Racionalidad Computacional

"Racionalidad Computacional" de Luis Pineda es una obra de gran profundidad y alcance. Su punto de partida es el libre albedrío y la racionalidad. A lo largo de los capítulos pasa revista a temas centrales del desarrollo de la inteligencia artificial, las ciencias cognitivas, la teoría de la comunicación y la teoría de la computación que le permiten postular un nuevo modelo de computación estocástico y entrópico que denomina Computación Relacional Indeterminada (CRI). Se trata de una propuesta muy ambiciosa que no se queda en consideraciones teóricas, sino que propone un modelo de arquitectura de máquina centrada en una memoria asociativa que trabaja junto con módulos de percepción y acción. La CRI se plantea como una generalización de la noción de computación en donde la computación determinista de la máquina de Turing es un caso singular. Este planteamiento recuerda lo sucedido en la física de inicios del siglo XX con los planteamientos de Boltzmann y Gibbs que crearon la mecánica estadística como generalización de la mecánica de Newton al introducir planteamientos estadísticos y entrópicos. (Christian Lemaitre)

¿Qué significa ser racional? Es una pregunta crítica tanto para los seres vivos, como para las máquinas, que Luis aborda en este libro desde una perspectiva novedosa, desde el punto de vista computacional. Para ello, nos lleva a través de un viaje muy interesante sobre diversos aspectos claves de la computación, como son la Máquina de Turing, la racionalidad limitada y su relación a la teoría de juegos, la toma de decisiones y su relación con la entropía, la memoria asociativa, las arquitecturas de la cognición computacional y natural, entre otros. En este camino, nos invita a reflexionar sobre las diferentes concepciones de la racionalidad, así como las diferencias entre las capacidades de la inteligencia humana y la inteligencia artificial, y sobre las limitaciones de esta última, un tema sin duda complejo y controversial. Los invito a recorrer este muy interesante viaje por el mundo de la computación, seguro van a disfrutarlo como yo (Enrique Sucar).

ISBN 978-607-98941-3-9

Racionalidad computacional | Academia Mexicana de Computación

amexcomp

RACIONALIDAD COMPUTACIONAL

Luis Alberto Pineda Cortés

amexcomp

Academia Mexicana de Computación A.C.

Racionalidad Computacional

Luis Alberto Pineda Cortés

Academia Mexicana de Computación A.C.

.

Autor: Luis Alberto Pineda Cortés.

Primera edición: 2021
Academia Mexicana de Computación

ISBN: 978-607-98941-3-9
Diseño de portada: Mario Alberto Vélez Sánchez.

Impreso en México.
Printed in Mexico.

Racionalidad Computacional

Luis Alberto Pineda Cortés

Este libro se imprimió con el apoyo de la DGAPA-UNAM
Proyecto PAPIIT UNAM IN112819

Para Nydia,
por las decisiones que hemos tomado
a lo largo del camino

ACADEMIA MEXICANA DE COMPUTACIÓN A. C.

Dr. Carlos Artemio Coello Coello
Presidente

Dr. Eduardo F. Morales Manzanares
Vicepresidente

Dr. Efrén Mezura Montes
Tesorero

Dra. María del Pilar Gómez Gil
Secretaria

Dr. Hugo Terashima-Marín
Secretario

Dra. Marcela Quiroz Castellanos
Vocal

Agradecimientos

En este texto presento una investigación acerca del pensamiento y la conducta racional desde la perspectiva de las máquinas computacionales y la Inteligencia Artificial. La inspiración inicial provino de mis investigaciones acerca de la formalización de las representaciones diagramáticas y del descubrimiento y prueba de teoremas basados en diagramas, como el Teorema de Pitágoras, donde el razonamiento conceptual se entrelaza estrechamente con la percepción. Agradezco a John Lee, Ewan Klein, Henk Zeevat, Aart Bijl, Rafael Morales, Gaby Garza, B. Chandrasekaran e Ivan Bratko por acompañarme en el desarrollo de estas ideas como interlocutores, colegas y amigos.

La reflexión se enriqueció con la experiencia del Proyecto Golem para la construcción de robots de servicio, desarrollado en el IIMAS-UNAM. Agradezco a Arturo Rodríguez, Noé Hernández, Iván Torres, Mauricio Reyes y Ricardo Cruz por la implementación del ciclo de inferencia de la vida cotidiana y de la base de conocimiento no-monotónica en el robot Golem-III que se presentan en los capítulos 3 y 4; y a los demás colegas y estudiantes que participaron en este proyecto a lo largo de los años; así como el apoyo continuo del CONACyT y del PAPIIT-UNAM.

A Rafael Morales por su apoyo en el diseño e implementación de los experimentos de la memoria asociativa que se presenta en el Capítulo 7, así como por la generación de las gráficas con los resultados; a Gibrán Fuentes por el diálogo continuo a lo largo de varios años y por su apoyo en el diseño de los experimentos; y a Iván Torres y Raúl Peralta por la implementación de los experimentos preliminares de reconocimiento. A Thomas Eiter y Magdalena Ortiz quienes fueron los anfitriones de mi sabático en la Universidad Técnica de Viena en 2011 durante el cual concebí las ideas originales del sistema de memoria y la entro-

pía computacional. A Tom Froese por las estimulantes discusiones acerca de la cognición; a Carlos Velarde por sus comentarios y sugerencias acerca de la sección de tablas y computabilidad; a Ignacio Peñaloza por sus sugerencias acerca del espacio de la acción; a Mike Posner por sus comentarios acerca del modo de computación y la memoria asociativa; y a Ed Hovy por las muy intensas discusiones acerca de la computación por tablas y el modo de computación, y por sus palabras de aliento en momentos muy difíciles. A los colegas de la comunidad mexicana de computación que me han permitido compartir muchas de las ideas expresadas en este texto en diversos foros de la UNAM y en varias otras instituciones nacionales en los últimos años. Estas pláticas siempre han sido un reto y una experiencia de crecimiento.

A mi esposa Nydia y a mi hermano Juan Carlos por su cuidadosa lectura y por sus comentarios y sugerencias que fueron muy valiosos para enriquecer el texto; a mi hermano Daniel, a Gerardo León, a Pepe Franco y a Philippe Ollé-Laprune, cuyas impresiones fueron también de gran utilidad; a mis estudiantes Dennis Mendoza y Eduardo Acuña por su lectura y comentarios del primer borrador y a Noé Hernández y Homero Buenrostro Trujillo por su lectura de la versión final.

Agradezco de manera muy especial a Enrique Sucar y a Christian Lemaitre por la rigurosa revisión y comentarios del primer borrador de este libro, que fueron muy útiles para redefinir su orientación y alcance, así como por la lectura y comentarios de la versión final; a la Academia Mexicana de Computación por permitir que esta publicación lleve su sello editorial; y al proyecto PAPIIT-UNAM IN112819 que hizo posible su impresión y difusión.

Prólogo

La racionalidad ha sido tema de estudio y reflexión desde el origen del pensamiento humano. Se relaciona estrechamente con el libre albedrío. Ser racional es tomar buenas decisiones y disfrutar sus consecuencias. De forma recíproca, tomar malas decisiones o actuar sin prever lo que pueda acontecer es irracional. Esta facultad está muy desarrollada en la especie humana pero también la disfrutan una amplia variedad de especies animales y el término se aplica metafóricamente a la naturaleza. Asimismo, la tecnología computacional y en especial la Inteligencia Artificial habilitan a entes inanimados a tomar decisiones, cuyas consecuencias pueden ser de gran impacto.

La racionalidad computacional se refiere al pensamiento y a la toma de decisiones que tienen como agente causal y esencial una máquina computacional. El modelo teórico general de las computadoras digitales lo presentó originalmente Alan Turing (1912-1954) en 1936, y muy pronto se denominó como *La Máquina de Turing*. En ausencia a una noción de computación diferente a la propuesta por Turing, la racionalidad computacional es aquella que se hace posible gracias a dicha máquina.

El estudio de esta facultad se ha abordado tradicionalmente desde tres perspectivas de gran aliento: i) la doctrina que sostiene que la racionalidad se debe a mecanismos generales basados en primeros principios; ii) la que sostiene que

ésta se debe a una gran variedad de mecanismos específicos producto de la evolución que permiten a los agentes interactuar con el entorno de manera muy efectiva; y iii) la doctrina dual que sostiene que el razonamiento emplea ambas estrategias.

Los mecanismos paradigmáticos del primer tipo incluyen a la lógica, que sostiene que el conocimiento tiene un carácter proposicional y declarativo, y se enfoca al estudio de los argumentos válidos; la lógica se puede ver también como la labor de análisis que permite investigar los efectos a partir de las causas. Un segundo mecanismo de carácter general es el razonamiento Bayesiano; éste se enfoca a la inferencia inversa orientada a determinar las causas a partir de los efectos, donde la noción lógica de validez se sustituye por la noción de probabilidad, en el sentido de qué tan probable es que una observación se deba a cierta causa; este tipo de inferencia está estrechamente ligada al aprendizaje y se orienta a la síntesis del conocimiento. Un tercer mecanismo general, mucho más reciente, al menos de forma explícita, es la teoría de juegos, el cual se enfoca a ponderar las decisiones, y lo racional se entiende como escoger la decisión más valiosa entre un conjunto de decisiones potenciales.

El estudio de los mecanismos particulares se ha abordado desde diversas disciplinas. En la Inteligencia Artificial se refleja en el desarrollo de las llamadas representaciones procedurales, donde el conocimiento y las habilidades perceptuales se expresan como programas de cómputo directamente. Esta perspectiva también tiene una gran tradición en la psicología cognitiva enfocada al estudio de las heurísticas y los sesgos conductuales. Las heurísticas son estrategias de conducta con un alto componente genético aunque se pueden aprender por analogía y aplicarse a situaciones novedosas; sin embargo, la adaptación puede ser imprecisa o inapropiada y conducir a conductas irracionales. Una forma adi-

cional y más reciente en esta concepción es la racionalidad ecológica que se basa también en heurísticas pero enfatiza la interacción entre el agente y el entorno de manera muy estrecha.

La tercera vía sostiene que el agente actúa por mecanismos específicos de manera cotidiana, pero si tiene el tiempo y la energía puede recurrir a mecanismos generales para mejorar sus decisiones.

El problema es cómo abordar esta problemática de manera integral. Aquí se parte de la Inteligencia Artificial y los mecanismos computacionales. La pregunta es en qué medida la computación ofrece una explicación del fenómeno de la racionalidad y de la inteligencia. Hay procesos racionales que se pueden modelar claramente con las máquinas computacionales, como los programas que juegan juegos racionales, pero la racionalidad se relaciona también con la interpretación, la experiencia y la consciencia, ante las cuales la computación, al menos en la teoría estándar, tiene poco que decir. En este texto se ilustra la gama de posturas y posibilidades haciendo un recorrido en tres partes principales.

La primera presenta de manera muy sucinta el programa original de la Inteligencia Artificial como lo propuso el mismo Turing en 1950 y como se dio en su primer impulso, principalmente con la teoría de la Racionalidad Limitada, en los sesentas y setentas del siglo pasado. El método se centró en la llamada manipulación simbólica; este enfoque tomó elementos de la lógica y de la teoría de juegos y se aplicó a la prueba de teoremas lógicos y matemáticos, al diseño de métodos generales de solución de problemas, y a la construcción de programas de cómputo capaces de jugar juegos racionales, especialmente el ajedrez.

Este programa tuvo un éxito inicial considerable y generó grandes expectativas, pero se criticó desde los primeros años ya que es puramente mentalista, incluso etéreo, y el agente no se liga al mundo a través de la percepción y la acción.

Ante esta limitación hubo varias respuestas como los sistemas conexionistas y las redes neuronales artificiales, los sistemas embebidos, los modelos causales Bayesianos, la computación corpórea (*embodied cognition*) e incluso el enactivismo, cada una de las cuales ha desarrollado escuelas de pensamiento con gran impacto científico y tecnológico. Sin embargo, en estos enfoques el pensamiento dejó de ser el centro de atención y pasó a ser un módulo marginal e incluso ausente en la cognición.

En la segunda parte del libro presento mi propia respuesta a este dilema. Mi punto de partida es el libre albedrío. Aquí propongo la conjetura de que hay una relación entre la comunicación, la toma de decisiones y el cambio de conducta. El propósito de la comunicación es cambiar, al menos potencialmente, las creencias, los deseos, los sentimientos, las intenciones y las acciones que los agentes llevarían a cabo sin la nueva información. La comunicación es una fuerza en el plano de los contenidos análoga a las fuerzas físicas en el plano material. De la misma forma que los cuerpos se mantienen en un estado inercial a menos que haya una fuerza que los desvíe de su curso, los agentes que gozamos del libre albedrío hacemos las cosas por la inercia de nuestra cognición y del entorno social a menos que la información novedosa nos motive a cambiar, y para este efecto hay que tener la opción de decidir.

Sin embargo, para que haya cambios de conducta se requiere que haya algo de indeterminación. Sin ésta la toma de decisiones es una ilusión. En los entornos de comunicación la indeterminación se mide con la entropía. Consecuentemente, debe haber un nivel de entropía en el entorno físico y en el social. Si la entropía es nula o muy baja todo está determinado y no se puede cambiar; de manera recíproca, si la entropía es muy alta el entorno es caótico y las decisiones no se pueden llevar a cabo; pero hay un nivel de entropía moderado en que las

decisiones se pueden traducir en acciones que cambien el entorno de manera productiva. En este texto presento y discuto esta conjetura a la que me refiero como *La productividad Potencial de las Decisiones*; para que ésta tenga un valor adecuado u óptimo debe haber un buen *compromiso de la entropía*.

Por su parte, la mente del agente que toma las decisiones debe también tener un grado de indeterminación. Es decir, la máquina computacional que es causal y esencial a la conducta debe ser indeterminada. El problema es que las computadoras digitales estándar están predeterminadas por necesidad. Turing mismo comparó las predicciones de la Máquina con el determinismo de Laplace. Para enfrentar este dilema se hace aquí un análisis del determinismo de la computación tradicional y propongo un modo más general de computación con carácter estocástico al que denomino *Computación Relacional Indeterminada*.

A diferencia de la Máquina de Turing, que se orienta al cómputo algorítmico, es fundamentalmente serial, enfrenta los problemas de complejidad, no se puede predecir si una computación terminará eventualmente y no es entrópica, la Computación Relacional Indeterminada se orienta a la memoria, utiliza algoritmos mínimos que siempre terminan, es masivamente paralela y tiene una entropía implícita, a la que llamo *entropía computacional*.

Esta forma de computación se utiliza para modelar una memoria asociativa, que es a su vez declarativa, distribuida y constructiva, y se opone a las memorias simbólicas de las computadoras digitales estándar, que no son distribuidas ni asociativas y más que constructivas son reproductivas; y también a las "memorias" creadas con redes neuronales que no son declarativas, por lo que no pueden registrar recuerdos para después retribuirlos y, al carecer de esta propiedad, no son memorias propiamente. La capacidad de la memoria asociativa depende del

nivel de entropía, y el agente computacional requiere adoptar el compromiso de la entropía para ser funcional.

El sistema de memoria asociativa incluye a la percepción y a la acción. Los eventos que son motivo del recuerdo ingresan a la memoria vía las modalidades de la percepción; la información concreta que se sensa e interpreta se mapea a una representación abstracta que a su vez se almacena de forma distribuida en la memoria; de manera recíproca, la información que se retribuye de la memoria se mapea a una representación concreta en alguna modalidad de la acción. El canal percepción, memoria y acción constituye la columna vertebral de la cognición.

Esta discusión concluye con una reflexión acerca del impacto de la Computación Relacional Indeterminada en la teoría estándar de la computación así como en el cómputo que, hipotéticamente, se lleva a cabo por los seres humanos y por los animales no humanos que cuentan con un sistema nervioso y un cerebro suficientemente desarrollados, al que nos referimos como *Computación Natural*. Para este efecto se presenta la noción de *El Modo de Computación*. Éste es el método artificial o el fenómeno natural que lleva a cabo un proceso físico en el plano material, pero que recibe una interpretación y adquiere un significado en el plano mental. El modo de computación y la interpretación siempre vienen juntos: no hay computación sin interpretación y viceversa, o más directamente, computar e interpretar son dos caras del mismo fenómeno.

En particular, si los procesos de la mente son procesos computacionales la computación natural debe tener un modo de computación, todavía no identificado. Sin éste la computación natural es simplemente una metáfora. La computación artificial es un invento humano, pero la computación natural, si existe, debe haber aparecido muy temprano en la naturaleza y con mucha antelación a la computación artificial. La noción del modo de computación sugiere que la

computación natural empieza cuando aparece la interpretación y, consecuentemente, cuando aparecen la experiencia y la consciencia. Un corolario de esta noción es que si no hay seres humanos o animales no humanos que hagan interpretaciones, las computadoras no se distinguen de la maquinaria ordinaria en ninguna dimensión significativa. Esta distinción se aborda en extenso en el Capítulo 8.

