediante la experimentación. Dado que esto no es aplicable en la mayor parte del ámbito de la metafísica, como era intención, el positivismo lógico se hizo impopular en algunos círculos.

la idea de un criterio de decisión racional en todos los ámbitos de la actividad humana. En la siguiente sección se explica una teoría de la decisión más formalmente.

Matemáticas (aproximadamente desde el año 800 al presente)

- ¿Qué reglas formales son las adecuadas para obtener conclusiones válidas?
- ¿Qué se puede computar?
- ¿Cómo razonamos con información incierta?

Los filósofos delimitaron las ideas más importantes de la IA, pero para pasar de ahí a una ciencia formal es necesario contar con una formulación matemática en tres áreas fundamentales: lógica, computación y probabilidad.

El concepto de lógica formal se remonta a los filósofos de la antigua Grecia (véase el Capítulo 7), pero su desarrollo matemático comenzó realmente con el trabajo de George Boole (1815-1864) que definió la lógica proposicional o Booleana (Boole, 1847). En 1879, Gottlob Frege (1848-1925) extendió la lógica de Boole para incluir objetos y relaciones, y creó la lógica de primer orden que se utiliza hoy como el sistema más básico de representación de conocimiento⁵. Alfred Tarski (1902-1983) introdujo una teoría de referencia que enseña cómo relacionar objetos de una lógica con objetos del mundo real. El paso siguiente consistió en definir los límites de lo que se podía hacer con la lógica y la informática.

ALGORITMO

Se piensa que el primer **algoritmo** no trivial es el algoritmo Euclídeo para el cálculo del máximo común divisor. El considerar los algoritmos como objetos en sí mismos se remonta a la época de al-Khowarazmi, un matemático persa del siglo IX, con cuyos escritos también se introdujeron los números arábigos y el álgebra en Europa. Boole, entre otros, presentó algoritmos para llevar a cabo deducciones lógicas y hacia el final del siglo XIX se llevaron a cabo numerosos esfuerzos para formalizar el razonamiento matemático general con la lógica deductiva. En 1900, David Hilbert (1862-1943) presentó una lista de 23 problemas que acertadamente predijo ocuparían a los matemáticos durante todo ese siglo. En el último de ellos se preguntaba si existe un algoritmo que permita determinar la validez de cualquier proposición lógica en la que aparezcan números naturales (el famoso *Entscheidungsproblem*, o problema de decisión). Básicamente, lo que Hilbert se preguntaba es si hay límites fundamentales en la capacidad de los procedimientos efectivos de demostración. En 1930, Kurt Gödel (1906-1978) demostró que existe un procedimiento eficiente para demostrar cualquier aseveración verdadera en la lógica de primer orden de Frege y Russell, sin embargo con la lógica de primer orden no era posible capturar el principio de inducción matemática necesario para la caracterización de los números naturales. En 1931, demostró que, en efecto, existen límites reales. Mediante su **teorema de incompletitud** demostró que en cualquier lenguaje que tuviera la capacidad suficiente para expresar las propiedades de los números naturales, existen aseveraciones verdaderas no decidible en el sentido de que no es posible decidir su validez mediante ningún algoritmo.

TEOREMA DE INCOMPLETITUD

⁵ La notación para la lógica de primer orden propuesta por Frege no se ha aceptado universalmente, por razones que son aparentemente obvias cuando se observa el ejemplo que aparece en la cubierta de este libro.

El resultado fundamental anterior se puede interpretar también como la indicación de que existen algunas funciones de los números enteros que no se pueden representar mediante un algoritmo, es decir no se pueden calcular. Lo anterior llevó a Alan Turing (1912-1954) a tratar de caracterizar exactamente aquellas funciones que sí *eran* susceptibles de ser caracterizadas. La noción anterior es de hecho problemática hasta cierto punto, porque no es posible dar una definición formal a la noción de cálculo o procedimiento efectivo. No obstante, la tesis de Church-Turing, que afirma que la máquina de Turing (Turing, 1936) es capaz de calcular cualquier función computable, goza de aceptación generalizada ya que proporciona una definición suficiente. Turing también demostró que existen algunas funciones que no se pueden calcular mediante la máquina de Turing. Por ejemplo, ninguna máquina puede decidir *en general* si un programa dado producirá una respuesta a partir de unas entradas, o si seguirá calculando indefinidamente.

Si bien ser no decidable ni computable son importantes para comprender el proceso del cálculo, la noción de **intratabilidad** tuvo repercusiones más importantes. En términos generales se dice que un problema es intratable si el tiempo necesario para la resolución de casos particulares de dicho problema crece exponencialmente con el tamaño de dichos casos. La diferencia entre crecimiento polinomial y exponencial de la complejidad se destacó por primera vez a mediados de los años 60 (Cobham, 1964; Edmonds, 1965). Es importante porque un crecimiento exponencial implica la imposibilidad de resolver casos moderadamente grandes en un tiempo razonable. Por tanto, se debe optar por dividir el problema de la generación de una conducta inteligente en subproblemas que sean tratables en vez de manejar problemas intratables.

¿Cómo se puede reconocer un problema intratable? La teoría de la **NP-completitud**, propuesta por primera vez por Steven Cook (1971) y Richard Karp (1972) propone un método. Cook y Karp demostraron la existencia de grandes clases de problemas de razonamiento y búsqueda combinatoria canónica que son NP completos. Toda clase de problema a la que la clase de problemas NP completos se pueda reducir será seguramente intratable (aunque no se ha demostrado que los problemas NP completos son necesariamente intratables, la mayor parte de los teóricos así lo creen). Estos resultados contrastan con el optimismo con el que la prensa popular recibió a los primeros computadores, «Supercerebros Electrónicos» que eran «¡Más rápidos que Einstein!». A pesar del rápido incremento en la velocidad de los computadores, los sistemas inteligentes se caracterizarán por el uso cuidadoso que hacen de los recursos. De manera sucinta, ¡el mundo es un ejemplo de problema *extremadamente* grande! Recientemente la IA ha ayudado a explicar por qué algunos ejemplos de problemas NP completos son difíciles de resolver y otros son fáciles (Cheeseman *et al.*, 1991).

Además de la lógica y el cálculo, la tercera gran contribución de las matemáticas a la IA es la teoría de la **probabilidad**. El italiano Gerolamo Cardano (1501-1576) fue el primero en proponer la idea de probabilidad, presentándola en términos de los resultados de juegos de apuesta. La probabilidad se convirtió pronto en parte imprescindible de las ciencias cuantitativas, ayudando en el tratamiento de mediciones con incertidumbre y de teorías incompletas. Pierre Fermat (1601-1665), Blaise Pascal (1623-1662), James Bernoulli (1654-1705), Pierre Laplace (1749-1827), entre otros, hicieron avanzar esta teoría e introdujeron nuevos métodos estadísticos. Thomas Bayes (1702-1761) propuso una

INTRATABILIDAD**NP-COMPLETITUD****PROBABILIDAD**

regla para la actualización de probabilidades subjetivas a la luz de nuevas evidencias. La regla de Bayes y el área resultante llamado análisis Bayesiano conforman la base de las propuestas más modernas que abordan el razonamiento incierto en sistemas de IA.

Economía (desde el año 1776 hasta el presente)

- ¿Cómo se debe llevar a cabo el proceso de toma de decisiones para maximizar el rendimiento?
- ¿Cómo se deben llevar a cabo acciones cuando otros no colaboren?
- ¿Cómo se deben llevar a cabo acciones cuando los resultados se obtienen en un futuro lejano?

La ciencia de la economía comenzó en 1776, cuando el filósofo escocés Adam Smith (1723-1790) publicó *An Inquiry into the Nature and Causes of the Wealth of Nations*. Aunque los antiguos griegos, entre otros, habían hecho contribuciones al pensamiento económico, Smith fue el primero en tratarlo como una ciencia, utilizando la idea de que las economías pueden concebirse como un conjunto de agentes individuales que intentan maximizar su propio estado de bienestar económico. La mayor parte de la gente cree que la economía sólo se trata de dinero, pero los economistas dicen que ellos realmente estudian cómo la gente toma decisiones que les llevan a obtener los beneficios esperados. Léon Walras (1834-1910) formalizó el tratamiento matemático del «beneficio deseado» o **utilidad**, y fue posteriormente mejorado por Frank Ramsey (1931) y después por John von Neumann y Oskar Morgenstern en su libro *The Theory of Games and Economic Behavior* (1944).

TEORÍA DE LA DECISIÓN

La **teoría de la decisión**, que combina la teoría de la probabilidad con la teoría de la utilidad, proporciona un marco completo y formal para la toma de decisiones (económicas o de otra índole) realizadas bajo incertidumbre, esto es, en casos en los que las descripciones probabilísticas capturan adecuadamente la forma en la que se toman las decisiones en el entorno; lo cual es adecuado para «grandes» economías en las que cada agente no necesita prestar atención a las acciones que lleven a cabo el resto de los agentes individualmente. Cuando se trata de «pequeñas» economías, la situación se asemeja más a la de un **juego**: las acciones de un jugador pueden afectar significativamente a la utilidad de otro (tanto positiva como negativamente). Los desarrollos de von Neumann y Morgenstern a partir de la **teoría de juegos** (véase también Luce y Raiffa, 1957) mostraban el hecho sorprendente de que, en algunos juegos, un agente racional debía actuar de forma aleatoria o, al menos, aleatoria en apariencia con respecto a sus contrincantes.

TEORÍA DE JUEGOS

La gran mayoría de los economistas no se preocuparon de la tercera cuestión mencionada anteriormente, es decir, cómo tomar decisiones racionales cuando los resultados de las acciones no son inmediatos y por el contrario se obtienen los resultados de las acciones de forma *secuencial*. El campo de la **investigación operativa** persigue este objetivo; dicho campo emergió en la Segunda Guerra Mundial con los esfuerzos llevados a cabo en el Reino Unido en la optimización de instalaciones de radar, y posteriormente en aplicaciones civiles relacionadas con la toma de decisiones de dirección complejas. El trabajo de Richard Bellman (1957) formaliza una clase de problemas de decisión secuencial llamados **procesos de decisión de Markov**, que se estudiarán en los Capítulos 17 y 21.

INVESTIGACIÓN OPERATIVA

El trabajo en la economía y la investigación operativa ha contribuido en gran medida a la noción de agente racional que aquí se presenta, aunque durante muchos años la investigación en el campo de la IA se ha desarrollado por sendas separadas. Una razón fue la **complejidad** aparente que trae consigo el tomar decisiones racionales. Herbert Simon (1916-2001), uno de los primeros en investigar en el campo de la IA, ganó el premio Nobel en Economía en 1978 por su temprano trabajo, en el que mostró que los modelos basados en **satisfacción** (que toman decisiones que son «suficientemente buenas», en vez de realizar cálculos laboriosos para alcanzar decisiones óptimas) proporcionaban una descripción mejor del comportamiento humano real (Simon, 1947). En los años 90, hubo un resurgimiento del interés en las técnicas de decisión teórica para sistemas basados en agentes (Wellman, 1995).

SATISFACCIÓN

Neurociencia (desde el año 1861 hasta el presente)

- ¿Cómo procesa información el cerebro?

NEUROCIENCIA

La **Neurociencia** es el estudio del sistema neurológico, y en especial del cerebro. La forma exacta en la que en un cerebro se genera el pensamiento es uno de los grandes misterios de la ciencia. Se ha observado durante miles de años que el cerebro está de alguna manera involucrado en los procesos de pensamiento, ya que fuertes golpes en la cabeza pueden ocasionar minusvalía mental. También es ampliamente conocido que los cerebros humanos son de alguna manera diferentes; aproximadamente en el 335 a.C. Aristóteles escribió, «de entre todos los animales el hombre tiene el cerebro más grande en proporción a su tamaño»⁶. Aunque, no fue hasta mediados del siglo XVIII cuando se aceptó mayoritariamente que el cerebro es la base de la conciencia. Hasta este momento, se pensaba que estaba localizado en el corazón, el bazo y la glándula pineal.