En la tercera parte del libro se exploran las consecuencias de incluir a la Computación Relacional Indeterminada y a la entropía en la arquitectura cognitiva. Desde esta perspectiva la arquitectura cognitiva es una extensión de la arquitectura de la memoria asociativa. Ésta hace posible la percepción, el pensamiento y la conducta intencional. Surge una noción de interpretación, de pensamiento, de acción, y un nuevo principio de racionalidad con una orientación Bayesiana: se es racional en la medida en que las hipótesis de interpretación que se realizan a través de la percepción son adecuadas al entorno; en la medida en que las decisiones que se toman reflejan las necesidades y deseos del agente; en la medida en que las acciones son consistentes con las intenciones; en la medida que éstas tienen los efectos esperados; y en el grado en que la entropía y la productividad potencial de las decisiones proveen el espacio para que las decisiones se puedan llevar a cabo en el entorno. La percepción, el diagnóstico, la toma de decisión, la planeación y la acción se realizan siempre bajo supuestos hipotéticos, frecuentemente inconscientes, que dependen de la evidencia y el conocimiento previo.

Las decisiones racionales están asociadas a las acciones racionales. Las buenas decisiones se traducen en acciones que tienden a mejorar las condiciones de vida y las malas tienden a empeorarlas. Más que oposiciones absolutas, qué tan racionales son las decisiones y las acciones depende de qué tan buenas son sus consecuencias. Lo irracional, por su parte, se da cuando la decisión y la acción son

dañinas. Pero el problema para el agente es que no sabe de antemano cuáles serán las consecuencias de sus decisiones y sus acciones. Si lo supiera, las decisiones estarían predeterminadas y las decisiones serían meras ilusiones. Por supuesto, el agente racional se informa y utiliza su experiencia, y sus expectativas pueden ser fundadas, pero no hay realmente nadie que evalúe la acción de antemano y el único juez es la naturaleza misma, que es entrópica. Como en la evolución, lo racional prevalece y lo irracional se extingue. Por supuesto, "la valoración" se da en la perspectiva de corto, mediano y largo plazo, en las dimensiones del tiempo y del espacio, y también en lo social, que se centra en el individuo pero se amplia a lo humano y al espectro de la vida. En este sentido lo más racional es el altruismo y la preservación de la naturaleza.

Esta forma del principio Bayesiano se implementa con algoritmos mínimos y/o con modos de computación no algorítmicos. La noción de algoritmo involucra el uso de representaciones externas, conceptos matemáticos y sistemas métricos, los cuales son constructos culturales e históricos muy posteriores a la aparición de la computación en la naturaleza. Los algoritmos son explicaciones a posteriori pero no causales a la conducta, de la misma forma que explicar cómo se maneja una bicicleta no es causal a andar en bicicleta. El cerebro/mente hace otra cosa, directa y eficiente, pero todavía desconocida. Elucidarla es descubrir el modo o los modos de la computación natural. Desde la perspectiva computacional la pregunta no es cuál es el modelo de la racionalidad que se puede computar con la Máquina de Turing, sino más bien cuál es la racionalidad que tiene como causa una u otra máquina.

La ciencia cognitiva y la cultura popular consideran que el cerebro es una máquina computacional. De acuerdo con esta metáfora el cerebro tiene las características de la Máquina de Turing y la conducta se debe a la ejecución de

algoritmos. Sin embargo, la Máquina de Turing no es entrópica –o su entropía es cero. Esto no ocurre en la naturaleza. Un organismo con entropía cero es inerte y está sujeto pasivamente al vaivén de las fuerzas materiales.

La entropía del cerebro se ha estudiado recientemente en las neurociencias y se sostiene que éste es una máquina entrópica. Aquí sugiero que la Computación Relacional Indeterminada es un modelo más apropiado para caracterizar a la computación natural, y que efectivamente, el cerebro es una máquina entrópica que se conforma al compromiso de la entropía, y propongo algunas hipótesis que se podrían investigar de manera empírica en la psicología evolutiva, la sociología, las neurociencias y la psicología cognitiva.

Esta discusión conlleva a una conjetura aún más intrigante: que la mente evolucionó a partir de la comunicación. Si el agente no se comunica su interacción con el entorno se da sólo en el plano material. Ésta es la condición de lo inanimado. La comunicación da lugar al plano de los contenidos, primero en el entorno y luego interiorizados en la mente. Conforme aumenta la entropía se adquiere la flexibilidad de enfrentar entornos más complejos y variables con el concurso del otro. Con la entropía aparece la interpretación, primero como la experiencia y en una fase más avanzada como consciencia. Pero el incremento de la entropía tiene un límite ya que la comunicación sin cambios de conducta productivos no se fomenta o estimula en la naturaleza. La racionalidad refleja en última instancia qué tan importante es la comunicación y la interpretación para el individuo y la especie.

Luis A. Pineda

Febrero de 2021

Índice general

Capítulo 1

Racionalidad y Toma de Decisiones

La racionalidad es la facultad de anticipar y evaluar las consecuencias de lo que se dice o se hace, de tomar decisiones y de realizar acciones para alcanzar los estados deseados del mundo. Dichas conductas son *racionales*. La acción racional se opone a las conductas que no son consistentes con los intereses de quienes las llevan a cabo, que son *irracionales*. No hay un juicio o un juez que dictamine qué tan racional es la acción más que sus consecuencias para el propio agente y su entorno. Como en la evolución, acciones racionales tenderán a beneficiarlo y las irracionales se traducirán en empeorar sus condiciones de vida, y en última instancia a perecer como individuo y a extinguirse como especie.

La presente discusión se centra en la racionalidad humana, pero siempre teniendo en cuenta que ésta es simplemente el extremo más acabado de una continuidad en el espectro de la vida. Incluso los organismos más básicos realizan acciones para lograr efectos favorables y desde esta perspectiva son "racionales"; y si realizan acciones que atentan contra su integridad son "irracionales". Este continuo se puede abstraer para efectos de análisis en estadios discretos, que corresponden a niveles de desarrollo de la especie o individuos, donde los entes

más sencillos o primitivos se limitan a responder de manera reactiva ante el entorno, y el nivel más acabado corresponde a la racionalidad humana, con toda su riqueza y variedad de acciones. El nivel de racionalidad de los mecanismos artificiales se puede conceptualizar de manera análoga.

En este texto se aborda la racionalidad desde una perspectiva computacional y la visión que esta metáfora ofrece a la racionalidad natural. Se trata de ver en qué grado las máquinas computacionales pueden ser racionales y que aspectos del fenómeno trascienden a esta metáfora científica y tecnológica.

El estudio de la racionalidad computacional se ha abordado en el contexto de la Inteligencia Artificial desde el inicio de esta disciplina. Éste se ha centrado en la síntesis y el análisis de la toma de decisiones y el entorno en que se lleva a cabo. Posiblemente el esfuerzo más explícito en esta línea de investigación ha sido el desarrollo de programas de cómputo que juegan juegos racionales, en particular el ajedrez, considerado en el mundo occidental como el juego racional por excelencia.

Los juegos se definen en relación a un entorno y un conjunto de reglas, incluyendo los turnos en que cada jugador "hace una movida" que contribuye a que gane, empate o pierda. En el ajedrez el entorno consiste del tablero y las piezas, y las reglas definen los movimientos posibles. Cada movida se hace en un estado del juego y la calidad del jugador consiste en analizar las consecuencias de sus movidas, cómo puede responder el adversario y qué tanto puede adelantar la situación para lograr una posición ventajosa. Cada jugador tiene conocimiento completo del estado del juego y en principio puede analizar las consecuencias de todas las movidas de los demás jugadores. El propósito es por supuesto ganar y la movida es la culminación de un proceso de toma de decisión o la manifestación de una decisión.

La calidad del jugador depende de su capacidad de análisis, de su memoria de trabajo, del control de la atención y de su capacidad para la corrección de errores, y los requerimientos de cálculo pueden ser muy significativos. Esta demanda se ilustra claramente en la leyenda del tablero y los granos de trigo. De acuerdo con ésta Sissa inventó el ajedrez hace mucho tiempo en una provincia de la India. El rey, llamado Sheram, estaba muy triste y deprimido porque había perdido a su hijo en una guerra. Un día Sissa se presentó en el palacio y le enseñó a Sheram el juego para mitigar sus penas. Al rey le gustó tanto que le ofreció recompensarlo con lo que él deseara. Sissa lo meditó como si estuviera pensando una movida y le pidió al rey que le diera un grano de trigo por el primer cuadrado del tablero; dos por el segundo; cuatro por el tercero; y así hasta agotar los 64 cuadros; la cantidad de trigo solicitada fue de $\sum_{i=0}^{63} 2^i = 18, 446, 744, 073, 709, 551, 615$ granos. Se dice que con ésta se llenaría un silo de 10 m^2 de base que se extendiera desde la superficie de la tierra hasta el sol. No sabemos el castigo que recibió Sissa por su osadía, pero sí que no pudo cobrar su premio y que Sheram no pudo pagar su deuda. Esta cifra está en el orden de las que hay que analizar en el ajedrez para cubrir 32 turnos hacia adelante, considerando que en cada posición hubiera sólo dos 2 movidas posibles.

Sin embargo, la racionalidad no se reduce a la capacidad de análisis. En escenarios más complejos la toma de decisiones se hace con conocimiento incompleto e incertidumbre. Por ejemplo, los juegos de azar tienen también un entorno y un conjunto de reglas bien definidas, y los jugadores tienen turnos en los que hacen movidas y toman decisiones, pero se tiene conocimiento incompleto o nulo de las cartas o fichas del adversario y hay un elemento aleatorio, como tirar los dados, revolver las cartas o “hacer la sopa”.

Los juegos se pueden ver como una metáfora de la postura egocéntrica del agente, que ve al mundo y/o a otros agentes como adversarios que hay que superar, para lo cual es necesario tomar decisiones que le permitan maximizar sus beneficios de manera individual. Sin embargo, se pueden plantear también escenarios cooperativos en los cuales las decisiones tienen por objetivo mejorar la ganancia o utilidad no sólo del agente que las toma sino también de otros agentes y en el corto, mediano y largo plazo. Las decisiones desde esta perspectiva responden a sus valores e implican una lógica afectiva donde el interés colectivo prevalece por encima del individual y conllevan a un nivel superior de racionalidad.

El estudio de la toma de decisiones y consecuentemente de la racionalidad debe considerar también las emociones y la afectividad. El jugador debe controlarlas y ser sensible a las del adversario y cómo éstas inciden en su juicio. Esta dimensión incluye también a los intereses, las preferencias, los valores y la voluntad, y afecta a las decisiones en un nivel más profundo que la capacidad de análisis. Antes de querer ganar un juego de ajedrez o un juego de póquer hay que tener el interés de jugar. Hacerlo con ánimo o desánimo tiene un impacto mayor en la calidad de las decisiones y el resultado del juego. La calidad del jugador depende también de la habilidad de imaginar las creencias e intenciones del adversario, junto con su capacidad de cálculo y memoria, y sus estados afectivos. Aunque es concebible incluir en el análisis la afectividad y los estados mentales de los adversarios, al menos desde el punto de vista informacional, esta empresa está aún en una etapa muy temprana.

A pesar de todas estas limitaciones, los modelos computacionales de racionalidad, incluyendo los juegos de azar, son como los experimentos biológicos en un tubo de ensayo: nos presentan la toma de decisiones en un mundo ideal

pero dejan fuera muchos de los fenómenos que se pueden encontrar en la vida real. Sin embargo, son experimentos muy útiles que permiten investigar aspectos cuantitativos y objetivos del fenómeno, que además tienen o pueden tener aplicaciones muy productivas de la tecnología computacional.

1.1. Perspectivas de la Racionalidad

La racionalidad se ha estudiado desde los orígenes del pensamiento humano. Las teorías de racionalidad se pueden clasificar en al menos tres tipos principales: i) aquellas que sostienen que la racionalidad se debe a un mecanismo general que opera o funciona de acuerdo a primeros principios; ii) aquellas que sostienen que el pensamiento racional se lleva a cabo por medio de una variedad muy rica de mecanismos específicos o esquemas que permiten al agente interactuar con el entorno de manera efectiva; y iii) aquellas que sostienen una posición dual que admite que la racionalidad se sostiene en ambos modos.

Las teorías que sostienen que la racionalidad emplea mecanismos generales tiene cuando menos tres vertientes: i) el razonamiento lógico y la postura que el conocimiento tiene un carácter proposicional y lingüístico; ii) el razonamiento Bayesiano que sostiene de manera muy general que razonar consiste en inducir las causas o los eventos que ocurren en el mundo y/o tomar decisiones a partir de las observaciones y del conocimiento que se tiene acerca del entorno; esta forma de razonar se plantea normalmente en términos probabilísticos y la inferencia se lleva a cabo mediante el Teorema de Bayes; y iii) las teorías que sostienen que el pensamiento y la toma de decisiones se basan en maximizar la utilidad de las decisiones y las acciones, como en la teoría de juegos. A continuación se describen brevemente estas tres visiones.

1.1.1. Razonamiento Lógico

El razonamiento lógico consiste en analizar si los argumentos que se expresan en el lenguaje son válidos. Los argumentos consisten en un conjunto de premisas y una conclusión, y son válidos si nunca es el caso que cuando las premisas son verdaderas la conclusión es falsa. Los silogismos, cuyo estudio se remonta a Aristóteles (385–323 AC), constituyen posiblemente el caso paradigmático de los argumentos. Las pruebas de los teoremas matemáticos son también argumentos en los que el teorema es la conclusión y las proposiciones que los justifican son las premisas. Los argumentos no se tienen que expresar necesaria o exclusivamente a través del lenguaje; por ejemplo, las pruebas de los teoremas que se presentan en Los Elementos de Euclides (330–275 AC) se expresan en parte mediante símbolos lingüísticos y en parte mediante diagramas y, sin embargo, constituyen argumentos válidos. Hay premisas cuya verdad se impone de manera directa a la mente, los llamados axiomas, y hay otras cuya verdad no es evidente y se tienen que apoyar en otros argumentos, los llamados teoremas. En una postura extrema el objeto de análisis es la totalidad del conocimiento, como el que tendría un ser omnisciente.

Los argumentos lógicos o matemáticos son casos particulares de los argumentos que se expresan en el uso del lenguaje. En la conversación cotidiana se expresan argumentos y los interlocutores aceptan las conclusiones, asumiendo que tienen una disposición cooperativa, siempre y cuando los argumentos sean válidos; sin embargo, cuando éste no es el caso, los interlocutores se dan cuenta que *el argumento no se sigue* y pueden interrumpir y refutar a quienes los expresan. Esta habilidad va de la mano de la competencia del lenguaje y se usa de manera muy efectiva en la conversación natural. Es como si la mente llevara un

registro de las premisas y la conclusión de manera continua e hiciera una prueba para verificar su validez en "tiempo real" durante el intercambio lingüístico.

El objetivo de la lógica, como disciplina de estudio, es descubrir la estructura de los argumentos válidos. Para este efecto se requiere formular o descubrir un método de análisis que permita determinar si un argumento dado es válido o no. Es deseable asimismo que esta prueba se aplique a todos los argumentos que se puedan expresar en el lenguaje. En esta visión, la racionalidad es esencialmente una labor de análisis y el objeto a analizar es el conocimiento que se expresa en el lenguaje. En el imaginario colectivo los argumentos se expresan en representaciones externas, como los textos escritos sobre una hoja de papel, y el analista realiza la prueba de validez apoyado en la representación textual con un borrador y un lápiz. En el escenario hablado los argumentos se vierten en ondas sonoras y se almacenan en la memoria, y el analista verifica su validez mentalmente.

Esta tarea ha ocupado a los lógicos desde los griegos, pasando por los escolásticos en la edad media; por los grandes avances que se dieron en los siglos XIX y XX; y por la intensa actividad que hay en dicha disciplina actualmente. Ha sido también central al estudio de la Inteligencia Artificial que tuvo como uno de sus objetivos iniciales la creación de programas de cómputo capaces de probar teoremas lógicos y matemáticos a través del análisis, que culminó en buena medida con el desarrollo de los lenguajes de programación declarativos como Lisp, y con la programación lógica, ejemplificada por el lenguaje de programación Prolog.

Sin embargo, desde esta perspectiva, la racionalidad es una facultad puramente mental y se abstrae de la interacción del agente con el mundo a través de la percepción y la acción, y del aprendizaje con base en la experiencia. La labor de análisis requiere interpretaciones perceptuales y acciones intencionales, pero éstas se asumen como dadas y no son objeto del análisis.

1.1.2. Aprendizaje y Razonamiento Bayesiano

El aprendizaje se abordó originalmente de manera explícita en la era moderna por el filósofo escocés David Hume (1711–1776) quien planteó la necesidad de contar con un mecanismo que pudiese *inducir* proposiciones generales a partir de experiencias particulares. Por ejemplo, a partir de un número de observaciones de manzanas rojas se pude inducir la proposición general que todas las manzanas son rojas. Ésta se puede sumar al acervo de conocimiento una vez que está disponible y se puede usar en el análisis, pero su generación propiamente es una tarea sintética.

Hume pensó que este tipo de inferencias se basan en un mecanismo de asociación básico; sin embargo, no le fue posible concebirlo, aunque pronosticó que a pesar de su complejidad se iba descubrir en el futuro. Esta predicción se ha materializado hasta cierto punto con las máquinas computacionales de aprendizaje de hoy en día, como las redes neuronales y los árboles de decisión, entre muchos otros mecanismos, que generan una proposición general a partir de un conjunto finito de observaciones. En Inteligencia Artificial este proceso se conoce como *aprendizaje inductivo*.

Sin embargo, la inducción no es una inferencia segura; es posible que la similitud de las observaciones sea contingente y que haya excepciones. Más aún, en las inferencias inductivas las cosas no son todo o nada y conviene hablar de grados de consistencia o coherencia, o de forma más general, del grado en que las conclusiones de un argumento se siguen de sus premisas. Por ejemplo, qué tan válido es decir que las manzanas son rojas a partir de observar nueve rojas y una verde.

Un caso esencial para la construcción del conocimiento es el aprendizaje de la relación de causalidad. Este proceso consiste en inducir una proposición ge-

neral a partir de observar un número de instancias de un efecto precedido de su causa. Esta inferencia es fundamental a la ciencia pues permite inducir las leyes generales a partir de observar las relaciones causales entre los eventos naturales y sus efectos en el mundo, y es también esencial para razonar acerca de las relaciones causales de la vida cotidiana.

La pregunta en aquella época era cómo definir argumentos cuyas consecuencias sean coherentes con los antecedentes y sean "válidos", pero tomando en cuenta la naturaleza inductiva del conocimiento. Este tipo de argumentos sería particularmente útil para modelar las relaciones causales. Sin embargo, se requería ampliar la noción de racionalidad para incluir la síntesis del conocimiento además de su análisis, y en última instancia poner al agente en el entorno a través de la percepción y la acción. En esta visión la racionalidad no es sólo un atributo de la mente sino de la conducta integral del agente.

Una solución a este problema la avanzó Tomas Bayes (1701–1761) con el teorema que lleva su nombre. Éste establece que la probabilidad que un evento sea la causa de una observación guarda una relación directa con la probabilidad que dicho evento produzca la observación y con la probabilidad que el evento ocurra, y una relación inversa a la probabilidad de que se dé la observación. Por ejemplo, el grado en que se tiene tifoidea dado que se presenta fiebre es proporcional al grado en que la tifoidea produce fiebre y al grado en que la tifoidea prolifera en la comunidad, e inversamente proporcional a la cantidad de gente que presenta fiebre, ya sea porque tiene tifoidea o porque padece otra enfermedad. Esta forma de argumentación se conoce como *Razonamiento Bayesiano* y ha tenido también una presencia muy amplia en la Inteligencia Artificial tanto en la construcción de máquinas de aprendizaje como en el descubrimiento de leyes causales.

1.1.3. Maximización de la Utilidad

La tercera forma de racionalidad basada en mecanismos generales es la teoría de juegos y la optimización. En esta vertiente la racionalidad se enfoca a la toma de decisiones, ya sea de manera individual o en entornos donde hay varios agentes cuyas decisiones afectan y son afectados por las decisiones que toman los demás. El caso paradigmático son los juegos racionales. En este paradigma el estudio de la racionalidad se centra en evaluar el valor de las decisiones para realizar la acción más favorable. Un antecedente central a esta perspectiva es el algoritmo Minimax, el cual se describe en el Capítulo 3.

La teoría de juegos se utilizó en la Inteligencia Artificial desde sus inicios en conjunto con la propuesta de la Racionalidad Limitada de H. Simon [1]. Ésta se asienta firmemente en el análisis y se asemeja a la racionalidad lógica, pero rechaza explícitamente la omnisciencia del agente, ya que una vez que el conocimiento del dominio está disponible, las consecuencias se hacen explícitas con recursos de memoria y tiempo de cómputo limitados. En esta perspectiva las soluciones de los problemas no son óptimas, y más bien el propósito del razonamiento es encontrar soluciones satisfactorias. Como en el análisis lógico, las interpretaciones se expresan simbólicamente y se asume que las acciones que realiza el agente logran sus efectos necesariamente; asimismo, se asume que se dispone de mecanismos de percepción y acción que realizan las interpretaciones y las acciones, como en el ajedrez computacional y programas similares, que simulan al mundo pero están fuera del mundo; por lo mismo, este tipo de pensamiento computacional está desligado del entorno.

Esta limitación del programa original de la Inteligencia Artificial se ha abordado desde diversas perspectivas, como las redes neuronales y los sistemas conexionistas [2] que rechazan las representaciones simbólicas; y las arquitecturas

embebidas [3], la cognición corporal –*embodied cognition*– (e.j., [4]) y el enactivismo (e.j., [5]), que rechazan a las representaciones y el cómputo simbólico de manera explícita. Una respuesta adicional al programa simbólico es el programa de los modelos causales probabilistas (e.j., [6, 7]). En éste se modela al agente racional mediante el uso de gráficas dirigidas acíclicas cuyos nodos representan conceptos y cuyas aristas representan relaciones causales, ambos expresados como probabilidades. La inferencia se centra en maximizar el valor de las decisiones como función de las observaciones que se hacen en el entorno y el conocimiento previo. Estos modelos se pueden conceptualizar como una fusión de la teoría de juegos y el razonamiento Bayesiano. Esta estrategia se relaciona estrechamente con el llamado aprendizaje por refuerzo, y hace eco de la relación entre el aprendizaje inductivo propuesto por Hume y la visión de la causalidad de Bayes. Sin embargo, estas propuestas rechazan implícita o explícitamente a las representaciones y consecuentemente al razonamiento analítico y al pensamiento conceptual y deliberativo.

1.1.4. Mecanismos Específicos y Teorías Duales

La racionalidad también se ha abordado desde una postura pragmática que postula que el pensamiento y la toma de decisiones se sostiene en el uso de mecanismos específicos que operan de manera directa, y que permiten al agente interactuar de manera muy efectiva con el entorno. Estos mecanismos se pueden pensar como heurísticas que se aplican en contextos específicos de manera sistemática y productiva aunque en ocasiones puedan llevar a conductas irracionales [8]. La motivación es simplemente que si una heurística se seleccionó por la evolución es que su uso es productivo para el individuo y la especie.

En esta visión las heurísticas o esquemas tienen que estar alineadas a las intensiones del agente y a la estructura ecológica. La distinción entre estos dos tipos de informaciones se hizo explícitamente desde el planteamiento original de la racionalidad limitada por el propio Simon [9]. Sin embargo, la computación simbólica, al no estar realmente aterrizada a través de la percepción y la acción, se enfocó a modelar las primeras y sólo de manera muy marginal a caracterizar la estructura ecológica.

En la práctica, la relación con el entorno se abordó de manera más directa con las llamadas representaciones procedurales, ejemplificadas por los llamados Marcos de Minsky [10]; este modelo fue paradigmático en su momento y se opuso frontalmente a los modelos lógicos y también a los sistemas basados en reglas de producción, que se pueden considerar como modelos basados en principios generales.

La necesidad de contar con un programa de racionalidad en que la mente se acople al entorno se ha planteado explícitamente [11, 12] y se ha abordado de manera más reciente en el contexto de la llamada Racionalidad Ecológica –*Ecological Rationality*– (e.j., [13]).

De forma más general, la racionalidad ecológica es una respuesta psicológica al racionalismo analítico para abordar la interacción con el mundo, pero se opone también a los sistemas sub-simbólicos, como las redes neuronales, la computación embebida, las redes causales, etc., ya que se implementa con módulos particulares que se activan cuando se dan configuraciones específicas del entorno, pero no se prestan a la sistematización. En este enfoque las heurísticas se implementan como algoritmos especializados.

Hay que distinguir también a las conductas esquemáticas de las reactivas: mientras estas últimas relacionan directamente los estímulos con las respuestas,

las heurísticas operan sobre las interpretaciones que produce la percepción e inciden en la conducta intencional. Es plausible que especies no-humanas con un cerebro suficientemente desarrollado cuenten con un repertorio de esquemas para interactuar con el entorno, como para cazar o jugar, como los perros cuando cachan una pelota. Estas conductas se adquieren a través de la práctica y es posible lograr niveles de competencias muy altos.

Es asimismo plausible que conductas análogas realizadas por seres humanos se lleven acabo de forma esquemática. Sin embargo, las tareas creativas y de pensamiento analítico requieren métodos generales, y se requieren ambas estrategias. En este caso la conducta heurística es el modo preferencial o de *default*. Los esquemas se activan directamente cuando sus condiciones de aplicación se presentan en el entorno; sin embargo, si no se cuenta con un esquema apropiado, o el esquema seleccionado conduce a conductas irracionales [8], pero el agente puede anticipar los resultados y tiene el conocimiento, interés y la energía para pensar explícitamente (e.j., [14]), se puede inhibir el uso del esquema y emplear mecanismo de razonamiento general. Asimismo, si las decisiones involucran aspectos éticos y una lógica afectiva, y sus consecuencias son de peso en el mediano y largo plazo, es más plausible que se recurra a un método de pensamiento de carácter general.

Pero una vez que se aceptan los dos modos, el primero constituido por un número posiblemente significativo de heurísticas y el segundo por un método general, se pueden admitir varios métodos generales que interactúan de manera muy flexible entre sí, y que capitalizan la riqueza de esquemas. En el Capítulo 9 se describe una arquitectura cognitiva con estas características. Ésta sugiere que la riqueza del razonamiento surge de una configuración heterogénea, que esta-

blece un compromiso entre la riqueza de los esquemas particulares y el poder y flexibilidad de los métodos generales.

1.2. La Perspectiva Computacional

Los estudios de la racionalidad se pueden ver desde dos enfoques muy generales: el descriptivo, en el que la racionalidad es el objeto de estudio con fines filosóficos o científicos, y el causal y tecnológico, donde lo que se pretende es construir una máquina esencial y causal a la conducta del agente.

Hasta antes de la invención de las computadoras digitales los estudios de la racionalidad tenían por necesidad un enfoque descriptivo. Éste sigue siendo de interés en muchas disciplinas científicas y las humanidades. Estos estudios pueden ser empíricos o analíticos y sus resultados pueden tener un interés puramente intelectual y/o pueden tener aplicaciones en diversas áreas del quehacer humano. En particular, pueden informar a la construcción de agentes artificiales racionales. Dada la dimensión del problema de la racionalidad y la inteligencia, estos estudios continuarán siendo vigentes por mucho tiempo.

El enfoque causal, por su parte, es mucho más restrictivo. En la conceptualización, diseño y construcción de estos dispositivos tecnológicos se adoptan posturas hacia la racionalidad y la inteligencia como las que se describen en este capítulo de manera explícita o implícita. Los modelos se implementan en una máquina computacional la cual refleja la teoría de racionalidad correspondiente. Para abordar este enfoque empezamos por estudiar a la Máquina de Turing, cuáles son sus ingredientes estructurales y funcionales, así como sus capacidades y limitaciones. Esta discusión es el tema del Capítulo 2.

Capítulo 2

La Máquina de Turing

La racionalidad computacional es un proceso completamente mecánico que surge de la operación de una computadora. El modelo teórico general de estas máquinas se presentó originalmente por Alan Turing en 1936 [15] y desde entonces nos referimos al mismo como la *Máquina de Turing* (MT). Las calculadoras creadas previamente, pasando por las de Pascal (presentada en 1642) y Leibniz (desarrollada entre 1671–1694), así como la sumadora patentada por Burroughs en 1888 y la máquina de tarjetas perforadas patentada por Hollerith en 1889, utilizada ya en el censo de los Estados Unidos de 1890, se pueden considerar como computadoras de propósito particular orientadas a las operaciones aritméticas básicas y a registrar opciones binarias; y la Máquina Analítica de Babbage, desarrollada desde 1837 hasta su muerte en 1871, aunque nunca se construyó completamente, se podía reconfigurar para hacer diversos cálculos, y se considera que era ya una computadora de propósito general; pero no fue sino hasta la introducción de la MT cuando estuvo disponible un modelo teórico completamente abstracto que describiera a las computadoras digitales de forma independiente de sus diseños y construcciones particulares.

2.1. Estructura y Funcionalidad

Los elementos constitutivos de la Máquina de Turing son una cinta dividida en celdas, un escáner y un mecanismo de control de estados. La cinta se puede ver como un renglón de una hoja de papel cuadriculado. Este sustento material se denomina aquí *el medio* de la computación. En cada celda se lee o escribe un símbolo de un alfabeto, como los numerales del 0 al 9 y las letras de la *a* a la *z*. El escáner corresponde a un lápiz con goma y se utiliza para leer, escribir y borrar los símbolos. Con éste se apunta a una celda y se lee o sustituye su contenido. Adicionalmente, el escáner se puede mover a la izquierda o a la derecha una sola celda a la vez. El mecanismo de control administra el estado en que está la máquina en un momento dado. Éste dirige al escáner y determina las operaciones sobre la cinta. El control es discreto y dado un estado y el símbolo que se inspecciona en la cinta, selecciona una operación, la lleva a cabo y cambia a otro estado. El control se puede especificar como una *tabla de transición* con una columna para cada símbolo del alfabeto y un renglón para cada estado, y la operación que se realiza cuando se inspecciona un símbolo en un estado se codifica en la celda que intersecta el renglón y la columna respectivas. El control se puede pensar como el proceso de la mente que guía a la mano que sostiene al lápiz. La máquina inicia su trabajo en un estado inicial designado y hay un subconjunto de estados llamados "de paro" tales que si la máquina llega a uno de éstos se detiene y la computación termina. Toda computación consiste en partir del estado inicial con el escáner inspeccionando una celda dada, y realizar las operaciones que se especifican en la tabla de transición hasta llegar a un estado final. El trabajo de la máquina se limita a sustituir la cadena de símbolos en la cinta, *la entrada*, por la cadena que queda al final, *la salida*.

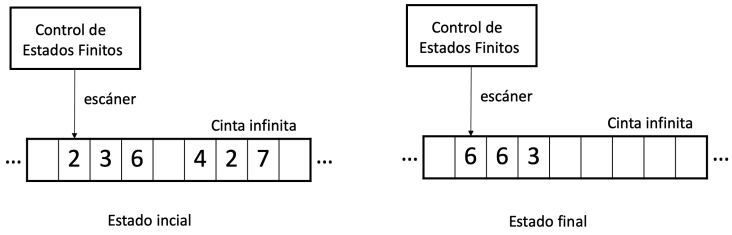


Figura 2.1: Máquina de Turing

La Figura 2.1 ilustra los elementos de la MT al inicio y al final de una computación –los diagramas a la izquierda y a la derecha. El algoritmo se codifica en la tabla de transición del control de estados, en este caso el algoritmo de la suma aritmética; los argumentos se representan como cadenas de símbolos en el estado inicial –las cadenas "236" y "427" separadas por un espacio en blanco– y el valor en el estado final –la cadena "663". El escáner apunta al símbolo más izquierdo en ambos estados.

Para comprender la máquina de manera intuitiva se puede generalizar a dos dimensiones substituyendo la cinta por un papel cuadriculado e incluyendo las operaciones para mover el escáner una celda hacia arriba y hacia abajo. Supóngase que se quiere diseñar una máquina para sumar dos números decimales, donde los sumandos están uno sobre otro y alineados por la derecha, como el algoritmo elemental que se enseña en la escuela primaria en México. En el estado inicial el lápiz está sobre el símbolo en el extremo derecho del sumando superior y la operación es leer dicho símbolo, mover el lápiz hacia abajo y cambiar a un estado que "recuerde" el símbolo leído; en el nuevo estado el lápiz apunta al símbolo más a la derecha del segundo sumando, y la operación consiste nuevamente en leer dicho símbolo y moverse hacia abajo, donde hay una celda en blanco, y pasar a un nuevo estado que "recuerda" la ruta seguida; en éste se escribe el símbolo

que corresponda a la tabla elemental de la suma. Por ejemplo, si los números sumandos superior e inferior son "236" y "427", en el tercer estado el lápiz quedará abajo del "7" y se escribirá un "3". Posteriormente el escáner se mueve para apuntar el símbolo superior de la siguiente columna hacia la izquierda y se repite el proceso de manera recurrente, considerando que puede haber "acarreos" de la columna anterior (e. j., si los dígitos son "236" y "427" se suman seis y siete, se escribe un tres y se lleva uno). Se invita al lector a diseñar la tabla de transición para la suma aritmética, para números representados por una cadena finita de dígitos. Esta tabla especifica el *algoritmo* que se enseña a los niños para sumar en la escuela primaria, y la computación desde el estado inicial hasta el final se puede considerar como un modelo del proceso mental que realiza el niño cuando suma. Los elementos materiales y conceptuales de este proceso motivan claramente el diseño de la máquina y el concepto de computación de Turing. Turing pensó que la dimensión del medio es contingente y escogió la cinta lineal y no la cuadrícula, pero los elementos conceptuales son los mismos. Sin embargo, comparar los dos formatos permite ilustrar tanto los aspectos conceptuales del modelo ideal como los prácticos de las máquinas físicas.