El estudio de Paul Broca (1824-1880) sobre la afasia (dificultad para hablar) en pacientes con el cerebro dañado, en 1861, le dio fuerza a este campo y convenció a la sociedad médica de la existencia de áreas localizadas en el cerebro responsables de funciones cognitivas específicas. En particular, mostró que la producción del habla se localizaba en una parte del hemisferio izquierdo; hoy en día conocida como el área de Broca⁷. En esta época ya se sabía que el cerebro estaba formado por células nerviosas o **neuronas**, pero no fue hasta 1873 cuando Camillo Golgi (1843-1926) desarrolló una técnica de coloración que permitió la observación de neuronas individuales en el cerebro (véase la Figura 1.2). Santiago Ramón y Cajal (1852-1934) utilizó esta técnica en sus estudios pioneros sobre la estructura neuronal del cerebro⁸.

En la actualidad se dispone de información sobre la relación existente entre las áreas del cerebro y las partes del cuerpo humano que controlan o de las que reciben impulsos

⁶ Desde entonces, se ha descubierto que algunas especies de delfines y ballenas tienen cerebros relativamente grandes. Ahora se piensa que el gran tamaño de los cerebros humanos se debe en parte a la mejora reciente en su sistema de refrigeración.

⁷ Muchos citan a Alexander Hood (1824) como una fuente posiblemente anterior.

⁸ Golgi insistió en la creencia de que las funciones cerebrales se desarrollaron inicialmente en el medio continuo en el que las neuronas estaban inmersas, mientras que Cajal propuso la «doctrina neuronal». Ambos compartieron el premio Nobel en 1906 pronunciando un discurso de aceptación antagónico.

NEURONAS

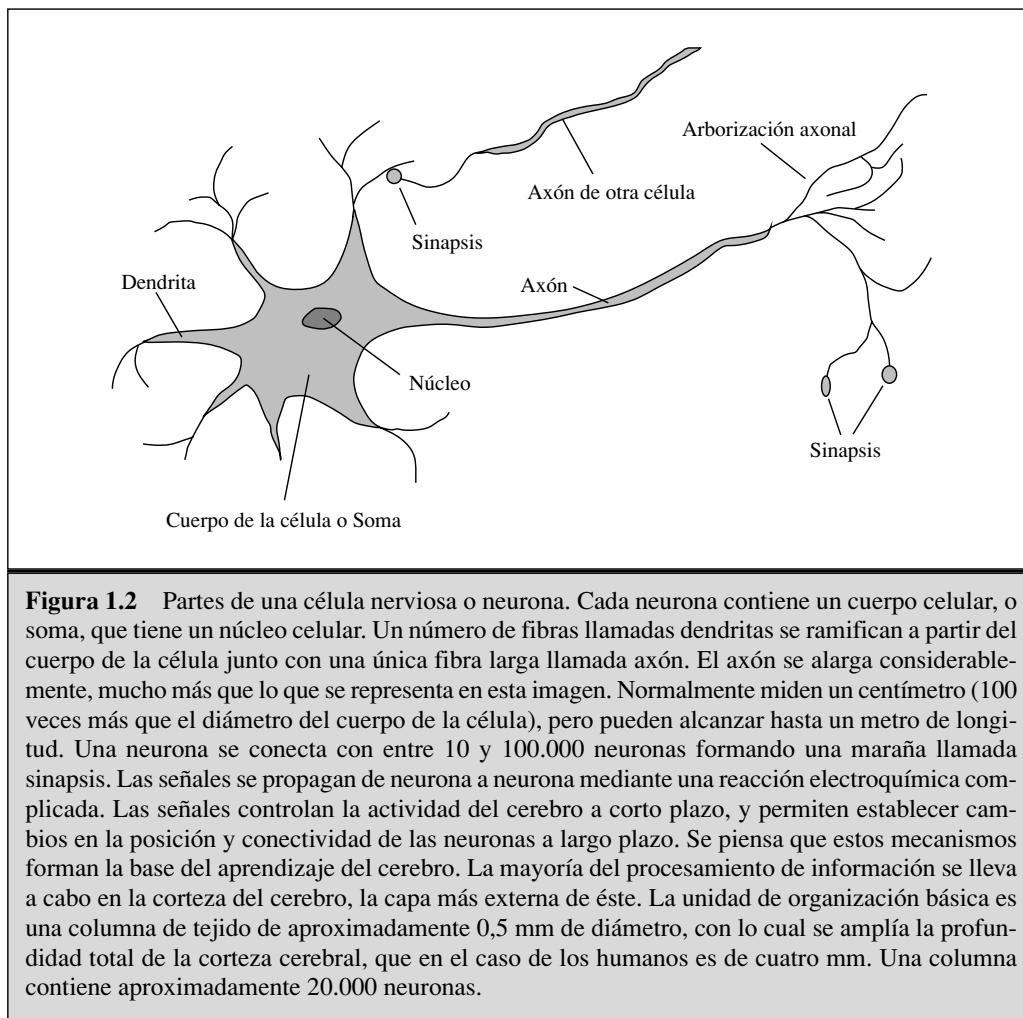


Figura 1.2 Partes de una célula nerviosa o neurona. Cada neurona contiene un cuerpo celular, o soma, que tiene un núcleo celular. Un número de fibras llamadas dendritas se ramifican a partir del cuerpo de la célula junto con una única fibra larga llamada axón. El axón se alarga considerablemente, mucho más que lo que se representa en esta imagen. Normalmente miden un centímetro (100 veces más que el diámetro del cuerpo de la célula), pero pueden alcanzar hasta un metro de longitud. Una neurona se conecta con entre 10 y 100.000 neuronas formando una maraña llamada sinapsis. Las señales se propagan de neurona a neurona mediante una reacción electroquímica complicada. Las señales controlan la actividad del cerebro a corto plazo, y permiten establecer cambios en la posición y conectividad de las neuronas a largo plazo. Se piensa que estos mecanismos forman la base del aprendizaje del cerebro. La mayoría del procesamiento de información se lleva a cabo en la corteza del cerebro, la capa más externa de éste. La unidad de organización básica es una columna de tejido de aproximadamente 0,5 mm de diámetro, con lo cual se amplía la profundidad total de la corteza cerebral, que en el caso de los humanos es de cuatro mm. Una columna contiene aproximadamente 20.000 neuronas.

sensoriales. Tales relaciones pueden cambiar de forma radical incluso en pocas semanas, y algunos animales parecen disponer de múltiples posibilidades. Más aún, no se tiene totalmente claro cómo algunas áreas se pueden encargar de ciertas funciones que eran responsabilidad de áreas dañadas. No hay prácticamente ninguna teoría que explique cómo se almacenan recuerdos individuales.

Los estudios sobre la actividad de los cerebros intactos comenzó en 1929 con el descubrimiento del electroencefalograma (EEG) desarrollado por Hans Berger. El reciente descubrimiento de las imágenes de resonancia magnética funcional (IRMF) (Ogawa *et al.*, 1990) está proporcionando a los neurólogos imágenes detalladas de la actividad cerebral sin precedentes, permitiéndoles obtener medidas que se corresponden con procesos cognitivos en desarrollo de manera muy interesante. Este campo está evolucionando gracias a los avances en los estudios en celdas individuales y su actividad neuronal. A pesar de estos avances, nos queda un largo camino para llegar a comprender cómo funcionan todos estos procesos cognitivos.



La conclusión verdaderamente increíble es que *una colección de simples células pueden llegar a generar razonamiento, acción, y conciencia* o, dicho en otras palabras, *los cerebros generan las inteligencias* (Searle, 1992). La única teoría alternativa es el misticismo: que nos dice que existe alguna esfera mística en la que las mentes operan fuera del control de la ciencia física.



Cerebros y computadores digitales realizan tareas bastante diferentes y tienen propiedades distintas. La Figura 1.3 muestra cómo hay 1.000 veces más neuronas en un cerebro humano medio que puertas lógicas en la UCP de un computador estándar. La ley de Moore⁹ predice que el número de puertas lógicas de la UCP se igualará con el de neuronas del cerebro alrededor del año 2020. Por supuesto, poco se puede inferir de esta predicción; más aún, la diferencia en la capacidad de almacenamiento es insignificante comparada con las diferencias en la velocidad de intercambio y en paralelismo. Los circuitos de los computadores pueden ejecutar una instrucción en un nanosegundo, mientras que las neuronas son millones de veces más lentas. Las neuronas y las sinapsis del cerebro están activas simultáneamente, mientras que los computadores actuales tienen una o como mucho varias UCP. Por tanto, incluso sabiendo que un computador es un millón de veces más rápido en cuanto a su velocidad de intercambio, el cerebro acaba siendo 100.000 veces más rápido en lo que hace.

Psicología (desde el año 1879 hasta el presente)

- ¿Cómo piensan y actúan los humanos y los animales?

La psicología científica se inició con los trabajos del físico alemán Hermann von Helmholtz (1821-1894), según se referencia habitualmente, y su discípulo Wilhelm Wundt (1832-1920). Helmholtz aplicó el método científico al estudio de la vista humana, y su obra *Handbook of Physiological Optics*, todavía en nuestros días, se considera como «el tratado actual más importante sobre la física y la fisiología de la vista humana» (Nalwa, 1993, p. 15). En 1879, Wundt abrió el primer laboratorio de psicología experimental en

	Computador	Cerebro Humano
Unidades computacionales	1 UCP, 10^8 puertas	10^{11} neuronas
Unidades de Almacenamiento	10^{10} bits RAM	10^{11} neuronas
	10^{11} bits disco	10^{14} sinapsis
Duración de un ciclo	10^{-9} sec	10^{-3} sec
Ancho de banda	10^{10} bits/sec	10^{14} bits/sec
Memoria actualización/sec	10^9	10^{14}

Figura 1.3 Comparación básica entre los recursos de cómputo generales de que disponen los computadores (*circa 2003*) y el cerebro. Las cifras correspondientes a los computadores se han incrementado en al menos un factor 10 desde la primera edición de este libro, y se espera que suceda lo mismo en esta década. Las cifras correspondientes al cerebro no han cambiado en los últimos 10.000 años.

⁹ La ley de Moore dice que el número de transistores por pulgada cuadrada se duplica cada año o año y medio. La capacidad del cerebro humano se dobla aproximadamente cada dos o cuatro millones de años.

CONDUCTISMO

la Universidad de Leipzig. Wundt puso mucho énfasis en la realización de experimentos controlados cuidadosamente en la que sus operarios realizaban tareas de percepción o asociación al tiempo que sometían a introspección sus procesos mentales. Los meticolosos controles evolucionaron durante un largo período de tiempo hasta convertir la psicología en una ciencia, pero la naturaleza subjetiva de los datos hizo poco probable que un investigador pudiera contradecir sus propias teorías. Biólogos, estudiando el comportamiento humano, por el contrario, carecían de datos introspectivos y desarrollaron una metodología objetiva, tal y como describe H. S. Jennings (1906) en su influyente trabajo *Behavior of the Lower Organisms*. El movimiento **conductista**, liderado por John Watson (1878-1958) aplicó este punto de vista a los humanos, rechazando *cualquier* teoría en la que interviniéran procesos mentales, argumentando que la introspección no aportaba una evidencia fiable. Los conductistas insistieron en el estudio exclusivo de mediciones objetivas de percepciones (o *estímulos*) sobre animales y de las acciones resultantes (o *respuestas*). Construcciones mentales como conocimientos, creencias, objetivos y pasos en un razonamiento quedaron descartadas por ser consideradas «psicología popular» no científica. El conductismo hizo muchos descubrimientos utilizando ratas y palomas, pero tuvo menos éxito en la comprensión de los seres humanos. Aún así, su influencia en la psicología fue notable (especialmente en Estados Unidos) desde aproximadamente 1920 hasta 1960.