2.2. Consideraciones Teóricas

Desde el punto de vista teórico toda MT evalúa una función matemática. En este sentido una función es una relación entre dos conjuntos, el primero llamado dominio y el segundo codominio, en la que cada elemento del primero se relaciona a lo más con un miembro del segundo. Los elementos del dominio y el codominio se designan como los *argumentos* y los *valores* de la función, respectivamente. Si todos los argumentos están asociados a un valor la función es

total pero puede haber argumentos para los que no se defina su valor y en este caso la función es *parcial*. No existe ninguna restricción para los miembros de ambos conjuntos –pueden ser objetos físicos, como los animales y los coches, abstractos, como los números, los conjuntos y las propias funciones, o conceptuales como los objetos de conocimiento o conceptos que residen en la mente– siempre y cuando se puedan concebir como objetos individuales. Se asume que la cinta es infinita y siempre hay celdas adicionales disponibles por los dos lados; asimismo, que las operaciones se hacen con una velocidad infinita, por lo que la cadena de símbolos de salida se escribe inmediatamente una vez que se inicia la computación. Una consideración adicional es que, dado su carácter ideal, la máquina nunca falla.

Además de los elementos físicos y conceptuales de la máquina se requiere asignar un significado a las cadenas de símbolos que se escriben en el medio. Para este efecto se establecen las llamadas *Convenciones de Interpretación* [16]. Hasta ahora se han utilizado de manera implícita pero es necesario hacerlas explícitas. La más básica es que la Máquina de Turing evalúa una función. Esta convención es fundamental, pero no es necesaria, ya que la máquina podría llevar a cabo su función sin que los símbolos en la cinta se interpretaran, como en los escenarios distópicos donde las computadoras funcionan normalmente aunque ya no haya intérpretes humanos. Asimismo, bajo otras condiciones de interpretación la máquina podría hacer otra cosa. Por ejemplo, se podría interpretar como un juguete o como un reloj que llevara la cuenta de un tiempo muy extraño. Otra convención básica asociada a la primera es la estipulación que las cadenas al inicio y al final de la computación representan al argumento y al valor de la función.

Se requiere asimismo establecer *la notación*: el código o lenguaje de las cadenas en la cinta. En la Figura 2.1 se ha asumido hasta este momento que la notación es *decimal* pero esto es una contingencia, pues las cadenas podrían ser numerales especificados en cualquier base. En general las mismas cadenas reciben diferentes interpretaciones en relación a diferentes notaciones. Por ejemplo, la cadena "111" se interpreta como tres en notación monádica –en la que toda cadena de 1s de longitud n se interpreta como el número n– siete en binaria y ciento once en decimal.

El modelo exige adicionalmente incluir una referencia convencional para manipular las cadenas de símbolos en el medio. Por ejemplo, en la llamada *configuración estándar* en el formato lineal [16] se exige que al inicio y al final de la computación el escáner inspeccione la celda que contenga el símbolo más a la izquierda de la cadena y que todas las demás celdas estén en blanco. El contenido de las celdas en blanco se piensa convencionalmente como un símbolo especial de todo alfabeto –"el símbolo vacío"– que se lee y escribe de manera estándar y cuando se escribe "pone" en blanco a la celda. Asimismo, es el contenido inicial de todas las celdas, cuando la cinta o la cuadrícula "están en blanco". Por supuesto, el proceso computacional tiene que inspeccionar todos los argumentos para que la computación no sea trivial.

Aunque el objeto de la computación, es decir la función matemática, es un objeto abstracto y se puede concebir independientemente de una representación, los algoritmos se diseñan en relación a una notación, a un medio, a las convenciones de interpretación y, en particular, a una configuración estándar. Aunque éstas son contingentes, la especificación e implementación de un algoritmo requiere siempre una elección particular.

La convención de que toda máquina computa una función junto con la definición de la configuración estándar tiene una implicación directa en la teoría de la computabilidad. Esta teoría tiene por objetivo definir un modelo o mecanismo para avaluar todas las funciones para todos sus argumentos. Si este objetivo se lograra, dicha teoría sería completa. Sin embargo, hay funciones que no se pueden computar, las cuales se conocen como funciones *no computables*. Para apreciar esta distinción considere que toda computación tiene tres posibles conclusiones con sus respectivas consecuencias:

1. Si la máquina se detiene o *para* en la configuración estándar, la cadena en la cinta representa al valor de la función.

2. Si la máquina se detiene en una configuración diferente de la estándar, la función no tiene valor para el argumento dado; en este caso la función es parcial.

3. Si la máquina no se detiene, la función no tiene valor para el argumento dado y la función es parcial.

El problema es que para satisfacer los tres casos se tendría que contar con una MT llamada la Máquina de Paro (*the Halting Machine* o H) que computara la función de paro h. Esta función tiene como argumentos la descripción o nombre de la MT en cuestión m y el argumento particular n para el que se desea saber si dicha máquina se detiene o no se detiene, es decir $h(m, n)$. Si se contara con la máquina H se sabría si la máquina m bajo investigación se detiene para el argumento n, y se caería en 1) o en 2); o no se detiene, y se caería en 3); pero para este efecto $h(m, n)$ tendría que detenerse para todo m y n, es decir, siempre. Se sabe, sin embargo, que H no existe (e.j., [16]). Esta limitación se conoce uni-

versalmente como *El Problema de Paro* (*the halting problem*). Por lo mismo, la función *h* es un ejemplo de función no computable.

La imposibilidad de contar con la máquina de paro se traduce en que no se puede conocer el valor de todas las funciones para todos sus argumentos, y la teoría de la computabilidad es incompleta. Para completarla se tendría que resolver el problema del paro, pero como éste no se puede resolver por una MT, se tendría que resolver por otro tipo de máquina computacional que fuera más poderosa que la MT. Por otra parte, si se considera que la MT trabaja a velocidad infinita y cuando se detiene lo hace instantáneamente, éste no sería un problema para un ser omnisciente, quien tendría conocimiento completo de todas las funciones, y no se puede descartar que haya otra forma de computar, alternativa a la MT, que pueda resolver el problema del paro. Sin embargo, esta pregunta ha estado abierta desde su formulación original [17] y a la fecha no se ha encontrado dicha máquina, y las posibilidades de encontrarla parecen ser muy remotas.

La investigación de formalismos para caracterizar el conjunto de todas las funciones así como de mecanismos para computarlas ha sido muy activa. Tres de éstos son la teoría de las funciones recursivas, la teoría de las funciones ábacus o máquinas de registro, que son el modelo teórico de la arquitectura de Von Neumann, y la propia MT. Sin embargo, todas estas formas de computación se pueden traducir a la MT y viceversa por medio de un proceso reductivo puntal. Es decir, dada una función expresada en cualquiera de estos formalismos se puede encontrar la MT que computa a la misma función y viceversa, para todas las funciones computables. Los procedimientos para traducir entre sí a las Máquinas de Turing, las funciones recursivas y las máquinas ábacus se muestran de manera concisa y elegante por Boolos y Jeffreys [16]. Otro formalismo equivalente desarrollado por Alonzo Church, quien fuera el director de la tesis

doctoral de Turing, es el *cálculo-λ*; en éste se basa el lenguaje de programación *Lisp* –por *List Processing*– desarrollado por John McCarthy a finales de la década de los cincuenta y posiblemente el lenguaje de programación más popular en la historia de la Inteligencia Artificial. La correspondencia entre la MT y el *cálculo-λ* fue demostrada parcialmente por el propio Turing en un apéndice al artículo original y posteriormente de forma rigurosa por él mismo y por Kleene. Por estas razones hay una corriente de opinión muy sólida en Ciencias de la Computación que sostiene que: 1) La Máquina de Turing computa el conjunto completo de las funciones computables; 2) todo formalismo computacional suficientemente general es equivalente a la MT; y 3) este conjunto corresponde con el conjunto de funciones que pueden ser evaluadas intuitivamente por los seres humanos. Esta hipótesis se conoce como la Tesis de Church o la Tesis Church–Turing. En su versión más fuerte, la tesis establece que la MT es el formalismo computacional más poderoso que puede existir en cualquier sentido posible. Una consideración teórica adicional que aparece ya en el artículo de Turing de 1936 es la definición de la Máquina Universal. La tabla de transición de cada MT codifica un algoritmo que computa a una función particular, pero Turing planteó que dicha tabla también se puede especificar como una cadena de símbolos, poniendo en secuencia cada uno de sus tuplos <*estado actual, símbolo escaneado, operación, siguiente estado*>. Esta secuencia es una representación del programa que computa la función correspondiente. Con base en esta observación, Turing propuso una Máquina con un control que inicialmente lee o *carga* (*up load*) la tabla de transición y configura dinámicamente una máquina particular. Como todas las tablas de transición se pueden poner en dicho formato, la Máquina Universal puede computar todas las funciones. La Máquina Universal es programable y es el modelo teórico de todas las computadoras digitales.

2.3. Consideraciones Prácticas

En oposición a la máquina ideal, las máquinas reales cuentan con recursos finitos de memoria y velocidad de cómputo. Éstos dependen del tipo de tecnología empleada en su construcción y aunque la velocidad y capacidad de memoria de las máquinas actuales es muy significativa, hay limitaciones que deben considerarse en el diseño de algoritmos prácticos, como se ilustra en la leyenda del tablero de ajedrez y los gramos de trigo.

La primera consideración es que puede haber una gran variedad de algoritmos para computar la misma función. Por ejemplo, la definición del algoritmo para sumar es más sencilla si se utiliza la cuadrícula en vez de la cinta lineal. Asimismo, los pasos que se requieren para completar el cálculo son mucho menos con el medio cuadriculado, como se puede verificar diseñando ambos algoritmos utilizando la notación decimal. En general la geometría del medio importa porque determina las trayectorias que tiene que seguir el escáner para llevar a cabo el proceso. Adicionalmente, la notación tiene un impacto muy significativo en la definición de algoritmos. La monádica es la más simple para definir las funciones sucesor, suma e identidad, como se puede verificar diseñando los algoritmos correspondientes.

Por su parte, la configuración estándar se requiere no sólo para interpretar a las cadenas sino también para concatenar o acoplar computaciones; por ejemplo, para permitir que el valor de una computación sea el argumento de la siguiente. La necesidad de establecer una configuración estándar impacta también en el diseño de algoritmos, que se puede dividir en dos partes: el procedimiento para evaluar la función propiamente y los procedimientos auxiliares para asegurar que las computaciones empiecen y terminen en la configuración estándar. La arquitectura Von Neumann y la memoria de acceso random (RAM), que

asigna una dirección a cada localidad de memoria, permitieron la definición de algoritmos prácticos, y de ahí su utilidad.

En computaciones prácticas se requiere también determinar el número de operaciones o pasos computacionales y la cantidad de localidades de memoria necesarias para llevar a cabo la computación. Dado que estos números pueden crecer de manera muy rápida, fácilmente se podría llegar a cifras extraordinarias, y el cómputo requeriría miles o millones de años en las computadoras más rápidas que existen en la actualidad, por lo que es necesario cuantificar de antemano estos parámetros. Para este efecto es importante distinguir a la teoría de la computabilidad de la teoría de la complejidad algorítmica; mientras que la primera asume que se cuenta con recursos infinitos de memoria y las computaciones se hacen instantáneamente, la segunda trata de los algoritmos que se pueden computar de manera efectiva. La teoría de la complejidad permite abstraer hasta cierto punto sobre los medios, notaciones y configuraciones, y determinar la complejidad en términos de la forma de las funciones directamente, pero de cualquier forma es conveniente tener presente los elementos involucrados en la definición de algoritmos y su relación a las máquinas prácticas. Esto es particularmente relevante si el objetivo no es sólo realizar cálculos matemáticos complejos sino modelar las funciones de la mente.

2.4. Consideraciones Interpretativas

Una consideración final es que la máquina opera sobre formas y su trabajo consiste en transformar representaciones: tan sólo manipula símbolos mediante el escáner de manera local a celdas específicas y nunca tiene acceso a toda la cinta. Sus ingredientes estructurales incluyen a la geometría del medio y al alfa-

beto, pero no la notación ni la configuración estándar, que son implícitos en la especificación de los algoritmos. Asimismo, los contenidos o interpretaciones le son ajenos y residen sólo en la mente de los intérpretes humanos. Por lo mismo, la máquina no sabe cuál es la función que se representa en la tabla de transición o se computa por el algoritmo. No sabe tampoco que las cadenas en la entrada y la salida representan al argumento y al valor de la función, respectivamente. Las computadoras, como cualquier otra máquina, no saben nada ni son conscientes del trabajo que hacen o la información que procesan para el consumo humano.

Sin embargo, la racionalidad computacional se basa en última instancia en que el conocimiento se puede representar en computadoras digitales y que *razonar* o, de manera más general, *realizar inferencias*, son procesos computacionales que transforman a las representaciones. Consecuentemente, que todo objeto de conocimiento, incluyendo las habilidades perceptuales y motoras, se puede representar a través de funciones matemáticas, bajo un conjunto apropiado de convenciones de interpretación, y que la ejecución de los algoritmos es causal y esencial a la conducta del agente. Ésta es la hipótesis que dio lugar a la llamada Inteligencia Artificial simbólica y al proyecto de la Racionalidad Limitada.

Capítulo 3

Racionalidad Limitada

La Inteligencia Artificial simbólica se inició con la definición e implementación de modelos computacionales de la toma de decisiones en los juegos racionales cuando el conocimiento es completo –se conoce la información que tiene el adversario en cada movida. Estos programas se definen con base en las reglas del juego, cuyas consecuencias determinan el espacio del problema. El método consiste en recorrer este espacio para encontrar la mejor movida. Como el espacio es en general de dimensiones muy significativas se recurre a heurísticas para recorrerlo. Esta idea se generalizó rápidamente a otros problemas como la prueba de teoremas lógicos y, en general, a la solución de problemas que se pueden plantear en base a un conjunto de proposiciones dadas y a un conjunto de reglas con las que se pueden explorar sus consecuencias.

3.1. El Algoritmo Minimax

El algoritmo de referencia para este efecto es *Minimax*. Este algoritmo tiene un lugar distinguido en la historia del pensamiento humano. Existen alusiones

desde el siglo XVIII y se dice que George Babbage (1840) exploró la idea, aunque la referencia más temprana al ajedrez se debe a Zermelo (1913); la prueba del llamado Teorema Minimax se debe a von Neumann (1928) y este algoritmo figura prominentemente en el libro fundacional de la teoría de juegos *Theory of Games and Economic Behavior* por el propio von Neumann y Oskar Morgensterm. Norbert Wiener lo utilizó asimismo en un programa de ajedrez descrito en *Cybernetics* en 1948, y se usó también en Turochamp, el primer programa de cómputo capaz de jugar un juego de ajedrez completo, desarrollado por Turing y su colega D. G. Champernowne en 1948, aunque en dicha época no había todavía computadoras capaces de correr el programa y Turing mismo simulaba ser la máquina, para lo cual hacía cálculos durante media hora por cada movida. También se utilizó en una máquina de ajedrez hecha con relevadores electromagnéticos capaz de representar tableros hasta con seis piezas construida por Shannon en 1949, quien adicionalmente publicó el artículo *Programming a Computer for Playing Chess* en 1951. Posteriormente Newell y Simon lo utilizaron en su programa de ajedrez, desarrollado a mediados de los años cincuenta del siglo pasado. Desde entonces se ha utilizado en programas de cómputo que juegan juegos racionales y en general en la Inteligencia Artificial. Minimax está ampliamente documentado en la literatura de la Inteligencia Artificial y la Teoría de Juegos, y aquí se presenta solo la idea intuitiva.

Minimax se define en relación a una función de evaluación del estado del juego cuyo valor se desea maximizar por un jugador, en el caso del ajedrez por *blancas*, y minimizar por el otro, es decir por *negras*. La función de evaluación más sencilla para el ajedrez es la suma de los valores de las piezas blancas menos la suma de los valores de las negras en el estado de evaluación. La reina vale nueve, las torres cinco, los caballos y los alfiles tres, y los peones uno. El valor de la posi-

ción inicial es cero y varía cuando se come o se toma una pieza. Por ejemplo, en el tablero en la Figura 6.1[1] el turno es de negras y el valor de la función es -1 ya que las negras tienen un caballo de más pero dos peones de menos. Sin embargo, se puede ver fácilmente que negras moverá su reina a *h*3 en la siguiente movida y hay mate inevitable en tres movidas. Por lo mismo una función de evaluación de calidad daría un valor mucho más negativo. La función de evaluación es crucial en Minimax y en algunos dominios, como el ajedrez, ha sido objeto de un esfuerzo de diseño muy significativo.

El algoritmo propiamente desarrolla el espacio del problema como un árbol o jerarquía en la que cada nodo representa a una posición del tablero y sus hijos las posiciones a las que se puede llegar a partir de la misma, para todas las movidas posibles. La posición inicial del juego se representa por el nodo madre de todo el espacio del problema. Cada nodo tiene asignado un espacio de memoria en el que guarda el valor de la posición, considerando un número dado de movidas hacia adelante. El valor del nodo más profundo en el análisis se computa directamente por la función de evaluación.

La Figura 3.1 ilustra la situación de un turno hacia adelante. El nodo superior representa a la posición en la que se efectúa la movida y cada nivel representa una movida o *ply*. El jugador en turno es blancas, quien tiene dos movidas posibles; el oponente o negras tiene a su vez dos movidas posibles por cada movida de blancas. En la figura 3.1.a se ilustra el valor de la función de evaluación para cada una de las posiciones que se llega en las cuatro trayectorias posibles. El objetivo de blancas es maximizar y el de negras minimizar por lo que estos jugadores escogerán la movida que los lleva a la posición de máximo y mínimo valor res-

[1]Esta posición surgió en un juego jugado entre Turochamp (blancas) y Kasparov (negras) jugado en el *Alan Turing Centenary Exhibition (2012), University of Manchester, el 25 de junio.* El video está disponible en https://www.youtube.com/watch?v=yvanV9B6EBs

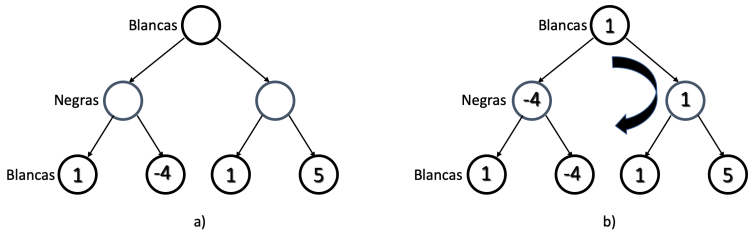


Figura 3.1: Algoritmo Minimax

pectivamente, como se ilustra en la figura 3.1.b. La flecha curva ilustra la mejor movida de negras dada la mejor movida de blancas. La búsqueda se realiza por profundidad de izquierda a derecha –*depth-first*– y los valores de las posiciones en los nodos correspondientes se registran conforme al orden del recorrido.

La idea es muy simple pero el costo de la evaluación crece de manera análoga al trigo que le pidió Sissa al rey Sheram por cada cuadro del tablero. En la Figura 3.2 se puede apreciar que el número de nodos o posiciones que hay que evaluar en el nivel n es 2^n, y que éste se duplica cada nivel menos uno, por el nodo inicial, por lo que el número total de nodos es $2^n + 2^n - 1 = 2^{n+1} - 1$. En la figura se asume que hay dos posibles movidas en cada posición,[2] pero se dice que las movidas legales posibles en cada posición en el ajedrez son aproximadamente 30 en promedio, por lo que el número de posiciones al inspeccionar n niveles o *plies*, es decir $n/2$ turnos, es del orden de 30^n. El promedio de hijos de un nodo se designa como *factor de ramificación* o r y el tamaño de un árbol o espacio de problema con factor de ramificación r y profundidad n es r^{n+1} –se evalúan directamente r^n tableros y Minimax determina el valor de r^n previos nodos. Si

[2] Esto es más o menos cierto en los juegos que juegan los maestros y los grandes maestros de ajedrez, donde se dice que en promedio hay tan sólo una o dos posiciones a lo más que no llevan a perder el juego, pero por supuesto, para saber cuáles son hay que ser un muy buen jugador.

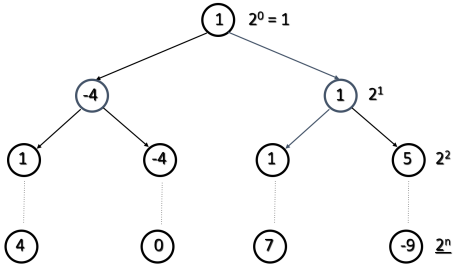


Figura 3.2: Complejidad del Ajedrez

se considera que el promedio de turnos en un juego es 50, es decir 100 *plies*, el espacio del problema es de 30^{101} posiciones. Números en este orden son muy grandes y su cálculo no es posible, incluso con la tecnología computacional más avanzada, y se requiere recurrir a estrategias adicionales para reducir el tamaño del espacio del problema.

Una estrategia muy socorrida es el llamado *recorte alfa-beta*. Aparentemente fue descubierta por John McCarthy y se utilizó muy pronto para el ajedrez por Allen Newell y Herbert Simon[3] a mediados de los cincuenta del siglo pasado, quienes comentaron que éste es un caso de múltiple descubrimiento; fue utilizada también por Arthur Samuel en su juego de Damas y su optimización fue objeto de estudio por Donald Knuth y Judea Pearl,[4] entre muchos otros. La idea intuitiva se ilustra en la Figura 3.3.a. En este punto del proceso de Minimax se ha analizado ya todo el árbol para la movida izquierda de blancas, así como la movida izquierda de negras para la movida derecha de blancas; se puede apreciar directamente que el valor mínimo de la movida de negras –quien minimiza– es cuando más el valor de la posición a la que se llega con su movida izquierda –es decir 2– pero este valor es menor que el valor de la movida izquierda de blancas

[3]Ganadores en forma conjunta del Premio Turing en 1975.

[4]Ganadores del Premio Turing en 1974 y 2012, respectivamente.

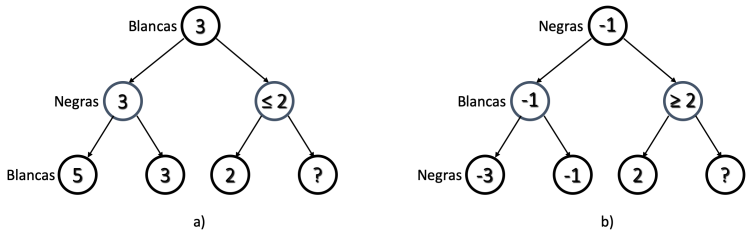


Figura 3.3: Recorte alfa-beta

–es decir 3– quien maximiza; consecuentemente, el valor de la movida derecha de negras no podrá ser mayor a 2, y blancas escogerá necesariamente su movida izquierda. Por lo mismo, ya no es necesario analizar la movida derecha de negras y toda esta rama se puede recortar del espacio del problema. Se debe asimismo considerar el caso en que el turno corresponde al jugador que minimiza en el que la situación se revierte, como se ilustra en la Figura 3.3.b. En este caso negras tiene asegurado -1 por su movida izquierda mientras que el valor de su movida derecha es al menos 2, por lo que podrá escoger su movida izquierda sin que sea necesario explorar el resto del árbol de su movida derecha. Se puede mostrar que el recorte alfa-beta optimizado reduce el espacio del problema a la mitad. Sin embargo, el espacio sigue siendo muy grande y se requieran estrategias adicionales para implementar juegos racionales de manera práctica.

El código del algoritmo Minimax es muy compacto y hay muchas implementaciones fácilmente accesibles. La expansión del espacio del problema y la estrategia de búsqueda son independientes del juego particular, como el ajedrez, las damas, el gato o el go, y se requiere definir la función de evaluación para cada juego. Ésta incluye las reglas del juego y la caracterización de las posiciones o tableros válidos, y permite generar el conjunto de movidas legales en la posición actual dado el jugador, y expandir los nodos. El número de nodos hijos prome-

dio será el valor del factor de ramificación y en principio todas las trayectorias se deben considerar. Sin embargo, la calidad del jugador humano no depende sólo de su conocimiento de las reglas y su capacidad para analizar las consecuencias de sus movidas, sino también de los conceptos estratégicos del juego. Por ejemplo, en el ajedrez es importante que las piezas tengan movilidad; controlar el centro del tablero; tener una estructura de peones sólida; que el rey esté protegido; etc. Estos conceptos se pueden representar como una rutina de análisis con un valor heurístico asociado. Cada posición tiene un valor heurístico total que toma en cuenta los valores aportados por los conceptos considerados y permite decidir si se expande el nodo que corresponde a la movida, reduciendo significativamente el espacio de búsqueda. La calidad de un programa de ajedrez depende de la implementación óptima del algoritmo de búsqueda, pero más fundamentalmente de la calidad de sus heurísticas. Éstas se desarrollan con la participación de jugadores muy sólidos incluyendo grandes maestros internacionales y se especifican en la función de evaluación.

3.2. El Ajedrez Computacional

A lo largo de la historia se han desarrollado un número considerable de programas de ajedrez, ya sea para jugar con humanos o contra otros programas, y desde hace muchos años se celebra un campeonato mundial de ajedrez entre máquinas, que sirve como ambiente de prueba para los programas, el hardware que utilizan y sus heurísticas. En la Figura 3.4 se muestra un breve resumen de algunos de los programas más sobresalientes, incluyendo el año o periodo en que estuvieron presentes en el panorama internacional, su nombre o el de su creador, su logro o relevancia y su grado o raiting FIDE internacional y/o su profundidad

Año	Programa	Logro	Nivel/*Raiting* y/o capacidad de análisis
1957	A. Berstein	Primer Programa	4-plies en 8 min.
1966– 1967	Mac Hack VI	1º. en ganar torneo	1243 a 1510
1977– 1983	BELLE	Nivel de Maestro Nacional	2203 160,000 posiciones/seg 8 plies
1981– 1986	CRAY-BLITZ	1º. en ganar a MI	2600
1985 – 1990	HITECH	1º en lograr nivel de GM	2530 2×10^5 pos/seg
1988	Deep Thought	1º. en ganar a un GMI en torneo oficial (B. Larsen)	2745 2×10^6 pos/seg
1996	Deep Blue	Match 4-2: Kasparov	100×10^9 pos/seg
1997	Deep Blue	Match 3 ½ - 2 ½: Deep Blue	200×10^9 pos/seg de 6 a 16 plies (max. 40)
2011-2013	Houdini 6	Campeón mundial 2011-12	3406
2013-2015	Komodo 11.3.1	Campeón mundial 2013-16	3408
2014-2018	StockFish 9	Campeón mundial 2018	3443 (35×10^6)

Figura 3.4: Breve historia de los programas de ajedrez

de búsqueda y velocidad. En particular se puede apreciar el progreso desde 1957 cuando se presentó el primer programa completamente programado –por Alex Berstein y colegas en una IBM 704– pasando por Deep Thought, el primero en ganarle a un gran maestro internacional, Deep Blue en su versión de 1996 que perdió ante Kasparov[5] hasta la versión de 1997 que finalmente lo venció. El panorama posterior lo dominaron los programas Houdini, Komodo y StockFish, campeones del mundo entre computadoras, que además se encuentran disponibles para computadoras personales. El pico de la capacidad de análisis y velocidad lo alcanzó Deep Blue en 1997, pero el nivel de competencia de StockFish subió significativamente a pesar de que sus requerimientos de búsqueda bajaron en tres órdenes de magnitud con respecto a Deep Blue debido a la calidad de sus heurísticas. Esta tradición tuvo un desarrollo constante hasta el 2017 cuando se

[5]Kasparov tuvo en 1999 el rating máximo de la Federación Internacional de Ajedrez (FIDE) de 2851; éste sólo se ha superado por Magnus Carlsen quien tuvo 2882 en 2014.

presentó una innovación cualitativa que utiliza redes neuronales profundas y aprendizaje por refuerzo en el contexto del programa AlphaZero [18], como se verá más adelante.

3.3. Inteligencia Artificial Simbólica

Las nociones de espacio del problema y búsqueda heurística se utilizaron desde el inicio de la IA para modelar otros tipos de inferencias, como el probador de teoremas lógico –*Logic Theorist*– considerado el primer programa de IA, que pudo probar varios teoremas de la *Principa Mathematica* de Bertrand Russell y Whitehead, y el solucionador general de problemas –*General Problem Solver*– capaz de resolver problemas arbitrarios que se pudieran formular en términos de hechos o proposiciones básicas, reglas de inferencia y metas, ambos desarrollados por Newell, Simon y J. C. Shaw en 1955 y 1959, respectivamente. Otro programa fundacional en esta línea fue el Probador de Teoremas Geométricos –*Geometry Theorem Proving*– presentado por H. Gelernter y Rochester también en 1959. Estos programas son modelos de racionalidad aunque enfocados al análisis deductivo y la inferencia válida, como en el razonamiento lógico y matemático, y a la solución de problemas a partir de un conjunto de hechos empíricos y las leyes del dominio de conocimiento. El estado inicial se expande con las reglas de inferencia aplicables, que corresponden a las movidas en los juegos o a las decisiones potenciales; el teorema o la solución se encuentra en uno de los nodos alcanzados, y corresponde a la posición deseada; y la trayectoria entre el nodo inicial y el nodo solución es la prueba del teorema o el método de solución y corresponde a la secuencia de movidas o decisiones. La investigación basada en la manipulación simbólica, la noción de espacio de estados y el uso de heurísti-

cas se desarrolló en diversos ámbitos y ha sido una de las más productivas en la Inteligencia Artificial.

Estas ideas motivaron la teoría de la Racionalidad Limitada –*bounded rationality*– propuesta originalmente por Herbert Simon en 1957 [1], en la que se establece que los seres humanos son esencialmente racionales y que para tomar decisiones o, de manera más general, realizar inferencias, exploran el espacio del problema mediante reglas de inferencia y heurísticas; sin embargo, como los recursos de memoria y velocidad de cómputo son limitados, no es posible normalmente encontrar la decisión óptima y se toma la decisión que satisface las restricciones del problema con los recursos disponibles. Esta teoría ha dado lugar a una literatura muy abundante, especialmente en el estudio de la toma de decisiones en economía y psicología –y eventualmente llevó a Simon a ganar el Premio Novel de Economía en 1978.

Simon, en conjunto con Newell, propusieron también la llamada *Hipótesis del Sistema de Símbolos Físicos*, que establece que un sistema de símbolos "aterrizados" provee las condiciones necesarias y suficientes para generar inteligencia general [19]. Posteriormente, Simon presentó una visión muy amplia de la inteligencia basada en el cómputo simbólico, la cual presentó en el libro Las Ciencias de lo Artificial –*The Sciences of the Artificial*– [20], que ha tenido un gran impacto en la llamada *IA dura* –o *strong-AI*.

Por su parte, Newell propuso que el conocimiento emerge del cómputo simbólico y reside en un plano o nivel de sistema al que llamó "el nivel del conocimiento" –*the knowledge level*– y que la única ley de comportamiento en este nivel es el *Principio de Racionalidad* [21]. Este programa se desarrolló posteriormente con las arquitecturas cognitivas SOAR [22] y ACT-R [23], la primera

enfocada a modelar tareas y aplicaciones de la IA, y la segunda a desarrollar una teoría de la mente y su relación con el cerebro.

Sin embargo, en esta visión de la racionalidad las interpretaciones del mundo están dadas y las decisiones se identifican directamente con las acciones. Es el caso de los programas de ajedrez que juegan contra seres humanos: el humano hace la movida y el programa despliega su respuesta en la pantalla. El programa lleva a cabo un "proceso mental" puro y el humano es quien realiza interpretaciones y acciones. El usuario humano está en el mundo pero el programa es una abstracción fuera del mundo. En el caso de programas contra programas la contienda es entre "agentes mentales" y los seres humanos tan sólo podemos ser testigos pasivos de la misma. Estas consideraciones plantearon la necesidad de extender el programa de la Inteligencia Artificial para relacionar al agente computacional con el entorno. A continuación se plantea una primera aproximación a este problema.

3.4. La Inferencia de la Vida Cotidiana

El razonamiento lógico y matemático, la solución de problemas, ya sea en la ciencia, la tecnología o la filosofía, y el razonamiento en el contexto de los juegos, son excepcionales. Estas actividades requieren aislarse para disminuir las distracciones y poder concentrarse. El filósofo sentado en su sillón y el jugador de ajedrez, que pasa horas en silencio frente al tablero moviendo una pieza después de varios minutos, son estereotipos de esta forma de pensar.