PSICOLOGÍA COGNITIVA

La conceptualización del cerebro como un dispositivo de procesamiento de información, característica principal de la **psicología cognitiva**, se remonta por lo menos a las obras de William James¹⁰ (1842-1910). Helmholtz también pone énfasis en que la percepción entraña cierto tipo de inferencia lógica inconsciente. Este punto de vista cognitivo se vio eclipsado por el conductismo en Estados Unidos, pero en la Unidad de Psicología Aplicada de Cambridge, dirigida por Frederic Bartlett (1886-1969), los modelos cognitivos emergieron con fuerza. La obra *The Nature of Explanation*, de Kenneth Craik (1943), discípulo y sucesor de Bartlett, re establece enérgicamente la legitimidad de términos «mentales» como creencias y objetivos, argumentando que son tan científicos como lo pueden ser la presión y la temperatura cuando se habla acerca de los gases, a pesar de que éstos estén formados por moléculas que no tienen ni presión ni temperatura. Craik establece tres elementos clave que hay que tener en cuenta para diseñar un agente basado en conocimiento: (1) el estímulo deberá ser traducido a una representación interna, (2) esta representación se debe manipular mediante procesos cognitivos para así generar nuevas representaciones internas, y (3) éstas, a su vez, se traducirán de nuevo en acciones. Dejó muy claro por qué consideraba que estos eran los requisitos idóneos para diseñar un agente:

Si el organismo tiene en su cabeza «un modelo a pequeña escala» de la realidad externa y de todas sus posibles acciones, será capaz de probar diversas opciones, decidir cuál es la mejor, planificar su reacción ante posibles situaciones futuras antes de que éstas surjan, emplear lo aprendido de experiencias pasadas en situaciones presentes y futuras, y en todo momento, reaccionar ante los imprevistos que acontezcan de manera satisfactoria, segura y más competente (Craik, 1943).

¹⁰ William James era hermano del novelista Henry James. Se comenta que Henry escribió novelas narrativas como si se tratara de psicología y William escribió sobre psicología como si se tratara de novelas narrativas.

CIENCIA COGNITIVA

Después de la muerte de Craik en un accidente de bicicleta en 1945, Donald Broadbent continuó su trabajo, y su libro *Perception and Communication* (1958) incluyó algunos de los primeros modelos de procesamiento de información del fenómeno psicológico. Mientras tanto, en Estados Unidos el desarrollo del modelo computacional llevó a la creación del campo de la **ciencia cognitiva**. Se puede decir que este campo comenzó en un simposio celebrado en el MIT, en septiembre de 1956 (como se verá a continuación este evento tuvo lugar sólo dos meses después de la conferencia en la que «nació» la IA). En este simposio, George Miller presentó *The Magic Number Seven*, Noam Chomsky presentó *Three Models of Language*, y Allen Newell y Herbert Simon presentaron *The Logic Theory Machine*. Estos tres artículos influyentes mostraron cómo se podían utilizar los modelos informáticos para modelar la psicología de la memoria, el lenguaje y el pensamiento lógico, respectivamente. Los psicólogos comparten en la actualidad el punto de vista común de que «la teoría cognitiva debe ser como un programa de computador» (Anderson, 1980), o dicho de otra forma, debe describir un mecanismo de procesamiento de información detallado, lo cual lleva consigo la implementación de algunas funciones cognitivas.

Ingeniería computacional (desde el año 1940 hasta el presente)

- ¿Cómo se puede construir un computador eficiente?

Para que la inteligencia artificial pueda llegar a ser una realidad se necesitan dos cosas: inteligencia y un artefacto. El computador ha sido el artefacto elegido. El computador electrónico digital moderno se inventó de manera independiente y casi simultánea por científicos en tres países involucrados en la Segunda Guerra Mundial. El equipo de Alan Turing construyó, en 1940, el primer computador *operacional* de carácter electromecánico, llamado Heath Robinson¹¹, con un único propósito: descifrar mensajes alemanes. En 1943 el mismo grupo desarrolló el Colossus, una máquina potente de propósito general basada en válvulas de vacío¹². El primer computador operacional *programable* fue el Z-3, inventado por Konrad Zuse en Alemania, en 1941. Zuse también inventó los números de coma flotante y el primer lenguaje de programación de alto nivel, Plankalkül. El primer computador *electrónico*, el ABC, fue creado por John Atanasoff junto a su discípulo Clifford Berry entre 1940 y 1942 en la Universidad Estatal de Iowa. Las investigaciones de Atanasoff recibieron poco apoyo y reconocimiento; el ENIAC, desarrollado en el marco de un proyecto militar secreto, en la Universidad de Pensilvania, por un equipo en el que trabajaban entre otros John Mauchly y John Eckert, puede considerarse como el precursor de los computadores modernos.

Desde mediados del siglo pasado, cada generación de dispositivos *hardware* ha llevado un aumento en la velocidad de proceso y en la capacidad de almacenamiento,

¹¹ Heath Robinson fue un caricaturista famoso por sus dibujos, que representaban artefactos de uso diario, caprichosos y absurdamente complicados, por ejemplo, uno para untar mantequilla en el pan tostado.

¹² En la postguerra, Turing quiso utilizar estos computadores para investigar en el campo de la IA, por ejemplo, desarrollando uno de los primeros programas para jugar a la ajedrez (Turing *et al.*, 1953). El gobierno británico bloqueó sus esfuerzos.

así como una reducción de precios. La potencia de los computadores se dobla cada 18 meses aproximadamente y seguirá a este ritmo durante una o dos décadas más. Después, se necesitará ingeniería molecular y otras tecnologías novedosas.

Por supuesto que antes de la aparición de los computadores ya había dispositivos de cálculo. Las primeras máquinas automáticas, que datan del siglo XVII, ya se mencionaron en la página seis. La primera máquina programable fue un telar, desarrollado en 1805 por Joseph Marie Jacquard (1752-1834) que utilizaba tarjetas perforadas para almacenar información sobre los patrones de los bordados. A mediados del siglo XIX, Charles Babbage (1792-1871) diseñó dos máquinas, que no llegó a construir. La «Máquina de Diferencias», que aparece en la portada de este manuscrito, se concibió con la intención de facilitar los cálculos de tablas matemáticas para proyectos científicos y de ingeniería. Finalmente se construyó y se presentó en 1991 en el Museo de la Ciencia de Londres (Swade, 1993). La «Máquina Analítica» de Babbage era mucho más ambiciosa: incluía memoria direccionable, programas almacenados y saltos condicionales; fue el primer artefacto dotado de los elementos necesarios para realizar una computación universal. Ada Lovelace, colega de Babbage e hija del poeta Lord Byron, fue seguramente la primera programadora (el lenguaje de programación Ada se llama así en honor a esta programadora). Ella escribió programas para la inacabada Máquina Analítica e incluso especuló acerca de la posibilidad de que la máquina jugara al ajedrez y compusiese música.

La IA también tiene una deuda con la parte *software* de la informática que ha proporcionado los sistemas operativos, los lenguajes de programación, y las herramientas necesarias para escribir programas modernos (y artículos sobre ellos). Sin embargo, en este área la deuda se ha saldado: la investigación en IA ha generado numerosas ideas novedosas de las que se ha beneficiado la informática en general, como por ejemplo el tiempo compartido, los intérpretes imperativos, los computadores personales con interfaces gráficas y ratones, entornos de desarrollo rápido, listas enlazadas, administración automática de memoria, y conceptos claves de la programación simbólica, funcional, dinámica y orientada a objetos.

Teoría de control y cibernética (desde el año 1948 hasta el presente)

- ¿Cómo pueden los artefactos operar bajo su propio control?

Ktesibios de Alejandría (250 a.C.) construyó la primera máquina auto controlada: un reloj de agua con un regulador que mantenía el flujo de agua circulando por él, con un ritmo constante y predecible. Esta invención cambió la definición de lo que un artefacto podía hacer. Anteriormente, solamente seres vivos podían modificar su comportamiento como respuesta a cambios en su entorno. Otros ejemplos de sistemas de control auto regulables y retroalimentados son el motor de vapor, creado por James Watt (1736-1819), y el termostato, inventado por Cornelis Drebbel (1572-1633), que también inventó el submarino. La teoría matemática de los sistemas con retroalimentación estables se desarrolló en el siglo XIX.

TEORÍA DE CONTROL

La figura central del desarrollo de lo que ahora se llama la **teoría de control** fue Norbert Wiener (1894-1964). Wiener fue un matemático brillante que trabajó en sistemas de control biológicos y mecánicos y en sus vínculos con la cognición. De la misma forma que Craik (quien también utilizó sistemas de control como modelos psicológicos), Wiener y sus colegas Arturo Rosenblueth y Julian Bigelow desafiaron la ortodoxia conductista (Rosenblueth *et al.*, 1943). Ellos veían el comportamiento determinista como algo emergente de un mecanismo regulador que intenta minimizar el «error» (la diferencia entre el estado presente y el estado objetivo). A finales de los años 40, Wiener, junto a Warren McCulloch, Walter Pitts y John von Neumann, organizaron una serie de conferencias en las que se exploraban los nuevos modelos cognitivos matemáticos y computacionales, e influyeron en muchos otros investigadores en el campo de las ciencias del comportamiento. El libro de Wiener, *Cybernetics* (1948), fue un *bestseller* y desveló al público las posibilidades de las máquinas con inteligencia artificial.

La teoría de control moderna, especialmente la rama conocida como control óptimo estocástico, tiene como objetivo el diseño de sistemas que maximizan una **función objetivo** en el tiempo. Lo cual se asemeja ligeramente a nuestra visión de lo que es la IA: diseño de sistemas que se comportan de forma óptima. ¿Por qué, entonces, IA y teoría de control son dos campos diferentes, especialmente teniendo en cuenta la cercana relación entre sus creadores? La respuesta está en el gran acoplamiento existente entre las técnicas matemáticas con las que estaban familiarizados los investigadores y entre los conjuntos de problemas que se abordaban desde cada uno de los puntos de vista. El cálculo y el álgebra matricial, herramientas de la teoría de control, se utilizaron en la definición de sistemas que se podían describir mediante conjuntos fijos de variables continuas; más aún, el análisis exacto es sólo posible en sistemas *lineales*. La IA se fundó en parte para escapar de las limitaciones matemáticas de la teoría de control en los años 50. Las herramientas de inferencia lógica y computación permitieron a los investigadores de IA afrontar problemas relacionados con el lenguaje, visión y planificación, que estaban completamente fuera del punto de mira de la teoría de control.

Lingüística (desde el año 1957 hasta el presente)

- ¿Cómo está relacionado el lenguaje con el pensamiento?

En 1957, B. F. Skinner publicó *Verbal Behavior*. La obra presentaba una visión extensa y detallada desde el enfoque conductista al aprendizaje del lenguaje, y estaba escrita por los expertos más destacados de este campo. Curiosamente, una revisión de este libro llegó a ser tan famosa como la obra misma, y provocó el casi total desinterés por el conductismo. El autor de la revisión fue Noam Chomsky, quien acababa de publicar un libro sobre su propia teoría, *Syntactic Structures*. Chomsky mostró cómo la teoría conductista no abordaba el tema de la creatividad en el lenguaje: no explicaba cómo es posible que un niño sea capaz de entender y construir oraciones que nunca antes ha escuchado. La teoría de Chomsky (basada en modelos sintácticos que se remontaban al lingüista indio Panini, aproximadamente 350 a.C.) sí podía explicar lo anterior y, a diferencia de teorías anteriores, poseía el formalismo suficiente como para permitir su programación.

La lingüística moderna y la IA «nacieron», al mismo tiempo y maduraron juntas, solapándose en un campo híbrido llamado **lingüística computacional o procesamiento del lenguaje natural**. El problema del entendimiento del lenguaje se mostró pronto mucho más complejo de lo que se había pensado en 1957. El entendimiento del lenguaje requiere la comprensión de la materia bajo estudio y de su contexto, y no solamente el entendimiento de la estructura de las sentencias. Lo cual puede parecer obvio, pero no lo fue para la mayoría de la comunidad investigadora hasta los años 60. Gran parte de los primeros trabajos de investigación en el área de la **representación del conocimiento** (el estudio de cómo representar el conocimiento de forma que el computador pueda razonar a partir de dicha representación) estaban vinculados al lenguaje y a la búsqueda de información en el campo del lenguaje, y su base eran las investigaciones realizadas durante décadas en el análisis filosófico del lenguaje.

1.3 Historia de la inteligencia artificial

Una vez revisado el material básico estamos ya en condiciones de cubrir el desarrollo de la IA propiamente dicha.

Génesis de la inteligencia artificial (1943-1955)

Warren McCulloch y Walter Pitts (1943) han sido reconocidos como los autores del primer trabajo de IA. Partieron de tres fuentes: conocimientos sobre la fisiología básica y funcionamiento de las neuronas en el cerebro, el análisis formal de la lógica proposicional de Russell y Whitehead y la teoría de la computación de Turing. Propusieron un modelo constituido por neuronas artificiales, en el que cada una de ellas se caracterizaba por estar «activada» o «desactivada»; la «activación» se daba como respuesta a la estimulación producida por una cantidad suficiente de neuronas vecinas. El estado de una neurona se veía como «equivalente, de hecho, a una proposición con unos estímulos adecuados». Mostraron, por ejemplo, que cualquier función de cómputo podría calcularse mediante alguna red de neuronas interconectadas, y que todos los conectores lógicos (*and*, *or*, *not*, etc.) se podrían implementar utilizando estructuras de red sencillas. McCulloch y Pitts también sugirieron que redes adecuadamente definidas podrían aprender. Donald Hebb (1949) propuso y demostró una sencilla regla de actualización para modificar las intensidades de las conexiones entre neuronas. Su regla, ahora llamada **de aprendizaje Hebbiano o de Hebb**, sigue vigente en la actualidad.

Dos estudiantes graduados en el Departamento de Matemáticas de Princeton, Marvin Minsky y Dean Edmonds, construyeron el primer computador a partir de una red neuronal en 1951. El SNARC, como se llamó, utilizaba 3.000 válvulas de vacío y un mecanismo de piloto automático obtenido de los desechos de un avión bombardero B-24 para simular una red con 40 neuronas. El comité encargado de evaluar el doctorado de Minsky veía con escepticismo el que este tipo de trabajo pudiera considerarse como matemático, pero se dice que von Newmann dijo, «Si no lo es actualmente, algún día lo será».

Minsky posteriormente probó teoremas influyentes que mostraron las limitaciones de la investigación con redes neuronales.

Hay un número de trabajos iniciales que se pueden caracterizar como de IA, pero fue Alan Turing quien articuló primero una visión de la IA en su artículo *Computing Machinery and Intelligence*, en 1950. Ahí, introdujo la prueba de Turing, el aprendizaje automático, los algoritmos genéricos y el aprendizaje por refuerzo.

Nacimiento de la inteligencia artificial (1956)

Princeton acogió a otras de las figuras señeras de la IA, John McCarthy. Posteriormente a su graduación, McCarthy se transladó al Dartmouth College, que se erigiría en el lugar del nacimiento oficial de este campo. McCarthy convenció a Minsky, Claude Shannon y Nathaniel Rochester para que le ayudaran a aumentar el interés de los investigadores americanos en la teoría de autómatas, las redes neuronales y el estudio de la inteligencia. Organizaron un taller con una duración de dos meses en Dartmouth en el verano de 1956. Hubo diez asistentes en total, entre los que se incluían Trenchard More de Princeton, Arthur Samuel de IBM, y Ray Solomonoff y Oliver Selfridge del MIT.

Dos investigadores del Carnegie Tech¹³, Allen Newell y Herbert Simon, acapararon la atención. Si bien los demás también tenían algunas ideas y, en algunos casos, programas para aplicaciones determinadas como el juego de damas, Newell y Simon contaban ya con un programa de razonamiento, el Teórico Lógico (TL), del que Simon afirmaba: «Hemos inventado un programa de computación capaz de pensar de manera no numérica, con lo que ha quedado resuelto el venerable problema de la dualidad mente-cuerpo»¹⁴. Poco después del término del taller, el programa ya era capaz de demostrar gran parte de los teoremas del Capítulo 2 de *Principia Matemática* de Russell y Whitehead. Se dice que Russell se manifestó complacido cuando Simon le mostró que la demostración de un teorema que el programa había generado era más corta que la que aparecía en *Principia*. Los editores de la revista *Journal of Symbolic Logic* resultaron menos impresionados y rechazaron un artículo cuyos autores eran Newell, Simon y el Teórico Lógico (TL).

El taller de Dartmouth no produjo ningún avance notable, pero puso en contacto a las figuras importantes de este campo. Durante los siguientes 20 años, el campo estuvo dominado por estos personajes, así como por sus estudiantes y colegas del MIT, CMU, Stanford e IBM. Quizá lo último que surgió del taller fue el consenso en adoptar el nuevo nombre propuesto por McCarthy para este campo: **Inteligencia Artificial**. Quizá «racionalidad computacional» hubiese sido más adecuado, pero «IA» se ha mantenido.

Revisando la propuesta del taller de Dartmouth (McCarthy *et al.*, 1955), se puede apreciar por qué fue necesario para la IA convertirse en un campo separado. ¿Por qué

¹³ Actualmente Universidad Carnegie Mellon (UCM).

¹⁴ Newell y Simon también desarrollaron un lenguaje de procesamiento de listas, IPL, para poder escribir el TL. No disponían de un compilador y lo tradujeron a código máquina a mano. Para evitar errores, trabajaron en paralelo, diciendo en voz alta números binarios, conforme escribían cada instrucción para asegurarse de que ambos coincidían.

no todo el trabajo hecho en el campo de la IA se ha realizado bajo el nombre de teoría de control, o investigación operativa, o teoría de la decisión, que, después de todo, persiguen objetivos similares a los de la IA? O, ¿por qué no es la IA una rama de las matemáticas? La primera respuesta es que la IA desde el primer momento abarcó la idea de duplicar facultades humanas como la creatividad, la auto-mejora y el uso del lenguaje. Ninguno de los otros campos tenían en cuenta esos temas. La segunda respuesta está relacionada con la metodología. La IA es el único de estos campos que es claramente una rama de la informática (aunque la investigación operativa comparte el énfasis en la simulación por computador), además la IA es el único campo que persigue la construcción de máquinas que funcionen automáticamente en medios complejos y cambiantes.

Entusiasmo inicial, grandes esperanzas (1952-1969)

Los primeros años de la IA estuvieron llenos de éxitos (aunque con ciertas limitaciones). Teniendo en cuenta lo primitivo de los computadores y las herramientas de programación de aquella época, y el hecho de que sólo unos pocos años antes, a los computadores se les consideraba como artefactos que podían realizar trabajos aritméticos y nada más, resultó sorprendente que un computador hiciese algo remotamente inteligente. La comunidad científica, en su mayoría, prefirió creer que «una máquina nunca podría hacer *tareas*» (véase el Capítulo 26 donde aparece una extensa lista de *tareas* recopilada por Turing). Naturalmente, los investigadores de IA responderían demostrando la realización de una *tarea* tras otra. John McCarthy se refiere a esta época como la era de «¡Mira, mamá, ahora sin manos!».

Al temprano éxito de Newell y Simon siguió el del sistema de resolución general de problemas, o SRGP. A diferencia del Teórico Lógico, desde un principio este programa se diseñó para que imitara protocolos de resolución de problemas de los seres humanos. Dentro del limitado número de puzzles que podía manejar, resultó que la secuencia en la que el programa consideraba que los subobjetivos y las posibles acciones eran semejantes a la manera en que los seres humanos abordaban los mismos problemas. Es decir, el SRGP posiblemente fue el primer programa que incorporó el enfoque de «pensar como un ser humano». El éxito del SRGP y de los programas que le siguieron, como los modelos de cognición, llevaron a Newell y Simon (1976) a formular la famosa hipótesis del **sistema de símbolos físicos**, que afirma que «un sistema de símbolos físicos tiene los medios suficientes y necesarios para generar una acción inteligente». Lo que ellos querían decir es que cualquier sistema (humano o máquina) que exhibiese inteligencia debería operar manipulando estructuras de datos compuestas por símbolos. Posteriormente se verá que esta hipótesis se ha modificado atendiendo a distintos puntos de vista.

En IBM, Nathaniel Rochester y sus colegas desarrollaron algunos de los primeros programas de IA. Herbert Gelernter (1959) construyó el demostrador de teoremas de geometría (DTG), el cual era capaz de probar teoremas que muchos estudiantes de matemáticas podían encontrar muy complejos de resolver. A comienzos 1952, Arthur Samuel escribió una serie de programas para el juego de las damas que eventualmente aprendieron a jugar hasta alcanzar un nivel equivalente al de un *amateur*. De paso, echó por

tierra la idea de que los computadores sólo pueden hacer lo que se les dice: su programa pronto aprendió a jugar mejor que su creador. El programa se presentó en la televisión en febrero de 1956 y causó una gran impresión. Como Turing, Samuel tenía dificultades para obtener el tiempo de cómputo. Trabajaba por las noches y utilizaba máquinas que aún estaban en período de prueba en la planta de fabricación de IBM. El Capítulo 6 trata el tema de los juegos, y en el Capítulo 21 se describe con detalle las técnicas de aprendizaje utilizadas por Samuel.

John McCarthy se trasladó de Dartmouth al MIT, donde realizó tres contribuciones cruciales en un año histórico: 1958. En el Laboratorio de IA del MIT Memo Número 1, McCarthy definió el lenguaje de alto nivel **Lisp**, que se convertiría en el lenguaje de programación dominante en la IA. Lisp es el segundo lenguaje de programación más antiguo que se utiliza en la actualidad, ya que apareció un año después de FORTRAN. Con Lisp, McCarthy tenía la herramienta que necesitaba, pero el acceso a los escasos y costosos recursos de cómputo aún era un problema serio. Para solucionarlo, él, junto a otros miembros del MIT, inventaron el tiempo compartido. También, en 1958, McCarthy publicó un artículo titulado *Programs with Common Sense*, en el que describía el Generador de Consejos, un programa hipotético que podría considerarse como el primer sistema de IA completo. Al igual que el Teórico Lógico y el Demostrador de Teoremas de Geometría, McCarthy diseñó su programa para buscar la solución a problemas utilizando el conocimiento. Pero, a diferencia de los otros, manejaba el conocimiento general del mundo. Por ejemplo, mostró cómo algunos axiomas sencillos permitían a un programa generar un plan para conducirnos hasta el aeropuerto y tomar un avión. El programa se diseñó para que aceptase nuevos axiomas durante el curso normal de operación, permitiéndole así ser competente en áreas nuevas, sin *necesidad de reprogramación*. El Generador de Consejos incorporaba así los principios centrales de la representación del conocimiento y el razonamiento: es útil contar con una representación formal y explícita del mundo y de la forma en que la acción de un agente afecta al mundo, así como, ser capaces de manipular estas representaciones con procesos deductivos. Es sorprendente constatar cómo mucho de lo propuesto en el artículo escrito en 1958 permanece vigente incluso en la actualidad.