Dichas conductas contrastan con las actividades de la vida cotidiana que se llevan a cabo por costumbre en un contexto temporal y espacial: levantarse, bañarse, desayunar, irse a trabajar, saludar a los colegas, etc., se hacen siguiendo

un "esquema de acción", casi sin pensar. En cada situación se tiene un conjunto de expectativas acerca de lo que puede pasar, incluyendo lo que pueden decir los interlocutores, y se mantiene un equilibrio dinámico con el entorno natural y social. Sin embargo, si ocurre un evento o se realiza una acción inesperada y las expectativas no se cumplen, se rompe el equilibro, es necesario tomar nota y surge el pensamiento natural.

Ante un hecho inesperado se hace primero un diagnóstico, para después tomar una decisión, que equivale al objetivo de un plan que hay que formular y ejecutar para atender la eventualidad y poder volver a la forma cotidiana de actuar. Si un individuo regresa a su casa por la tarde y ve un charco enfrente de la puerta, como se ilustra en la Figura 3.5, se puede preguntar qué pasó. Para responder hay que realizar una inferencia que va de los hechos observados –en este caso el charco– a sus causas posibles; por ejemplo, que llovió y se tapó la coladera o que se rompió una tubería. El diagnóstico tiene dos aspectos salientes: la síntesis de las hipótesis y la evaluación de la más plausible. La síntesis es un proceso creativo en el que el símbolo se presenta a la consciencia de manera espontánea como producto de la observación y del trabajo subconsciente de la mente, y se resiste a explicaciones analíticas. La evaluación es también compleja, pero ésta es una tarea de análisis y su solución es más accesible.

La selección de la hipótesis más probable da lugar a la pregunta de qué hacer al respecto. Responderla consiste en realizar una inferencia de toma de decisión. Al igual que el diagnóstico ésta tiene un aspecto sintético y otro analítico. La síntesis consiste en poner las alternativas sobre la mesa –qué hacer ante cada hipótesis– y el análisis en seleccionar la decisión que tiene el mejor costo-beneficio. El insumo principal de la toma de decisión es el diagnóstico: si la hipótesis es que llovió y se tapó la coladera, la decisión razonable es destaparla;

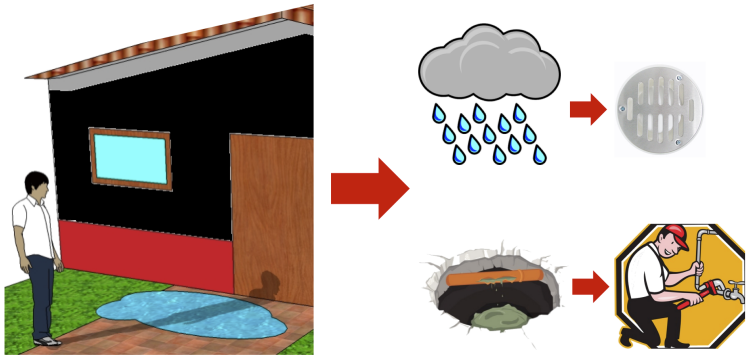


Figura 3.5: Diagnóstico y toma de deciones

pero si fue que se rompió la tubería lo razonable es repararla. Sin embargo, esta inferencia tiene otros insumos como las preferencias, los intereses e incluso los valores, y la lógica afectiva. La acción se lleva a cabo porque es valioso ahorrar y preservar el agua. El alcance de las decisiones así como la capacidad de realizarlas corresponden al grado de autonomía del agente y en última instancia a la medida en que es libre.

La decisión se convierte a su vez en el objetivo o meta de un plan, que se tiene que inducir para luego llevarlo a cabo. Es aquí donde el pensamiento permite anticipar al mundo. Si la decisión fue destapar la coladera hay que ir por la bomba y destaparla; si fue reparar la tubería hay que llamar al plomero, quien tendrá que localizar la fuga y repararla. Puede asimismo haber varios planes para cumplir con el objetivo, en cuyo caso se selecciona el mejor en términos de costo-beneficio.

Los tres tipos de inferencias se pueden pensar como módulos o procesos independientes con entradas y salidas provenientes de otros módulos y/o el entorno. La toma de decisiones en conjunto con la planeación habilitan al agente

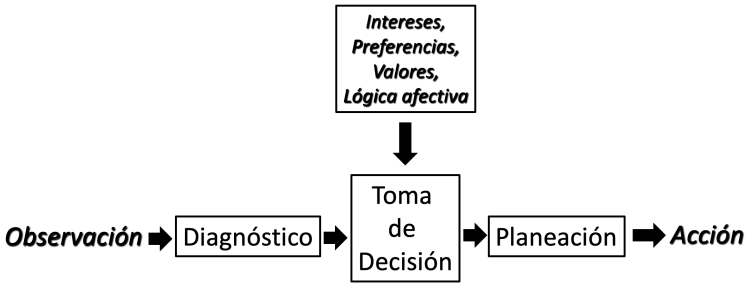


Figura 3.6: Cadena Inferencial

racional a anticipar al mundo. Esta cadena de inferencia se ilustra en la Figura 3.6.

Sin embargo, el mundo puede variar durante la ejecución del plan y es necesario verificar que el resultado de cada acción es el esperado; si éste es el caso para todas las acciones, la realización del plan conllevará a la solución del problema y el agente podrá retomar el esquema de acción cotidiano; pero si hay desviaciones, será necesario llevar a cabo el ciclo de inferencia de manera recurrente hasta recobrar el equilibrio.[6] Asimismo, toda acción conlleva al equilibro inmediato, pero la acción siempre se acompaña de la experiencia, para lograr un equilibrio más estable; en este caso, para hacer mejores diagnósticos, tomar mejores decisiones e inducir y realizar mejores planes. Para este efecto hay que agregar un cuarto tipo de inferencia: el aprendizaje. Este módulo evalúa los resultados esperados de la acción e incide sobre los otros tres tipos de inferencia.

Las inferencias de diagnóstico, planeación y aprendizaje tienen una larga tradición. La inferencia abductiva se puede conceptualizar como una elaboración y generalización de la inferencia Bayesiana, propuesta y desarrollada ampliamen-

[6] Este ciclo de inferencia se implementó en el robot de servicio Golem-III [24] exceptuando las entradas afectivas al módulo de toma de decisiones.

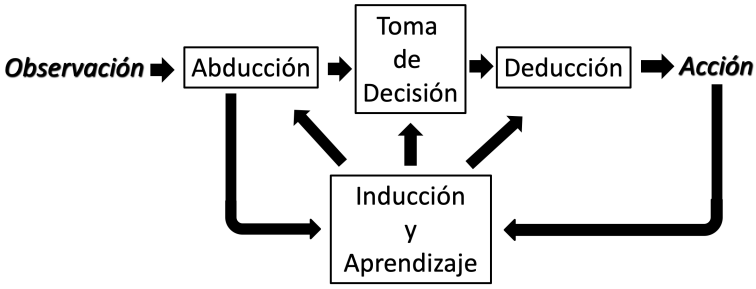


Figura 3.7: Ciclo de inferencia de la vida cotidiana

te por el filósofo pragmatista norteamericano Charles Sanders Pierce (1839-1914). La planeación, por su parte, es un caso de la inferencia deductiva, que relaciona las causas con sus efectos y por lo mismo es válida en el sentido lógico, como se discute en la Sección 1.1.1. Por su parte, el aprendizaje de hábitos se puede trazar a la inducción de aprendizaje, como se discute en la Sección 1.1.2; sin embargo hay otras formas de aprendizaje, como el conocimiento que se adquiere a través del lenguaje, que es posiblemente la forma de aprendizaje distintiva de los seres humanos. El aprendizaje cierra el ciclo de la cadena de inferencia como se ilustra en la Figura 3.7.

El ciclo de inferencia de la vida cotidiana amplia el estudio de la inferencia ya que se involucra no sólo al agente que lo realiza, incluyendo su afectividad, sino también a su interacción con el entorno, y aunque todavía "mete" a la racionalidad en un tubo de ensayo, ilustra aspectos adicionales al modelo ideal centrado en el agente. Esta perspectiva muestra directamente que los modelos de juegos enfocados a la toma de decisiones se especifican en un nivel de análisis que no distingue entre los diferentes tipos de inferencia de manera explícita. Por ejemplo, para hacer una movida en el ajedrez hay implícitamente un diagnóstico de la situación y un plan para lograr un objetivo, pero estas inferencias se subsumen

en Minimax, y se podría decir que la toma de decisiones es la inferencia esencial y que otras formas son subsidiarias. Asimismo, las inferencias lógicas, matemáticas y la solución de problemas enfatizan el análisis y asumen de manera implícita que en todo caso el diagnóstico y la toma de decisiones son inferencias implícitas y/o subordinadas. Se ha argüido también que la abducción subsume a todas las demás formas de inferencia; y el aprendizaje como se concibe ahora centrado en las redes neuronales profundas y el aprendizaje por refuerzo pretende también subsumir al resto de las formas de inferencia. Sin embargo, diferenciar explícitamente los diversos tipos de inferencia así como la forma en que se relacionan entre sí permite investigar más claramente diversos aspectos del conocimiento.

Una limitación adicional del programa inicial de la IA y la racionalidad limitada fue que no distinguieron explícitamente al pensamiento de la memoria. La memoria se requiere para desarrollar o expandir el espacio del problema pero estos programas guardan información conceptual de forma muy limitada. En una analogía muy burda es como si toda la carga recayera sobre la memoria de trabajo. Sin embargo, todo dominio conceptual consiste de un acervo de conocimiento, que incluye proposiciones, algunas de carácter particular y otras generales, que el agente tiene a su disposición y utiliza según se requiera en el pensamiento y el lenguaje. Por esta razón se requiere separar al proceso de pensamiento explícito o deliberativo de la memoria en que se almacena el conocimiento. Esta necesidad dio origen a las bases de conocimiento. El registro, verificación y retribución de informaciones corresponden a las operaciones de la memoria de largo plazo, y las respuestas que entrega el sistema de conocimiento corresponden a las inferencias conceptuales, que entregan al pensamiento las proposiciones básicas o sus consecuencias estructurales directamente. Este tipo de inferencia completa el marco conceptual básico de la IA simbólica.

Capítulo 4

Razonamiento Conceptual

El conocimiento se constituye por los conceptos generales y particulares en la mente de agentes plenamente racionales. Se puede expresar de forma pública a través del lenguaje y puede ser objeto de la introspección; por lo mismo tiene un carácter simbólico y declarativo, y se distingue de la información que habilita las habilidades perceptuales y motoras, que se embeben de forma sub-simbólica en las estructuras respectivas.

Se caracteriza asimismo porque el agente concibe al mundo como constituido por entes individuales, tanto concretos como abstractos, con sus propiedades y relaciones, que se presentan a la mente como un producto acabado de la interpretación perceptual y/o del pensamiento. El conjunto de estas entidades constituye el *universo*, *dominio* u *ontología* de individuos acerca de los que se tiene conocimiento.

Esta conceptualización se puede elaborar con otro tipo de entidades cuya extensión espacial y temporal es difusa, como los fluidos o los grumos, en los que no se distinguen claramente las partes del todo, o como los eventos, las acciones y los procesos, pero la noción de entidad individual, que tiene un conjunto de

propiedades particulares y establece relaciones con otras entidades, incluyendo a sí misma, es la intuición básica.

Las entidades individuales se presentan directamente a la consciencia de los seres humanos, y posiblemente de otros animales con un sistema nervioso suficientemente desarrollado, y pueden ser objeto de la atención –y señalarse con gestos ostensivos– así como de la manipulación mecánica y de la afectividad. De forma recíproca, los individuos o especies que no dividen al mundo en entidades individuales se enfrentan al entorno de forma holística e indiferenciada. Es posible también que esta distinción sea gradual y el grado de individuación sea incremental, y que en el plano ontológico se desarrolle y madure a lo largo del desarrollo mental. En todo caso, la capacidad de *individuación* depende de la dotación perceptual innata y del pensamiento imaginativo, y se puede modelar hasta cierta instancia con programas de visión computacional e interpretación del lenguaje natural, por ejemplo, pero su comprensión cabal es un misterio de la mente.

Aunada de manera muy cercana a la intuición básica de la entidad individual es la noción de *clase*. Ésta consiste en la partición del dominio en regiones mutuamente excluyentes, donde cada región contiene a individuos suficientemente similares desde la perspectiva conceptual que se adopte en el análisis. Estas subparticiones son las *subclases* del dominio. Las relaciones entre subclases y clases es de inclusión, y la de individuos y clases de membresía. Por ejemplo, el universo de todos los entes se puede partir primero entre los abstractos y los concretos; y los concretos entre los vivos y los inanimados. En esta jerarquía, la clase de los seres vivos se incluye en la clase de los objetos concretos y esta última en la de todos los entes; y un ser vivo particular es miembro de las clases de los seres vivos, de los concretos y de los entes. Por supuesto, la ontología puede variar con

el tiempo y en relación a los mundos posibles, y estas consideraciones se tienen también que tomar en cuenta.

La formulación de la ontología y su clasificación depende del dominio de conocimiento que se aborde, y una premisa básica es que el dominio sea completo, es decir, que incluya a todos los individuos materiales o abstractos del espacio del problema bajo estudio; de otra forma los modelos del conocimiento serían incompletos.

Los individuos tienen propiedades y relaciones. Por ejemplo, los objetos materiales tienen una extensión en el espacio y un ente puede estar adelante o atrás de otro. La extensión es una propiedad y estar enfrente o atrás son relaciones. Las propiedades pueden ser booleanas, como si un ave vuela o no vuela, o pueden tener otro tipo de valores, por ejemplo, si su color es blanco, donde "color" es la propiedad y "blanco" el valor. A diferencia de los individuos que de manera fundamental pertenecen a una clase y las clases que la contienen, las propiedades y las relaciones se puede adscribir a individuos de clases diferentes de manera muy arbitraria. Por ejemplo, los pingüinos y las águilas pueden compartir la propiedad de tener pecho blanco y cuerpo negro, aunque sean de diferentes clases.

La Figura 4.1 ilustra la ontología de los animales donde cada punto representa a un animal particular y las regiones demarcadas por líneas continuas a las clases y, en particular, el rectángulo exterior representa al dominio o universo. Las líneas más gruesas dividen al espacio en tres subclases principales que se indican con las letras de mayor tamaño: las aves, los peces y los mamíferos. Las clases en el siguiente nivel se demarcan con líneas más delgadas y letras más pequeñas; las aves se dividen en águilas y pingüinos, y se distinguen los ornitorrincos de los demás mamíferos. Los óvalos con líneas punteadas representan a las propiedades p_i, p_j y la relación r_i e indican que los puntos que contienen

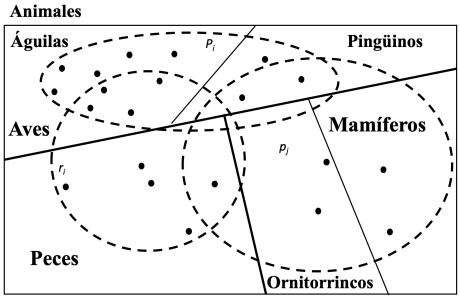


Figura 4.1: Individuos, clases, propiedades y relaciones

representan a individuos que tienen las propiedades o entran en las relaciones correspondientes. Los óvalos trascienden las fronteras de las regiones que representan a las clases arbitrariamente. El diagrama ilustra que los individuos son la intuición básica, que precede a las clases, y que las propiedades y las relaciones son más contingentes.

4.1. Representación del Conocimiento

El diagrama es una representación del conocimiento. Ésta se presenta de manera pública para su interpretación por parte de los seres humanos, y se distingue de la ontología propiamente, que se constituye por los objetos en el mundo. La representación diagramática se puede expresar también como una taxonomía estricta y representar computacionalmente por medio de jerarquías análogas a las que se utilizan en el algoritmo Minimax y programas similares, como se ilustra en la Figura 4.2. Los círculos representan a las clases y las flechas la relación de contención; los nombres de las clases se incluyen en negrillas en el círculo correspondiente, y las etiquetas asociadas a los círculos indican las propiedades y relaciones que tienen todos los miembros de la clase. El símbolo => en las eti-

quetas de forma α => β indica que la clase representada por el círculo está en la relación α con la clase β. La lectura es genérica y no se especifica quién tiene la propiedad o entra en la relación. Por ejemplo, en la Figura 4.2 se expresa que las aves vuelan, que los pingüinos nadan, que los ornitorrincos ponen huevos y que las águilas comen peces.

Asimismo los individuos particulares se representan por un rectángulo que incluye sus propiedades y relaciones, y las flechas que relacionan círculos con cuadrados representa la relación de membresía. Las propiedades booleanas se especifican directamente por su nombre; las etiquetas de forma α => β indican que el individuo representado por el cuadrado tiene la propiedad α cuyo valor es β, o que está en la relación α con el individuo β. En este caso la lectura es específica ya que las propiedades y relaciones son de y se establecen entre individuos concretos. El diagrama expresa que Pedro es grande y que Arturo es listo y es amigo o está en la relación de amistad con Pedro.