1958 fue el año en el que Marvin Minsky se trasladó al MIT. Sin embargo, su colaboración inicial no duró demasiado. McCarthy se centró en la representación y el razonamiento con lógica formal, mientras que Minsky estaba más interesado en lograr que los programas funcionaran y eventualmente desarrolló un punto de vista anti-lógico. En 1963 McCarthy creó el Laboratorio de IA en Stanford. Su plan para construir la versión más reciente del Generador de Consejos con ayuda de la lógica sufrió un considerable impulso gracias al descubrimiento de J. A. Robinson del método de resolución (un algoritmo completo para la demostración de teoremas para la lógica de primer orden; véase el Capítulo 9). El trabajo realizado en Stanford hacía énfasis en los métodos de propósito general para el razonamiento lógico. Algunas aplicaciones de la lógica incluyan los sistemas de planificación y respuesta a preguntas de Cordell Green (1969b), así como el proyecto de robótica de Shakey en el nuevo Instituto de Investigación de Stanford (Stanford Research Institute, SRI). Este último proyecto, comentado en detalle en el Capítulo 25, fue el primero que demostró la total integración del razonamiento lógico y la actividad física.

Minsky supervisó el trabajo de una serie de estudiantes que eligieron un número de problemas limitados cuya solución pareció requerir inteligencia. Estos dominios limi-

MICROMUNDOS

tados se conocen como **micromundos**. El programa SAINT de James Slagle (1963a) fue capaz de resolver problemas de integración de cálculo en forma cerrada, habituales en los primeros cursos de licenciatura. El programa ANALOGY de Tom Evans (1968) resolvía problemas de analogía geométrica que se aplicaban en las pruebas de medición de inteligencia, semejante al de la Figura 1.4. El programa STUDENT de Daniel Bobrow (1967) podía resolver problemas de álgebra del tipo:

Si el número de clientes de Tom es dos veces el cuadrado del 20 por ciento de la cantidad de anuncios que realiza, y éstos ascienden a 45, ¿cuántos clientes tiene Tom?

El micromundo más famoso fue el mundo de los bloques, que consiste en un conjunto de bloques sólidos colocados sobre una mesa (más frecuentemente, en la simulación de ésta), como se muestra en la Figura 1.5. Una tarea típica de este mundo es la reordenación de los bloques de cierta manera, con la ayuda de la mano de un robot que es capaz de tomar un bloque cada vez. El mundo de los bloques fue el punto de partida para el proyecto de visión de David Huffman (1971), la visión y el trabajo de propagación con restricciones de David Waltz (1975), la teoría del aprendizaje de Patrick Winston (1970), del programa para la comprensión de lenguaje natural de Terry Winograd (1972) y del planificador de Scott Fahlman (1974).

El trabajo realizado por McCulloch y Pitts con redes neuronales hizo florecer esta área. El trabajo de Winograd y Cowan (1963) mostró cómo un gran número de elementos podría representar un concepto individual de forma colectiva, lo cual llevaba consigo un aumento proporcional en robustez y paralelismo. Los métodos de aprendizaje de Hebb se reforzaron con las aportaciones de Bernie Widrow (Widrow y Hoff, 1960; Widrow, 1962), quien llamó **adalines** a sus redes, y por Frank Rosenblatt (1962) con sus **perceptrones**. Rosenblatt demostró el famoso teorema del perceptrón, con lo que mostró que su algoritmo de aprendizaje podría ajustar las intensidades de las conexiones de un perceptrón para que se adaptaran a los datos de entrada, siempre y cuando existiera una correspondencia. Estos temas se explicarán en el Capítulo 20.

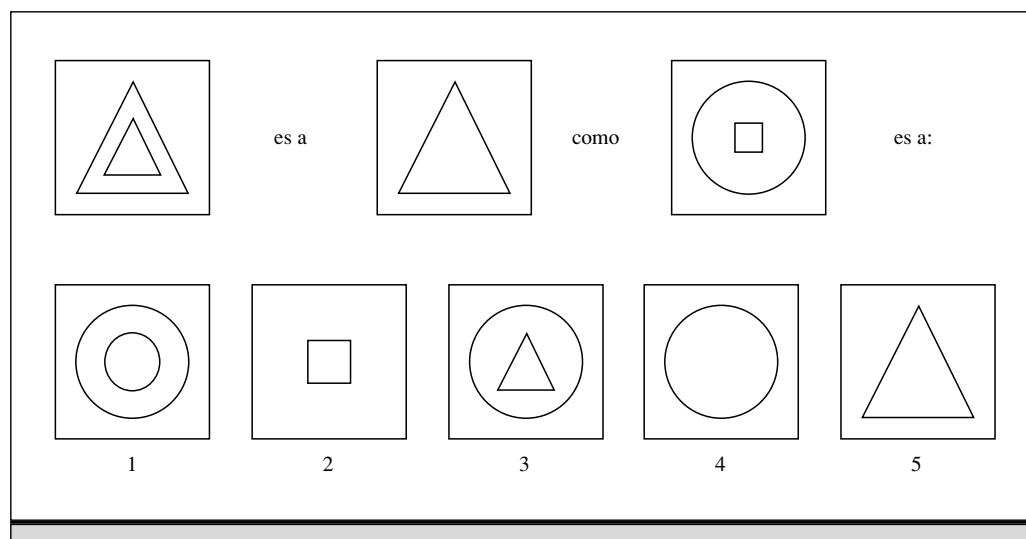
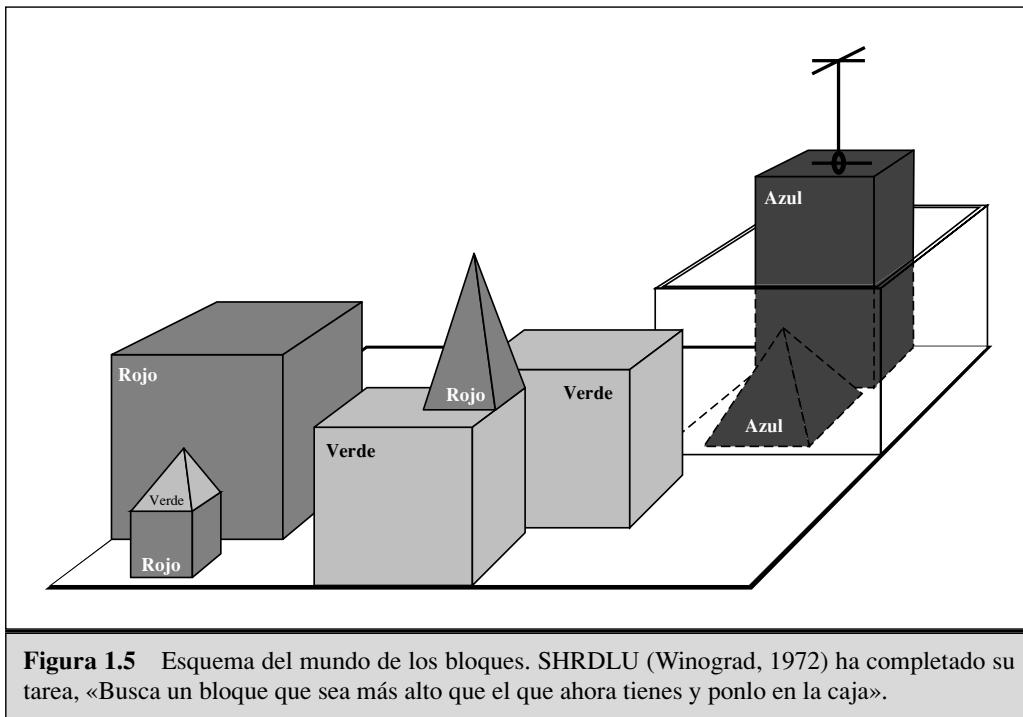


Figura 1.4 Ejemplo de un problema resuelto por el programa ANALOGY de Evans.



Una dosis de realidad (1966-1973)

Desde el principio, los investigadores de IA hicieron públicas, sin timidez, predicciones sobre el éxito que les esperaba. Con frecuencia, se cita el siguiente comentario realizado por Herbert Simon en 1957:

Sin afán de sorprenderlos y dejarlos atónitos, pero la forma más sencilla que tengo de resumirlo es diciéndoles que actualmente en el mundo existen máquinas capaces de pensar, aprender y crear. Además, su aptitud para hacer lo anterior aumentará rápidamente hasta que (en un futuro previsible) la magnitud de problemas que serán capaces de resolver irá a la par que la capacidad de la mente humana para hacer lo mismo.

Términos como «futuro previsible» pueden interpretarse de formas distintas, pero Simon también hizo predicciones más concretas: como que en diez años un computador llegaría a ser campeón de ajedrez, y que se podría demostrar un importante teorema matemático con una máquina. Estas predicciones se cumplirían (al menos en parte) dentro de los siguientes 40 años y no en diez. El exceso de confianza de Simon se debió a la prometedora actuación de los primeros sistemas de IA en problemas simples. En la mayor parte de los casos resultó que estos primeros sistemas fallaron estrepitosamente cuando se utilizaron en problemas más variados o de mayor dificultad.

El primer tipo de problemas surgió porque la mayoría de los primeros programas contaban con poco o ningún conocimiento de las materia objeto de estudio; obtenían resultados gracias a sencillas manipulaciones sintácticas. Una anécdota típica tuvo lugar cuando se comenzaba a trabajar en la traducción automática, actividad que recibía un

generoso patrocinio del Consejo Nacional para la Investigación de Estados Unidos en un intento de agilizar la traducción de artículos científicos rusos en vísperas del lanzamiento del Sputnik en 1957. Al principio se consideró que todo se reduciría a sencillas transformaciones sintácticas apoyadas en las gramáticas rusas e inglesa y al emplazamiento de palabras mediante un diccionario electrónico, lo que bastaría para obtener el significado exacto de las oraciones. La realidad es que para traducir es necesario contar con un conocimiento general sobre el tema, que permita resolver ambigüedades y así, precisar el contenido de una oración. La famosa retraducción del ruso al inglés de la frase «el espíritu es fuerte pero la carne es débil», cuyo resultado fue «el vodka es bueno pero la carne está podrida» es un buen ejemplo del tipo de dificultades que surgieron. En un informe presentado en 1966, el comité consultivo declaró que «no se ha logrado obtener ninguna traducción de textos científicos generales ni se prevé obtener ninguna en un futuro inmediato». Se canceló todo el patrocinio del gobierno estadounidense que se había asignado a los proyectos académicos sobre traducción. Hoy día, la traducción automática es una herramienta imperfecta pero de uso extendido en documentos técnicos, comerciales, gubernamentales y de Internet.

El segundo problema fue que muchos de los problemas que se estaban intentando resolver mediante la IA eran intratables. La mayoría de los primeros programas de IA resolvían problemas experimentando con diversos pasos hasta que se llegara a encontrar una solución. Esto funcionó en los primeros programas debido a que los micromundos con los que se trabajaba contenían muy pocos objetos, y por lo tanto muy pocas acciones posibles y secuencias de soluciones muy cortas. Antes de que se desarrollara la teoría de la complejidad computacional, se creía que para «aumentar» el tamaño de los programas de forma que estos pudiesen solucionar grandes problemas sería necesario incrementar la velocidad del *hardware* y aumentar las memorias. El optimismo que acompañó al logro de la demostración de problemas, por ejemplo, pronto se vio eclipsado cuando los investigadores fracasaron en la demostración de teoremas que implicaban más de unas pocas decenas de condiciones. *El hecho de que, en principio, un programa sea capaz de encontrar una solución no implica que tal programa encierre todos los mecanismos necesarios para encontrar la solución en la práctica.*

La ilusoria noción de una ilimitada capacidad de cómputo no sólo existió en los programas para la resolución de problemas. Los primeros experimentos en el campo de la **evolución automática** (ahora llamados **algoritmos genéticos**) (Friedberg, 1958; Friedberg *et al.*, 1959) estaban basados en la, sin duda correcta, premisa de que efectuando una adecuada serie de pequeñas mutaciones a un programa de código máquina se podría generar un programa con buen rendimiento aplicable en cualquier tarea sencilla. Después surgió la idea de probar con mutaciones aleatorias aplicando un proceso de selección con el fin de conservar aquellas mutaciones que hubiesen demostrado ser más útiles. No obstante, las miles de horas de CPU dedicadas, no dieron lugar a ningún avance tangible. Los algoritmos genéticos actuales utilizan representaciones mejores y han tenido más éxito.