Las relaciones de inclusión y membresía representan a la relación de herencia o *modus ponens* y todos los individuos tienen además de sus propiedades y relaciones particulares las de su clase y sus súper clases. Por ejemplo, el diagrama representa implícitamente que Pedro come peces y vuela; y que Arturo nada y vuela. Las propiedades y relaciones de una clase se pueden ver también como valores asumidos o *defaults* que tienen todos los miembros de la clase, ya sea por estipulación directa o por la relación de herencia.

La economía es una propiedad muy importante de las representaciones, que en general deben tener un buen compromiso entre lo que se expresa explícita *versus* implícitamente. Éste es también un compromiso entre la cantidad de memoria que se requiere para almacenar a la información y el esfuerzo o costo del proceso de cómputo para retribuirla. En el caso límite, si toda la información se

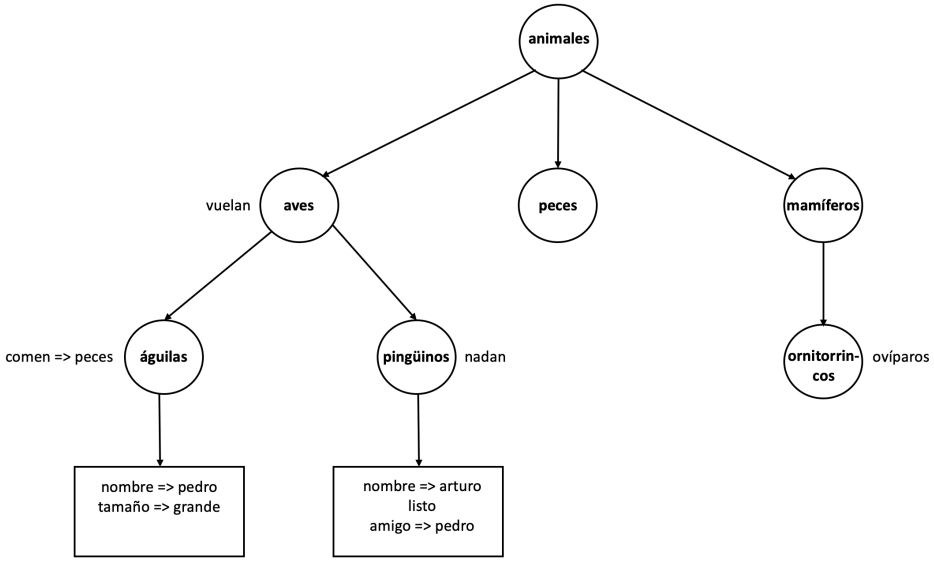


Figura 4.2: Taxonomía de los animales

expresara explícitamente se retribuiría por inspección directa; sin embargo, el conocimiento conceptual es un objeto muy grande y su almacenamiento explícito no es en general una opción viable. La partición del dominio en clases se motiva por un criterio de economía, para tener un compromiso más satisfactorio entre la memoria y el proceso.

Sin embargo, la estructura de la representación tiene efectos en la interpretación que pueden estar en conflicto con lo que se expresa explícitamente. Por ejemplo, la jerarquía conceptual en la Figura 4.2, bajo las convenciones de interpretación expuestas, expresa que los pingüinos, y en particular Arturo, vuelan, lo cual es falso. El diagrama representa una fase de la adquisición del conocimiento en la que ya se sabe que las aves vuelan y que los pingüinos son aves que nadan, pero todavía no se ha aprendido que no vuelan. Para expresar este hecho

se requiere poder expresar la negación; pero este incremento de la capacidad expresiva trae también conflictos interpretativos cuya resolución no es trivial.

4.2. Negación y Razonamiento No-Monotónico

Se conoce como *extensión* de la representación o de la *Base de Conocimiento* a la totalidad del conocimiento expresado explícita e implícitamente. En el caso de la taxonomía estricta el conocimiento implícito es todo lo que se sigue de la relación de herencia. La forma más básica de expresar la negación es asumir que lo que no se expresa es falso. Por ejemplo, si se pregunta si los peces vuelan en relación a la Figura 4.2, la respuesta será *no*, ya que la propiedad de volar no se estipula para los peces ni para los animales, su única súper clase, y en este caso es correcta.

A esta hipótesis de interpretación se le conoce como *Hipótesis del Mundo Cerrado* o *Closed World Assumption*. De manera más general ésta estipula que si algo no se sigue de la información expresada explícitamente y de los esquemas de inferencia del sistema de representación se considera que es falso. También se conoce como *negación bajo falla* en sistemas de prueba de teoremas y, en particular, en el lenguaje de programación Prolog. También se refiere en algunos contextos como *negación débil*.

La hipótesis del mundo cerrado es muy útil ya que normalmente se expresan las propiedades y relaciones que se tienen, y sólo se expresan las que no se tienen cuando es necesario marcarlas –además de que normalmente las propiedades y relaciones que un individuo no tiene, entre la totalidad posible, es mucho mayor que las que sí tiene. Sin embargo, es posible equivocarse; por ejemplo, si se pregunta si los peces comen animales, la respuesta será igualmente negativa, pe-

ro en este caso falsa. Consecuentemente, la hipótesis del mundo cerrado sólo se puede aplicar cuando se asume que la información expresada en la representación es completa; es decir, que no hay hechos relevantes del dominio que no se hayan incluido. Sólo en este caso la inferencia se puede considerar como válida o correcta, y de ahí el nombre de la hipótesis.

La negación es un gran salto en la representación del conocimiento, pero también trae consigo conflictos muy severos. En la Figura 4.3 se introduce la negación y se incluyen algunas proposiciones negativas relevantes, en particular que Pedro no es amigo de Arturo, lo que origina un pequeño drama. Se incluye además que los pingüinos no vuelan, que las aves no nadan y que los mamíferos no ponen huevos. La negación explícita, en oposición a la negación por falla, expresa la certeza de que la proposición que se niega es falsa. Para distinguirla de la negación débil a veces se refiere como *negación fuerte*. Esta es la negación del lenguaje cotidiano y de los lenguajes lógicos.

La negación explícita permite razonar acerca de dominios de conocimientos abiertos, y garantizar que las inferencias sean válidas aún si la representación es incompleta. En este tipo de sistemas la respuesta a una pregunta deberá estar en la extensión de la base de conocimientos, tanto si la proposición es positiva como negativa, y en caso contrario el sistema deberá responder *no sé*. Si se pregunta, por ejemplo, si las águilas nadan en relación a la Figura 4.3 la respuesta será *no*, pero si se pregunta si los peces vuelan en relación a la misma figura la respuesta deberá ser *no sé*.

Es posible también utilizar la hipótesis del mundo cerrado incluso en caso de que no se sepa, y responder *no* en vez de *no sé*. Esta conducta es psicológicamente plausible y aunque se confunde la negación fuerte con la débil, las respuestas podrían ser correctas frecuentemente, ya que hay mucho más proposiciones nega-

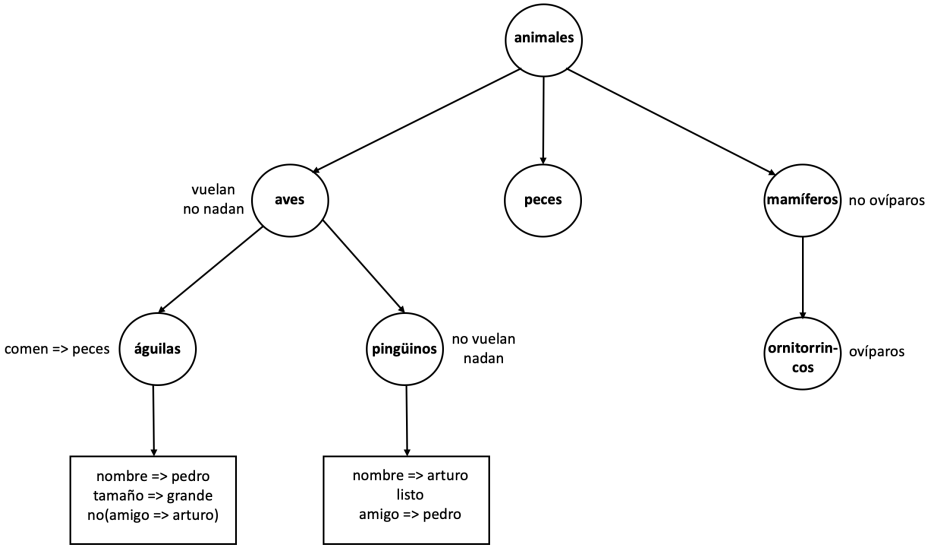


Figura 4.3: Razonamiento no-monotónico

tivas que positivas, pero también es posible equivocarse con consecuencias muy costosas. Por otra parte, hay una tendencia a la creatividad y la cortesía mexicana que se resiste a decir no sé y se prefiere responder que sí aunque no se sepa. Por ejemplo, ante la pregunta si los elefantes comen mariposas se respondería que sí, seguida de una discusión *ad nauseam*.

La riqueza expresiva que provee la negación tiene que enfrentar que las representaciones se pueden volver inconsistentes. Bajo las condiciones de interpretación estipuladas, la taxonomía en la Figura 4.3 expresa que los pingüinos –en particular Arturo– vuelan y no vuelan, que nadan y no nadan, y que los ornitorrincos ponen y no ponen huevos. Este es un problema muy serio para cualquier sistema representacional porque se puede siempre concluir cualquier cosa aunque el razonamiento no sea válido; por ejemplo, la proposición que los pingüinos no vuelan se apoya en la expresión directa de este hecho, pero su nega-

ción se apoya en que los pingüinos son aves y volar es una propiedad que tienen todas las aves.

Estas inconsistencias surgen de la estructura de la representación y para abordarlas hay que introducir convenciones, heurísticas o preferencias de interpretación adicionales que permitan afrontar el problema, aunque sea sólo parcialmente. Una heurística muy útil es el llamado *Principio de Especificidad* de acuerdo con el cual en caso de conflictos de conocimiento se prefiere siempre a la proposición más específica. Los individuos son más específicos que las clases, y las clases más específicas corresponden a las regiones más anidadas de la ontología, o a los nodos más inferiores de la taxonomía.

Bajo este principio el diagrama en la Figura 4.3 expresa que los pingüinos, en particular Arturo, no vuelan pero nadan, y que los ornitorrincos ponen huevos. Las propiedades generales como que las aves vuelan y no nadan, y que los mamíferos no ponen huevos, se mantienen pero "se bloquean" para las excepciones, que corresponden a la información más específica. Los *defaults* y las excepciones pueden ser tanto positivos como negativos; por ejemplo *vuelan* es un *default* positivo y los pingüinos es la excepción negativa, pero *no nadan* es un *default* negativo y los pingüinos son la excepción positiva. Con estas consideraciones la interpretación de la taxonomía se hace consistente. A este tipo de razonamiento se le conoce como *no monotónico* y es una aportación de la Inteligencia Artificial a la lógica [25].

El término *no monotónico* proviene de la negación de una propiedad general del conocimiento lógico y matemático; ésta consiste en que el valor de verdad de una proposición es permanente: si algo es verdadero es verdadero para siempre. Por ejemplo, cuando se prueba un teorema matemático, éste es verdadero necesariamente. Si posteriormente se muestra que el teorema es falso, es una falla

de la prueba original, pero no de la proposición propiamente. Sin embargo, el conocimiento conceptual que tiene una base empírica no tiene esta propiedad y el valor de verdad de las proposiciones puede cambiar porque se incremente el conocimiento, como cuando el niño aprende que los pingüinos son aves que no vuelan, o simplemente porque el mundo cambia todo el tiempo. Los sistemas de representación del conocimiento así como la memoria conceptual humana tienen que ser flexibles para acomodar estos cambios y reflejar al mundo lo más precisamente posible.

4.3. Preferencias y Justificaciones

Las contradicciones que dan lugar al razonamiento no monotónico dependen de la interacción de la negación con la estructura jerárquica de la taxonomía, pero pueden ocurrir por otras razones. En particular los *defaults* especificados hasta este punto son absolutos y se aplican necesariamente, pero es posible condicionarlos para que sólo se aplique en contextos específicos. Para este efecto se requiere aumentar la expresividad del sistema representacional con la implicación lógica, la cual se representa por las expresiones de forma α =>> β, donde =>> es el símbolo de implicación, α el antecedente y β el consecuente. Esta nueva expresividad se ilustra en la Figura 4.4 donde el ejemplo se enriquece con las etiquetas *trabaja*(z) =>> *vive*(z) y *nació*(z) =>> *vive*(z) asociadas a las clase *animales*, donde z es una variable cuyo rango son los lugares. Su interpretación es *si un animal trabaja en z entonces vive en z* y *si un animal nació en z entonces vive en z*, respectivamente. En el ejemplo los antecedentes y los consecuentes son positivos, pero no hay restricciones y pueden ser negativos.

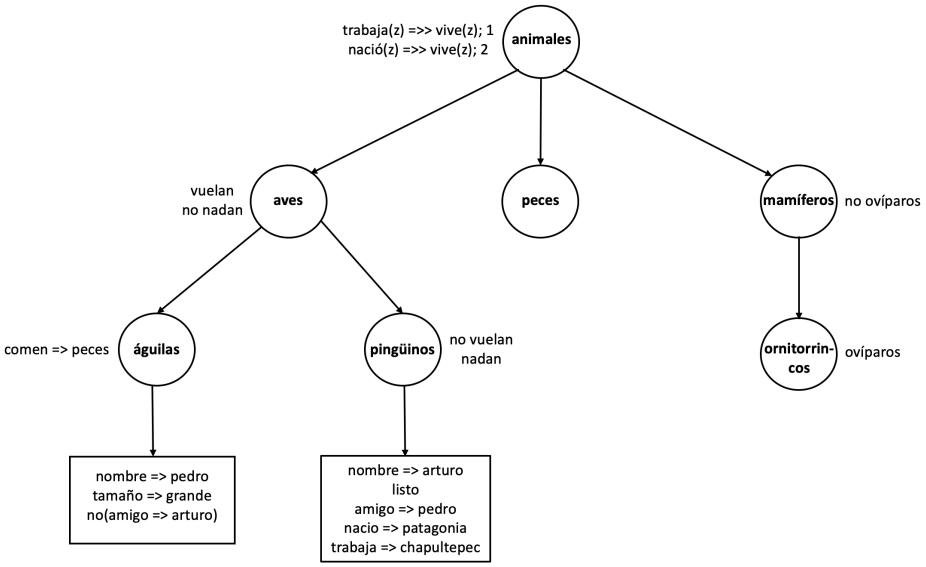


Figura 4.4: Preferencias y Justificaciones

La figura 4.4 muestra un nuevo estado de la representación en el que el agente ha adquirido o aprendido esta nueva información. En la figura se ilustra también que Arturo, el pingüino, nació en Patagonia y que trabaja en Chapultepec. Estos hechos se aprenden a través del lenguaje y posiblemente otras modalidades como la visión. La propiedad condicional se hereda a todas las clases abajo de *animales* y consecuentemente a las aves y a los pingüinos, y la extensión de la base de conocimiento incluye ahora dónde vive Arturo, pero no se aplica a Pedro, ya que para éste no se especifica dónde nació ni dónde trabaja.

Sin embargo, el aumento de la expresividad tiene también un costo y viene acompañado de un nuevo tipo de incoherencias y/o contradicciones. Se sigue en particular que Arturo vive en Chapultepec y en la Patagonia. Esto es directamente inocuo, pero si se considera la proposición de sentido común que un individuo no puede vivir en más de un lugar al mismo tiempo, las dos proposi-

ciones no se pueden sostener simultáneamente. Consecuentemente, una de las preposiciones *Arturo vive en Chapultepec* y *Arturo vive en la Patagonia* se tiene que descartar.

Esta distinción no se puede hacer con el principio de especificidad ya que los *defaults* condicionales están al mismo nivel o son igualmente específicos, y es necesario acudir a una heurística de interpretación adicional. Para este efecto se introduce el factor de preferencia, peso o prioridad, el cual se indica con el número a la derecha de los *defaults* condicionales. Éste se estipula aquí pero se puede aprender por experiencia. Se adopta la convención de que mientras más bajo es más prioritario. Se asume asimismo que todas las proposiciones tienen una prioridad, y las propiedades y relaciones tienen prioridad cero. Nos referimos aquí a toda proposición con un peso mayor que cero como *preferencia*. Los conflictos de conocimiento que no se pueden resolver por el principio de especificidad se resuelven por su valor de preferencia. En el ejemplo se concluye que Arturo vive en Chapultepec porque se prefiere a que vive en la Patagonia. La intuición es que es más plausible que alguien viva donde trabaja que donde nació, aunque esto es contingente y las preferencias pueden cambiar con la experiencia.