La incapacidad para manejar la «explosión combinatoria» fue una de las principales críticas que se hicieron a la IA en el informe de Lighthill (Lighthill, 1973), informe en el que se basó la decisión del gobierno británico para retirar la ayuda a las investigaciones sobre IA, excepto en dos universidades. (La tradición oral presenta un cuadro un

poco distinto y más animado, en el que se vislumbran ambiciones políticas y animadversiones personales, cuya descripción está fuera del ámbito de esta obra.)

El tercer obstáculo se derivó de las limitaciones inherentes a las estructuras básicas que se utilizaban en la generación de la conducta inteligente. Por ejemplo, en 1969, en el libro de Minsky y Papert, *Perceptrons*, se demostró que si bien era posible lograr que los perceptrones (una red neuronal simple) aprendieran cualquier cosa que pudiesen representar, su capacidad de representación era muy limitada. En particular, un perceptrón con dos entradas no se podía entrenar para que aprendiese a reconocer cuándo sus dos entradas eran diferentes. Si bien los resultados que obtuvieron no eran aplicables a redes más complejas multicapa, los fondos para la investigación de las redes neuronales se redujeron a prácticamente nada. Es irónico que los nuevos algoritmos de aprendizaje de retroalimentación utilizados en las redes multicapa y que fueron la causa del gran resurgimiento de la investigación en redes neuronales de finales de los años 80, en realidad, se hayan descubierto por primera vez en 1969 (Bryson y Ho, 1969).

Sistemas basados en el conocimiento: ¿clave del poder? (1969-1979)

El cuadro que dibujaba la resolución de problemas durante la primera década de la investigación en la IA estaba centrado en el desarrollo de mecanismos de búsqueda de propósito general, en los que se entrelazaban elementos de razonamiento básicos para encontrar así soluciones completas. A estos procedimientos se les ha denominado **métodos débiles**, debido a que no tratan problemas más amplios o más complejos. La alternativa a los métodos débiles es el uso de conocimiento específico del dominio que facilita el desarrollo de etapas de razonamiento más largas, pudiéndose así resolver casos recurrentes en dominios de conocimiento restringido. Podría afirmarse que para resolver un problema en la práctica, es necesario saber de antemano la correspondiente respuesta.

El programa DENDRAL (Buchanan *et al.*, 1969) constituye uno de los primeros ejemplos de este enfoque. Fue diseñado en Stanford, donde Ed Feigenbaum (discípulo de Herbert Simon), Bruce Buchanan (filósofo convertido en informático) y Joshua Lederberg (genetista ganador del Premio Nobel) colaboraron en la solución del problema de inferir una estructura molecular a partir de la información proporcionada por un espectrómetro de masas. El programa se alimentaba con la fórmula elemental de la molécula (por ejemplo, $C_6H_{13}NO_2$) y el espectro de masas, proporcionando las masas de los distintos fragmentos de la molécula generada después de ser bombardeada con un haz de electrones. Por ejemplo, un espectro de masas con un pico en $m = 15$, correspondería a la masa de un fragmento de metilo (CH_3).

La versión más simple del programa generaba todas las posibles estructuras que correspondieran a la fórmula, luego predecía el espectro de masas que se observaría en cada caso, y comparaba éstos con el espectro real. Como era de esperar, el método anterior resultó pronto inviable para el caso de moléculas con un tamaño considerable. Los creadores de DENDRAL consultaron con químicos analíticos y se dieron cuenta de que éstos trabajaban buscando patrones conocidos de picos en el espectro que sugerían estructuras comunes en la molécula. Por ejemplo, para identificar el subgrupo (con un peso de 28) de las cetonas ($C=O$) se empleó la siguiente regla:

MÉTODOS DÉBILES

si hay dos picos en x_1 y x_2 tales que

- a) $x_1 + x_2 = M + 28$ (siendo M la masa de toda la molécula);
- b) $x_1 - 28$ es un pico alto;
- c) $x_2 - 28$ es un pico alto;
- d) al menos una de x_1 y x_2 es alta.

entonces existe un subgrupo de cetonas

Al reconocer que la molécula contiene una subestructura concreta se reduce el número de posibles candidatos de forma considerable. La potencia de DENDRAL se basaba en que:

Toda la información teórica necesaria para resolver estos problemas se ha proyectado desde su forma general [componente predicho por el espectro] («primeros principios») a formas eficientes especiales («recetas de cocina»). (Feigenbaum *et al.*, 1971)

La trascendencia de DENDRAL se debió a ser el primer sistema de *conocimiento intenso* que tuvo éxito: su base de conocimiento estaba formada por grandes cantidades de reglas de propósito particular. En sistemas diseñados posteriormente se incorporaron también los elementos fundamentales de la propuesta de McCarthy para el Generador de Consejos, la nítida separación del conocimiento (en forma de reglas) de la parte correspondiente al razonamiento.

Teniendo en cuenta esta lección, Feigenbaum junto con otros investigadores de Stanford dieron comienzo al Proyecto de Programación Heurística, PPH, dedicado a determinar el grado con el que la nueva metodología de los **sistemas expertos** podía aplicarse a otras áreas de la actividad humana. El siguiente gran esfuerzo se realizó en el área del diagnóstico médico. Feigenbaum, Buchanan y el doctor Edward Shortliffe diseñaron el programa MYCIN, para el diagnóstico de infecciones sanguíneas. Con 450 reglas aproximadamente, MYCIN era capaz de hacer diagnósticos tan buenos como los de un experto y, desde luego, mejores que los de un médico recién graduado. Se distinguía de DENDRAL en dos aspectos principalmente. En primer lugar, a diferencia de las reglas de DENDRAL, no se contaba con un modelo teórico desde el cual se pudiesen deducir las reglas de MYCIN. Fue necesario obtenerlas a partir de extensas entrevistas con los expertos, quienes las habían obtenido de libros de texto, de otros expertos o de su experiencia directa en casos prácticos. En segundo lugar, las reglas deberían reflejar la incertidumbre inherente al conocimiento médico. MYCIN contaba con un elemento que facilitaba el cálculo de incertidumbre denominado **factores de certeza** (véase el Capítulo 13), que al parecer (en aquella época) correspondía muy bien a la manera como los médicos ponderaban las evidencias al hacer un diagnóstico.

La importancia del conocimiento del dominio se demostró también en el área de la comprensión del lenguaje natural. Aunque el sistema SHRDLU de Winograd para la comprensión del lenguaje natural había suscitado mucho entusiasmo, su dependencia del análisis sintáctico provocó algunos de los mismos problemas que habían aparecido en los trabajos realizados en la traducción automática. Era capaz de resolver los problemas de ambigüedad e identificar los pronombres utilizados, gracias a que se había diseñado especialmente para un área (el mundo de los bloques). Fueron varios los investigadores que, como Eugene Charniak, estudiante de Winograd en el MIT, opinaron que para una sólida comprensión del lenguaje era necesario contar con un conocimiento general sobre el mundo y un método general para usar ese conocimiento.

SISTEMAS EXPERTOS

En Yale, el lingüista transformado en informático Roger Schank reforzó lo anterior al afirmar: «No existe eso que llaman sintaxis», lo que irritó a muchos lingüistas, pero sirvió para iniciar un útil debate. Schank y sus estudiantes diseñaron una serie de programas (Schank y Abelson, 1977; Wilensky, 1978; Schank y Riesbeck, 1981; Dyer, 1983) cuyo objetivo era la comprensión del lenguaje natural. El foco de atención estaba menos en el lenguaje *per se* y más en los problemas vinculados a la representación y razonamiento del conocimiento necesario para la comprensión del lenguaje. Entre los problemas estaba el de la representación de situaciones estereotipo (Cullingford, 1981), la descripción de la organización de la memoria humana (Rieger, 1976; Kolodner, 1983) y la comprensión de planes y objetivos (Wilensky, 1983).

El crecimiento generalizado de aplicaciones para solucionar problemas del mundo real provocó el respectivo aumento en la demanda de esquemas de representación del conocimiento que funcionaran. Se desarrolló una considerable cantidad de lenguajes de representación y razonamiento diferentes. Algunos basados en la lógica, por ejemplo el lenguaje Prolog gozó de mucha aceptación en Europa, aceptación que en Estados Unidos fue para la familia del PLANNER. Otros, siguiendo la noción de **Marcos** de Minsky (1975), se decidieron por un enfoque más estructurado, al recopilar información sobre objetos concretos y tipos de eventos, organizando estos tipos en grandes jerarquías taxonómicas, similares a las biológicas.

MARCOS

La IA se convierte en una industria (desde 1980 hasta el presente)

El primer sistema experto comercial que tuvo éxito, R1, inició su actividad en Digital Equipment Corporation (McDermott, 1982). El programa se utilizaba en la elaboración de pedidos de nuevos sistemas informáticos. En 1986 representaba para la compañía un ahorro estimado de 40 millones de dólares al año. En 1988, el grupo de Inteligencia Artificial de DEC había distribuido ya 40 sistemas expertos, y había más en camino. Du Pont utilizaba ya 100 y estaban en etapa de desarrollo 500 más, lo que le generaba ahorro de diez millones de dólares anuales aproximadamente. Casi todas las compañías importantes de Estados Unidos contaban con su propio grupo de IA, en el que se utilizaban o investigaban sistemas expertos.

En 1981 los japoneses anunciaron el proyecto «Quinta Generación», un plan de diez años para construir computadores inteligentes en los que pudiese ejecutarse Prolog. Como respuesta Estados Unidos constituyó la Microelectronics and Computer Technology Corporation (MCC), consorcio encargado de mantener la competitividad nacional en estas áreas. En ambos casos, la IA formaba parte de un gran proyecto que incluía el diseño de chips y la investigación de la relación hombre máquina. Sin embargo, los componentes de IA generados en el marco de MCC y del proyecto Quinta Generación nunca alcanzaron sus objetivos. En el Reino Unido, el informe Alvey restauró el patrocinio suspendido por el informe Lighthill¹⁵.

¹⁵ Para evitar confusiones, se creó un nuevo campo denominado Sistemas Inteligentes Basados en Conocimiento (IKBS, *Intelligent Knowledge-Based Systems*) ya que la Inteligencia Artificial había sido oficialmente cancelada.

En su conjunto, la industria de la IA creció rápidamente, pasando de unos pocos millones de dólares en 1980 a billones de dólares en 1988. Poco después de este período llegó la época llamada «El Invierno de la IA», que afectó a muchas empresas que no fueron capaces de desarrollar los extravagantes productos prometidos.

Regreso de las redes neuronales (desde 1986 hasta el presente)

Aunque la informática había abandonado de manera general el campo de las redes neuronales a finales de los años 70, el trabajo continuó en otros campos. Físicos como John Hopfield (1982) utilizaron técnicas de la mecánica estadística para analizar las propiedades de almacenamiento y optimización de las redes, tratando colecciones de nodos como colecciones de átomos. Psicólogos como David Rumelhart y Geoff Hinton continuaron con el estudio de modelos de memoria basados en redes neuronales. Como se verá en el Capítulo 20, el impulso más fuerte se produjo a mediados de la década de los 80, cuando por lo menos cuatro grupos distintos reinventaron el algoritmo de aprendizaje de retroalimentación, mencionado por vez primera en 1969 por Bryson y Ho. El algoritmo se aplicó a diversos problemas de aprendizaje en los campos de la informática y la psicología, y la gran difusión que conocieron los resultados obtenidos, publicados en la colección *Parallel Distributed Processing* (Rumelhart y McClelland, 1986), suscitó gran entusiasmo.

CONEXIONISTAS

Aquellos modelos de inteligencia artificial llamados **conexionistas**¹⁶ fueron vistos por algunos como competidores tanto de los modelos simbólicos propuestos por Newell y Simon como de la aproximación lógica de McCarthy entre otros (Smolensky, 1988). Puede parecer obvio que los humanos manipulan símbolos hasta cierto nivel, de hecho, el libro *The Symbolic Species* (1997) de Terrence Deacon sugiere que esta es la *característica que define* a los humanos, pero los conexionistas más ardientes se preguntan si la manipulación de los símbolos desempeña algún papel justificable en determinados modelos de cognición. Este interrogante no ha sido aún clarificado, pero la tendencia actual es que las aproximaciones conexionistas y simbólicas son complementarias y no competidoras.

IA se convierte en una ciencia (desde 1987 hasta el presente)

En los últimos años se ha producido una revolución tanto en el contenido como en la metodología de trabajo en el campo de la inteligencia artificial.¹⁷ Actualmente es más usual el desarrollo sobre teorías ya existentes que proponer teorías totalmente novedo-

¹⁶ Se usa la traducción literal del término *connectionist* por no existir un término equivalente en español (*N. del RT*).

¹⁷ Hay quien ha caracterizado este cambio como la victoria de los pulcros (aquejlos que consideran que las teorías de IA deben basarse rigurosamente en las matemáticas) sobre los desaliñados (aquejlos que después de intentar muchas ideas, escriben algunos programas y después evalúan las que aparentemente funcionan). Ambos enfoques son útiles. Esta tendencia en favor de una mayor pulcritud es señal de que el campo ha alcanzado cierto nivel de estabilidad y madurez. Lo cual no implica que tal estabilidad se puede ver alterada con el surgimiento de otras ideas poco estructuradas.

sas, tomar como base rigurosos teoremas o sólidas evidencias experimentales más que intuición, y demostrar la utilidad de las aplicaciones en el mundo real más que crear ejemplos de juguete.

La IA se fundó en parte en el marco de una rebelión en contra de las limitaciones de los campos existentes como la teoría de control o la estadística, y ahora abarca estos campos. Tal y como indica David McAllester (1998),

En los primeros años de la IA parecía perfectamente posible que las nuevas formas de la computación simbólica, por ejemplo, los marcos y las redes semánticas, hicieran que la mayor parte de la teoría clásica pasara a ser obsoleta. Esto llevó a la IA a una especie de aislamiento, que la separó del resto de las ciencias de la computación. En la actualidad se está abandonando este aislamiento. Existe la creencia de que el aprendizaje automático no se debe separar de la teoría de la información, que el razonamiento incierto no se debe separar de los modelos estocásticos, de que la búsqueda no se debe aislar de la optimización clásica y el control, y de que el razonamiento automático no se debe separar de los métodos formales y del análisis estático.

En términos metodológicos, se puede decir, con rotundidad, que la IA ya forma parte del ámbito de los métodos científicos. Para que se acepten, las hipótesis se deben someter a rigurosos experimentos empíricos, y los resultados deben analizarse estadísticamente para identificar su relevancia (Cohen, 1995). El uso de Internet y el compartir repositorios de datos de prueba y código, ha hecho posible que ahora se puedan contrastar experimentos.

Un buen modelo de la tendencia actual es el campo del reconocimiento del habla. En la década de los 70 se sometió a prueba una gran variedad de arquitecturas y enfoques. Muchos de ellos fueron un tanto *ad hoc* y resultaban frágiles, y fueron probados sólo en unos pocos ejemplos elegidos especialmente. En años recientes, las aproximaciones basadas en los **modelos de Markov ocultos**, MMO, han pasado a dominar el área. Dos son las características de los MMO que tienen relevancia. Primero, se basan en una rigurosa teoría matemática, lo cual ha permitido a los investigadores del lenguaje basarse en los resultados de investigaciones matemáticas hechas en otros campos a lo largo de varias décadas. En segundo lugar, los modelos se han generado mediante un proceso de aprendizaje en grandes *corpus* de datos de lenguaje reales. Lo cual garantiza una funcionalidad robusta, y en sucesivas pruebas ciegas, los MMO han mejorado sus resultados a un ritmo constante. La tecnología del habla y el campo relacionado del reconocimiento de caracteres manuscritos están actualmente en transición hacia una generalizada utilización en aplicaciones industriales y de consumo.

Las redes neuronales también siguen esta tendencia. La mayor parte del trabajo realizado con redes neuronales en la década de los 80 se realizó con la idea de dejar a un lado lo que se podía hacer y de descubrir en qué se diferenciaban las redes neuronales de otras técnicas «tradicionales». La utilización de metodologías mejoradas y marcos teóricos, ha autorizado que este campo alcance un grado de conocimiento que ha permitido que ahora las redes neuronales se puedan comparar con otras técnicas similares de campos como la estadística, el reconocimiento de patrones y el aprendizaje automático, de forma que las técnicas más prometedoras pueden aplicarse a cualquier problema. Como resultado de estos desarrollos, la tecnología denominada **minería de datos** ha generado una nueva y vigorosa industria.

La aparición de *Probabilistic Reasoning in Intelligent Systems* de Judea Pearl (1988) hizo que se aceptara de nuevo la probabilidad y la teoría de la decisión como parte de la IA, como consecuencia del resurgimiento del interés despertado y gracias especialmente al artículo *In Defense of Probability* de Peter Cheeseman (1985). El formalismo de las **redes de Bayes** apareció para facilitar la representación eficiente y el razonamiento riguroso en situaciones en las que se disponía de conocimiento incierto. Este enfoque supera con creces muchos de los problemas de los sistemas de razonamiento probabilístico de las décadas de los 60 y 70; y ahora domina la investigación de la IA en el razonamiento incierto y los sistemas expertos. Esta aproximación facilita el aprendizaje a partir de la experiencia, y combina lo mejor de la IA clásica y las redes neuronales. El trabajo de Judea Pearl (1982a) y de Eric Horvitz y David Heckerman (Horvitz y Heckerman, 1986; Horvitz *et al.*, 1986) sirvió para promover la noción de sistemas expertos *normativos*: es decir, los que actúan racionalmente de acuerdo con las leyes de la teoría de la decisión, sin que intenten imitar las etapas de razonamiento de los expertos humanos. El sistema operativo WindowsTM incluye varios sistemas expertos de diagnóstico normativos para la corrección de problemas. En los Capítulos 13 y 16 se aborda este tema.

Revoluciones similares y suaves se han dado en robótica, visión por computador, y aprendizaje automático. La comprensión mejor de los problemas y de su complejidad, junto a un incremento en la sofisticación de las matemáticas ha facilitado el desarrollo de una agenda de investigación y de métodos más robustos. En cualquier caso, la formalización y especialización ha llevado también a la fragmentación: áreas como la visión y la robótica están cada vez más aislados de la «rama central» de la IA. La concepción unificadora de IA como diseño de agentes racionales puede facilitar la unificación de estos campos diferentes.

Emergencia de los sistemas inteligentes (desde 1995 hasta el presente)

Quizás animados por el progreso en la resolución de subproblemas de IA, los investigadores han comenzado a trabajar de nuevo en el problema del «agente total». El trabajo de Allen Newell, John Laird, y Paul Rosenbloom en SOAR (Newell, 1990; Laird *et al.*, 1987) es el ejemplo mejor conocido de una arquitectura de agente completa. El llamado «movimiento situado» intenta entender la forma de actuar de los agentes inmersos en entornos reales, que disponen de sensores de entradas continuas. Uno de los medios más importantes para los agentes inteligentes es Internet. Los sistemas de IA han llegado a ser tan comunes en aplicaciones desarrolladas para la Web que el sufijo «-bot» se ha introducido en el lenguaje común. Más aún, tecnologías de IA son la base de muchas herramientas para Internet, como por ejemplo motores de búsqueda, sistemas de recomendación, y los sistemas para la construcción de portales Web.

Además de la primera edición de este libro de texto (Russell y Norvig, 1995), otros libros de texto han adoptado recientemente la perspectiva de agentes (Poole *et al.*, 1998; Nilsson, 1998). Una de las conclusiones que se han extraído al tratar de construir agentes completos ha sido que se deberían reorganizar los subcampos aislados de la IA para

que sus resultados se puedan interrelacionar. En particular, ahora se cree mayoritariamente que los sistemas sensoriales (visión, sónar, reconocimiento del habla, etc.) no pueden generar información totalmente fidedigna del medio en el que habitan. Otra segunda consecuencia importante, desde la perspectiva del agente, es que la IA se ha ido acercando a otros campos, como la teoría de control y la economía, que también tratan con agentes.

1.4 El estado del arte

¿Qué es capaz de hacer la IA hoy en día? Responder de manera concisa es difícil porque hay muchas actividades divididas en muchos subcampos. Aquí se presentan unas cuantas aplicaciones; otras aparecerán a lo largo del texto.

Planificación autónoma: a un centenar de millones de millas de la Tierra, el programa de la NASA Agente Remoto se convirtió en el primer programa de planificación autónoma a bordo que controlaba la planificación de las operaciones de una nave espacial desde abordo (Jonsson *et al.*, 2000). El Agente Remoto generaba planes a partir de objetivos generales especificados desde tierra, y monitorizaba las operaciones de la nave espacial según se ejecutaban los planes (detección, diagnóstico y recuperación de problemas según ocurrían).

Juegos: Deep Blue de IBM fue el primer sistema que derrotó a un campeón mundial en una partida de ajedrez cuando superó a Garry Kasparov por un resultado de 3.5 a 2.5 en una partida de exhibición (Goodman y Keene, 1997). Kasparov dijo que había percibido un «nuevo tipo de inteligencia» al otro lado del tablero. La revista *Newsweek* describió la partida como «La partida final». El valor de las acciones de IBM se incrementó en 18 billones de dólares.

Control autónomo: el sistema de visión por computador ALVINN fue entrenado para dirigir un coche de forma que siguiese una línea. Se instaló en una furgoneta controlada por computador en el NAVLAB de UCM y se utilizó para dirigir al vehículo por Estados Unidos. Durante 2.850 millas controló la dirección del vehículo en el 98 por ciento del trayecto. Una persona lo sustituyó en el dos por ciento restante, principalmente en vías de salida. El NAVLAB posee videocámaras que transmiten imágenes de la carretera a ALVINN, que posteriormente calcula la mejor dirección a seguir, basándose en las experiencias acumuladas en los viajes de entrenamiento.

Diagnosis: los programas de diagnóstico médico basados en el análisis probabilista han llegado a alcanzar niveles similares a los de médicos expertos en algunas áreas de la medicina. Heckerman (1991) describe un caso en el que un destacado experto en la patología de los nodos linfáticos se mofó del diagnóstico generado por un programa en un caso especialmente difícil. El creador del programa le sugirió que le preguntase al computador cómo había generado el diagnóstico. La máquina indicó los factores más importantes en los que había basado su decisión y explicó la ligera interacción existente entre varios de los síntomas en este caso. Eventualmente, el experto aceptó el diagnóstico del programa.

Planificación logística: durante la crisis del Golfo Pérsico de 1991, las fuerzas de Estados Unidos desarrollaron la herramienta *Dynamic Analysis and Replanning Tool*

(DART) (Cross y Walker, 1994), para automatizar la planificación y organización logística del transporte. Lo que incluía hasta 50.000 vehículos, carga y personal a la vez, teniendo en cuenta puntos de partida, destinos, rutas y la resolución de conflictos entre otros parámetros. Las técnicas de planificación de IA permitieron que se generara un plan en cuestión de horas que podría haber llevado semanas con otros métodos. La agencia DARPA (*Defense Advanced Research Project Agency*) afirmó que esta aplicación por sí sola había más que amortizado los 30 años de inversión de DARPA en IA.

Robótica: muchos cirujanos utilizan hoy en día asistentes robot en operaciones de microcirugía. HipNav (DiGioia *et al.*, 1996) es un sistema que utiliza técnicas de visión por computador para crear un modelo tridimensional de la anatomía interna del paciente y después utiliza un control robotizado para guiar el implante de prótesis de cadera.

Procesamiento de lenguaje y resolución de problemas: PROVER B (Littman *et al.*, 1999) es un programa informático que resuelve crucigramas mejor que la mayoría de los humanos, utilizando restricciones en programas de relleno de palabras, una gran base de datos de crucigramas, y varias fuentes de información como diccionarios y bases de datos *online*, que incluyen la lista de películas y los actores que intervienen en ellas, entre otras cosas. Por ejemplo, determina que la pista «Historia de Niza» se puede resolver con «ETAGE» ya que su base de datos incluye el par pista/solución «Historia en Francia/ETAGE» y porque reconoce que los patrones «Niza X» y «X en Francia» a menudo tienen la misma solución. El programa no sabe que Niza es una ciudad de Francia, pero es capaz de resolver el puzzle.

Estos son algunos de los ejemplos de sistemas de inteligencia artificial que existen hoy en día. No se trata de magia o ciencia ficción, son más bien ciencia, ingeniería y matemáticas, para los que este libro proporciona una introducción.

1.5 Resumen

En este capítulo se define la IA y se establecen los antecedentes culturales que han servido de base. Algunos de los aspectos más destacables son:

- Cada uno tiene una visión distinta de lo que es la IA. Es importante responder a las dos preguntas siguientes: ¿Está interesado en el razonamiento y el comportamiento? ¿Desea modelar seres humanos o trabajar a partir de un ideal estándar?
- En este libro se adopta el criterio de que la inteligencia tiene que ver principalmente con las **acciones racionales**. Desde un punto de vista ideal, un **agente inteligente** es aquel que emprende la mejor acción posible ante una situación dada. Se estudiará el problema de la construcción de agentes que sean inteligentes en este sentido.
- Los filósofos (desde el año 400 a.C.) facilitaron el poder imaginar la IA, al concebir la idea de que la mente es de alguna manera como una máquina que funciona a partir del conocimiento codificado en un lenguaje interno, y al considerar que el pensamiento servía para seleccionar la acción a llevar a cabo.
- Las matemáticas proporcionaron las herramientas para manipular tanto las aseveraciones de certeza lógicas, como las inciertas de tipo probabilista. Asimismo,

prepararon el terreno para un entendimiento de lo que es el cálculo y el razonamiento con algoritmos.

- Los economistas formalizaron el problema de la toma de decisiones para maximizar los resultados esperados.
- Los psicólogos adoptaron la idea de que los humanos y los animales podían considerarse máquinas de procesamiento de información. Los lingüistas demostraron que el uso del lenguaje se ajusta a ese modelo.
- Los informáticos proporcionaron los artefactos que hicieron posible la aplicación de la IA. Los programas de IA tienden a ser extensos y no podrían funcionar sin los grandes avances en velocidad y memoria aportados por la industria informática.
- La teoría de control se centra en el diseño de dispositivos que actúan de forma óptima con base en la retroalimentación que reciben del entorno en el que están inmersos. Inicialmente, las herramientas matemáticas de la teoría de control eran bastante diferentes a las técnicas que utilizaba la IA, pero ambos campos se están acercando.
- La historia de la IA ha pasado por ciclos de éxito, injustificado optimismo y consecuente desaparición de entusiasmo y apoyos financieros. También ha habido ciclos caracterizados por la introducción de enfoques nuevos y creativos y de un perfeccionamiento sistemático de los mejores.
- La IA ha avanzado más rápidamente en la década pasada debido al mayor uso del método científico en la experimentación y comparación de propuestas.
- Los avances recientes logrados en el entendimiento de las bases teóricas de la inteligencia han ido aparejados con las mejoras realizadas en la optimización de los sistemas reales. Los subcampos de la IA se han integrado más y la IA ha encontrado elementos comunes con otras disciplinas.



NOTAS BIBLIOGRÁFICAS E HISTÓRICAS

El estatus metodológico de la inteligencia artificial se ha investigado en *The Sciences of the Artificial*, escrito por Herb Simon (1981), en el cual se analizan áreas de investigación interesadas en el desarrollo de artefactos complejos. Explica cómo la IA se puede ver como ciencia y matemática. Cohen (1995) proporciona una visión de la metodología experimental en el marco de la IA. Ford y Hayes (1995) presentan una revisión crítica de la utilidad de la Prueba de Turing.

Artificial Intelligence: The Very Idea, de John Haugeland (1985), muestra una versión amena de los problemas prácticos y filosóficos de la IA. La ciencia cognitiva está bien descrita en varios textos recientes (Johnson-Laird, 1988; Stillings *et al.*, 1995; Thagard, 1996) y en la *Encyclopedia of the Cognitive Sciences* (Wilson y Keil, 1999). Baker (1989) cubre la parte sintáctica de la lingüística moderna, y Chierchia y McConnell-Ginet (1990) la semántica. Jurafsky y Martin (2000) revisan la lingüística computacional.

Los primeros trabajos en el campo de la IA se citan en *Computers and Thought* (1963), de Feigenbaum y Feldman, en *Semantic Information Processing* de Minsky, y en la serie *Machine Intelligence*, editada por Donald Michie. Webber y Nilsson (1981)

y Luger (1995) han recogido una nutrida cantidad de artículos influyentes. Los primeros artículos sobre redes neuronales están reunidos en *Neurocomputing* (Anderson y Rosenfeld, 1988). La *Encyclopedia of AI* (Shapiro, 1992) contiene artículos de investigación sobre prácticamente todos los temas de IA. Estos artículos son muy útiles para iniciarse en las diversas áreas presentes en la literatura científica.

El trabajo más reciente se encuentra en las actas de las mayores conferencias de IA: la *International Joint Conference on AI* (IJCAI), de carácter bianual, la *European Conference on AI* (ECAI), de carácter bianual, y la *National Conference on AI*, conocida normalmente como AAAI por la organización que la patrocina. Las revistas científicas que presentan aspectos generales de la IA más importantes son *Artificial Intelligence*, *Computational Intelligence*, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *IEEE Intelligent Systems*, y la revista electrónica *Journal of Artificial Intelligence Research*. Hay también numerosas revistas y conferencias especializadas en áreas concretas, que se mencionarán en los capítulos apropiados. La asociaciones profesionales de IA más importantes son la American Association for Artificial Intelligence (AAAI), la ACM Special Interest Group in Artificial Intelligence (SIGART), y la Society for Artificial Intelligence and Simulation of Behaviour (AISB). La revista *AI Magazine* de AAAI contiene muchos artículos de interés general y manuales, y su página Web, aaai.org contiene noticias e información de referencia.



EJERCICIOS

El propósito de los siguientes ejercicios es estimular la discusión, y algunos de ellos se podrían utilizar como proyectos. Alternativamente, se podría hacer un esfuerzo inicial para solucionarlos ahora, de forma que una vez se haya leído todo el libro se puedan revisar estos primeros intentos.



1.1 Defina con sus propias palabras: (a) inteligencia, (b) inteligencia artificial, (c) agente.

1.2 Lea el artículo original de Turing sobre IA (Turing, 1950). En él se comentan algunas objeciones potenciales a su propuesta y a su prueba de inteligencia. ¿Cuáles de estas objeciones tiene todavía validez? ¿Son válidas sus refutaciones? ¿Se le ocurren nuevas objeciones a esta propuesta teniendo en cuenta los desarrollos realizados desde que se escribió el artículo? En el artículo, Turing predijo que para el año 2000 sería probable que un computador tuviera un 30 por ciento de posibilidades de superar una Prueba de Turing dirigida por un evaluador inexperto con una duración de cinco minutos. ¿Considera razonable lo anterior en el mundo actual? ¿Y en los próximos 50 años?



1.3 Todos los años se otorga el premio Loebner al programa que lo hace mejor en una Prueba de Turing concreta. Investigue y haga un informe sobre el último ganador del premio Loebner. ¿Qué técnica utiliza? ¿Cómo ha hecho que progrese la investigación en el campo de la IA?

1.4 Hay clases de problemas bien conocidos que son intratables para los computadores, y otras clases sobre los cuales un computador no pueda tomar una decisión. ¿Quiere esto decir que es imposible lograr la IA?

1.5 Supóngase que se extiende ANALOGY, el programa de Evans, como para alcanzar una puntuación de 200 en una prueba normal de cociente de inteligencia. ¿Quiere decir lo anterior que se ha creado un programa más inteligente que un ser humano? Explíquese.

1.6 ¿Cómo puede la introspección (revisión de los pensamientos íntimos) ser inexacta? ¿Se puede estar equivocado sobre lo que se cree? Discútase.



1.7 Consulte en la literatura existente sobre la IA si alguna de las siguientes tareas se puede efectuar con computadores:

- a)** Jugar una partida de tenis de mesa (ping-pong) decentemente.
- b)** Conducir un coche en el centro del Cairo.
- c)** Comprar comestibles para una semana en el mercado.
- d)** Comprar comestibles para una semana en la web.
- e)** Jugar una partida de *bridge* decentemente a nivel de competición.
- f)** Descubrir y demostrar nuevos teoremas matemáticos.
- g)** Escribir intencionadamente una historia divertida.
- h)** Ofrecer asesoría legal competente en un área determinada.
- i)** Traducir inglés hablado al sueco hablado en tiempo real.
- j)** Realizar una operación de cirugía compleja.

En el caso de las tareas que no sean factibles de realizar en la actualidad, trate de describir cuáles son las dificultades y calcule para cuándo se podrán superar.

1.8 Algunos autores afirman que la percepción y las habilidades motoras son la parte más importante de la inteligencia y que las capacidades de «alto nivel» son más bien parásitas (simples añadidos a las capacidades básicas). Es un hecho que la mayor parte de la evolución y que la mayor parte del cerebro se han concentrado en la percepción y las habilidades motoras, en tanto la IA ha descubierto que tareas como juegos e inferencia lógica resultan más sencillas, en muchos sentidos, que percibir y actuar en el mundo real. ¿Consideraría usted que ha sido un error la concentración tradicional de la IA en las capacidades cognitivas de alto nivel?

1.9 ¿Por qué la evolución tiende a generar sistemas que actúan racionalmente? ¿Qué objetivos deben intentar alcanzar estos sistemas?

1.10 ¿Son racionales las acciones reflejas (como retirar la mano de una estufa caliente)? ¿Son inteligentes?

1.11 «En realidad los computadores no son inteligentes, hacen solamente lo que le dicen los programadores». ¿Es cierta la última aseveración, e implica a la primera?

1.12 «En realidad los animales no son inteligentes, hacen solamente lo que le dicen sus genes». ¿Es cierta la última aseveración, e implica a la primera?

1.13 «En realidad los animales, los humanos y los computadores no pueden ser inteligentes, ellos sólo hacen lo que los átomos que los forman les dictan siguiendo las leyes de la física». ¿Es cierta la última aseveración, e implica a la primera?