Las implicaciones se pueden utilizar también de forma inversa para producir justificaciones o explicaciones de los hechos que se observan en el mundo. Supongamos, por ejemplo, que el agente, que tiene la representación en la Figura 4.4 en su memoria, aprende, ya sea porque lo vea o se lo digan, que Arturo vive en Chapultepec, y se le pregunta y/o quisiera saber o justificar este hecho. Para este efecto es posible utilizar las implicaciones en reversa, y hacer la hipótesis más plausible de las causas de los hechos observados. La explicación en este caso es porque ahí trabaja, que se prefiere a que ahí nació.

Las inferencias en reversa pueden estar encadenadas, y la premisa de una implicación puede ser la consecuencia de otra, y se puede generar un árbol con varias trayectorias "hacia atrás", cada una de las cuales representa una justificación posible. La prioridad de cada hipótesis es función de los pesos de las implicaciones consideradas. Esta inferencia en reversa, de las consecuencias a los antecedentes, corresponde al diagnóstico o la abducción del ciclo de inferencia de la vida cotidiana.

La expresividad del sistema representacional se incrementa mediante la inclusión de operadores como la negación y la implicación, pero es también posible aumentarla a nivel estructural. Por ejemplo, la jerarquía estricta de la taxonomía se puede relajar admitiendo retículas o *lattices* en las que una clase puede tener más de una madre; sin embargo, esta extensión tiene el costo de que la relación de herencia se puede dar por varias trayectorias y se puede heredar una proposición por una ruta y su negación por otra; este conflicto no se puede abordar con el principio de especificidad ni el uso preferencias, y su resolución es mucho más compleja.

La taxonomía corresponde a la partición del dominio en regiones mutuamente excluyentes, mientras que la retícula admite que las regiones estén traslapadas. Esta situación se puede conceptualizar como un conjunto de particiones paralelas –mutuamente excluyentes– que se proyectan en un sólo plano. Cada plano paralelo representa a una perspectiva perceptual diferente y la proyección a la perspectiva integrada. Esta situación se puede ilustrar con la interpretación de los planos de una casa: aunque las entidades básicas –los cuartos y pasillos con sus paredes, puertas, tuberías y conductos eléctricos, que se representan por áreas y líneas– son las mismas para todos, el albañil, el electricista, el plomero, el diseñador de interiores, etc., tienen perspectivas conceptuales diferentes, que

atienden a los objetivos y prácticas de su oficio, pero el arquitecto debe integrar todas las perspectivas, y resolver los conflictos que sólo se pueden apreciar en la visión global, pero que se reflejarían en la casa, e incomodarían profundamente a la dueña.

4.4. Memoria e Inferencia

La representación de los conceptos que constituyen al conocimiento simbólico se almacena en la memoria de largo plazo. Ésta incluye dos categorías principales: la semántica y la episódica [26]. La primera incluye el conocimiento lingüístico y terminológico, el enciclopédico, el que se tiene por pertenecer a una comunidad lingüística y cultural, entre muchos otros tipos posibles, y el segundo la historia de vida o autobiografía del propio agente. Los seres humanos somos capaces de almacenar todo este conocimiento y de retribuirlo de manera muy efectiva cuando se requiere en la experiencia cotidiana y, en particular, en el uso del lenguaje y el ejercicio del pensamiento.

La taxonomía de los animales ilustra que el conocimiento es un objeto muy complejo. La extensión de la base de conocimiento en la Figura 4.4 es limitada pero si se incluyera todo el conocimiento que se tiene acerca de los animales, y más aún, el conocimiento de la vida cotidiana de un ser humano, sería de dimensiones extraordinarias. El problema es mucho mayor si se considera que el conocimiento humano es no monotónico y que el incremento de la capacidad expresiva conlleva la introducción de diversos tipos de coherencias y contradicciones; sin embargo, los seres humanos tenemos la capacidad de razonar productivamente de manera muy efectiva.

Las contradicciones surgen del conflicto entre la estructura de la representación y los medios expresivos. La estructura jerárquica interactúa con la negación y la implicación lógica, y la interpretación se requiere complementar con el principio de especificidad y el uso de preferencias. Estos conflictos son contingentes a la estructura jerárquica,[1] pero toda estructura representacional tiene que enfrentar sus conflictos particulares. Otro formalismo no monotónico con una orientación lógica y semántica es la llamada Programación por Conjuntos de Respuestas (*Answer Set Programming* [31]).

Otros esquemas, aunque no necesariamente no monotónicos, son los Marcos de Minsky, las redes semánticas, las primitivas conceptuales y las lógicas descriptivas. Esta variedad es muy amplia y el desarrollo y aplicación de esquemas representacionales es el pan de cada día de la Inteligencia Artificial.

A pesar del esfuerzo de investigación muy intenso que se ha realizado desde el inicio de la Inteligencia Artificial con el artículo de Turing de 1950, todavía no se sabe cuál es la estructura de la memoria humana ni cuáles son sus recursos expresivos, ni qué conflictos de conocimiento surgen y cómo se resuelven. Es muy intuitivo que tanto en la memoria artificial como en las memorias naturales se tiene que abordar el registro, el reconocimiento y la retribución de la información, y cualquier modelo se debe referir a un formato o estructura particular, pero la creación de un modelo capaz de representar al conocimiento conceptual humano es un problema de investigación abierto.

En particular todo sistema de representación enfrenta el llamado compromiso de la representación del conocimiento (*the knowledge representation tradeoff*) [32, 33] que establece que interpretar representaciones con información con-

[1] La taxonomía presentada aquí se desarrolló en el contexto del proyecto Golem y se usó en el robot Golem-III [27, 28, 29, 24]; también se ha aplicado al desarrollo de sistemas de información no regimentados [30].

creta es fácil pero interpretar abstracciones es costoso, y que incrementos muy moderados de la expresividad pueden tener un efecto muy significativo en el costo computacional. Consecuentemente, el esquema de representación debe ser lo suficientemente expresivo para capturar al dominio pero, al mismo tiempo, lo menos expresivo posible para que se pueda interpretar con recursos computacionales limitados. El compromiso de la representación del conocimiento se aplica a todos los sistemas simbólicos artificiales, como los que se pueden representar en computadoras digitales del tipo ordinario; sin embargo, no es claro si se aplica a memorias naturales; es muy intuitivo que razonar con información concreta es fácil si la información es limitada, pero si ésta aumenta, o el dominio de conocimiento es infinito, como en el caso de los objetos matemáticos, se tienen que introducir abstracciones para razonar de manera efectiva, y se sigue que las abstracciones se forman precisamente para razonar fácilmente con estructuras complejas o con grandes cantidades de información. Esta oposición entre la computación simbólica y el razonamiento natural es una paradoja de la representación del conocimiento [34].

La inferencia conceptual se opone a la deliberativa, como la empleada en Minimax y en general en la racionalidad limitada, en que no se requiere definir un espacio del problema dinámicamente cada vez que se requiere hacer un diagnóstico, tomar una decisión o generar un plan. En la inferencia conceptual "se piensa en la memoria". Si se pregunta si las águilas son animales carnívoros o si los ornitorrincos ponen huevos, las respuestas se presentan inmediatamente a la consciencia. Las operaciones de la memoria no son accesibles a la introspección, de manera muy similar a las informaciones que se presentan a partir de la percepción, y para pensar conceptualmente no se requiere entrar en "modo juego".

La inferencia conceptual reorienta el planteamiento del ciclo de inferencia de la vida cotidiana, como se describe en la Sección 3.4, ya que el diagnóstico, la toma de decisiones e incluso los planes espontáneos se pueden realizar retribuyendo la información relevante de la memoria: el diagnóstico es una inferencia que brinda la mejor explicación usando las preferencias en reversa; la toma de decisiones es la acción que se prefiere ante el problema que se enfrenta, y en vez de planear explícitamente se pueden realizar las acciones preferidas en una interacción muy estrecha entre el agente y el entorno [24].

Hay un compromiso entre la inferencia conceptual y deliberativa, que el agente tiene que satisfacer dinámicamente; la mejor estrategia es actuar con base al razonamiento conceptual y embeberlo en la interacción con el entorno, principalmente en el uso del lenguaje, y permanecer "conectado"; sin embargo, a veces es necesario pensar deliberativamente y entrar al "modo juego". Esta es la imagen del ajedrecista, el filósofo sentado en su sofá, el matemático y el escritor trabajando con papel y lápiz, que puede ser muy productivo y gratificante, pero viene con el costo de cerrar los canales de la percepción y la acción, y de desconectarse del mundo.

El estudio de la inferencia deliberativa y conceptual embebidas en el ciclo de inferencia de la vida cotidiana constituye el modelo general de la Inteligencia Artificial simbólica y la racionalidad limitada, como se ha mostrado en la primera parte de este libro. Sin embargo, en este modelo las entradas y salidas son representaciones de las interpretaciones que produce la percepción y genera la motricidad, y es necesario abordar directamente estos procesos y la relación del agente con el entorno. Esta discusión se aborda en la segunda parte de este libro, la cual se inicia a continuación.

Capítulo 5

Entropía y Toma de Decisiones

Los seres vivos actuamos en el mundo. La acción se orienta en el plano individual a preservar la vida y en el colectivo a preservar la especie. La inteligencia y la racionalidad dependen de la variedad y coordinación de la acción y su estudio requiere contar con una noción general de la acción. Aquí se adopta la noción de acción propuesta por Jean Piaget (1896-1980) en el contexto de su teoría del desarrollo mental [35].

5.1. Concepto de la Acción

Piaget sostuvo que de manera completamente general toda acción, es decir, todo movimiento, todo pensamiento y todo sentimiento, es la respuesta a una necesidad (Piaget, 1970, pp. 15). La necesidad puede surgir del entorno exterior o del interior, y se atiende por la acción motora, como detenerse si se va a chocar contra un objeto o evadir un obstáculo; por un pensamiento, como diagnosticar la causa de un hecho observado; o por un sentimiento, como alegrarse por un recuerdo de un hecho feliz guardado en la memoria. Asimismo, toda nece-

sidad es la manifestación de un desequilibrio del organismo –o el sistema– y la acción tiende a recuperar el equilibrio. El desequilibrio puede ser entre el organismo y su entorno o entre sus propias estructuras internas. Los organismos están sumergidos en el entorno, siempre en proceso de cambio, y las estructuras internas se afectan por las propias percepciones, pensamientos y emociones. La acción tiende a restablecer el equilibrio de manera inmediata y, al mismo tiempo, a modificar a las estructuras para lograr un equilibrio más estable. Este es un proceso continuo y el equilibrio es móvil o dinámico. La acción se puede dar de forma automática o reactiva, como respuesta a los estímulos del mundo o a los mensajes que se reciben de otros agentes, o puede ser proactiva, causada por las creencias, intenciones y la voluntad del propio individuo, que le confieren su autonomía. Desde esta perspectiva, la vida y la acción humana se sitúan en una continuidad ecológica y evolutiva. Cada individuo confronta un espacio local –contenido en el espacio global– con un grado específico de variabilidad y complejidad ante el cual tiene que actuar continuamente. El individuo tiene también un entorno interior que incluye el intelecto, la vida social y la afectividad, que es también sujeto de desequilibrios que se tienen que atender. El individuo está dotado por contraparte de un conjunto de tipos de acciones cuyas instancias –o su realización– articulan su interacción con el entorno tanto externo como interno. Las acciones a su vez se pueden realizar de forma aislada para lograr objetivos concretos, posiblemente simples, o se pueden coordinar en diferentes grados para conseguir objetivos complejos. Mientras más amplio sea el abanico de acciones y mejor su coordinación, el individuo y la especie estarán mejor dotados para sobrevivir en entornos más complejos y variables. La variedad de los tipos de acciones, su grado de coordinación y el nivel de autonomía del individuo son estrictamente dimensiones diferentes, pero se pueden abstraer en

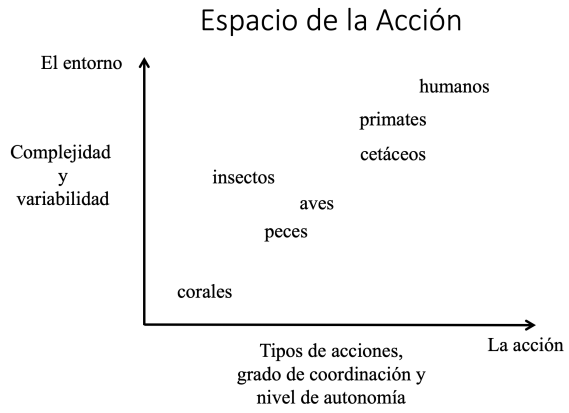


Figura 5.1: Espacio de la Acción

una sola a la que designamos aquí *la dimensión de la acción*. La complejidad y variabilidad del entorno, tanto externo como interno, las abstraemos asimismo en *la dimensión del entorno*; ambas tomadas en conjunto dan lugar al *Espacio de la Acción*. Todo individuo y especie en el mundo biológico tiene un lugar en este espacio, como se ilustra con algunos ejemplos en la Figura 5.1.

Piaget llama *asimilación* a la incorporación de la experiencia a las estructuras internas y *acomodación* a la modificación de dichas estructuras para adaptarse mejor al entorno. La adaptación puede ser pasiva como cuando se responde a un cambio inmediato del entorno, pero también activa, como cuando se modifican las estructuras causales de la conducta para anticipar cambios del entorno en el mediano y largo plazo. En el contexto computacional y aludiendo a la misma idea Turing propuso que la adaptación pasiva y activa corresponden al retorno al equilibrio después de la excitación de una masa subcrítica y a la reacción en cadena causada por la excitación de una masa supercrítica, respectivamente [36]. Un mismo individuo puede adaptarse de manera pasiva a los cambios que ocurren en el mundo de manera cotidiana y también de manera activa cuando

requiere resolver un problema para el que no tenga una conducta conocida o cuando se involucra en un proceso creativo. Los organismos más básicos despliegan conductas muy simples que permanecen fijas a lo largo de su vida, pero otros más complejos se desarrollan y pasan por una serie de estadios o planos de organización interior con un repertorio de acciones específicas, con mayor grado de estructura y coordinación, que los habilitan a atender desequilibrios más complejos y, de acuerdo con Piaget, corresponden a las etapas del desarrollo mental. Cada etapa le brinda al individuo un nuevo conjunto de conductas que le permiten adoptar una perspectiva novedosa sobre sí mismo y sobre el mundo, pero su uso requiere un amplio proceso de asimilación, que se da inicialmente de manera egocéntrica y que puede provocar desequilibrios momentáneos; sin embargo, la asimilación va aunada a un proceso de transformación de sus estructuras internas, la acomodación, que concluye al final de la etapa, cuando las nuevas habilidades se pueden usar de manera efectiva y el individuo es capaz de sostener un equilibrio más estable. De manera paralela, el proceso de acomodación engendra un nuevo conjunto de acciones que maduran a lo largo de la etapa, las cuales constituyen el inventario de conductas de la siguiente etapa. El paso de una a otra ocurre con una reorganización global de la conducta que da una nueva cualidad a la interacción del individuo consigo mismo y con el entorno –visible para un observador externo– que le permite establecer equilibrios más estables. La nueva etapa involucra a su vez un ciclo mayor de asimilación y acomodación hasta llegar a la siguiente; este proceso continúa de manera recurrente hasta alcanzar la etapa de la vida adulta. Cada etapa del desarrollo mental se puede considerar como un plano paralelo en el espacio de la acción ilustrado en la Figura 5.1 y también, desde un punto de vista evolutivo, como un estadio de desarrollo de la especie. Los equilibrios que se pueden alcanzar están supedi-

tados al repertorio y grado de coordinación de las acciones propias de la etapa. La inteligencia y el nivel de racionalidad de cada especie o individuo es función de la variedad, coordinación y nivel de autonomía de las acciones, en el plano de cada etapa y entre las diversas etapas. Aunque la acción racional se concibe tradicionalmente como aquella asociada al pensamiento humano, ésta es la culminación de un continuo que culmina en el cuadrante superior a la derecha del espacio de la acción. Los diversos estadios del desarrollo mental se pueden también conceptualizar como niveles de sistema, que corresponden a planos de análisis del fenómeno. Desde esta perspectiva cada nivel se puede analizar con sus propias funcionalidades, con sus respectivas entradas y salidas, de manera independiente de los otros. Asimismo, los niveles superiores surgen de los niveles inferiores. La relación de un nivel con los niveles inferiores se puede analizar por composición de las conductas, y por lo mismo reducirse a ellas, o la funcionalidad del nivel superior puede "emerger" por una reorganización radical de las conductas de los niveles inferiores, en cuyo caso no se puede analizar como "reductiva". Esta partición de labores permite enfocar el estudio y comprender mejor las acciones y equilibrios de cada etapa. Aunque las máquinas y los robots son creaciones humanas, se pueden también ubicar en el espacio de la acción y desde la presente perspectiva hay una continuidad entre la inteligencia natural y la artificial, salvo las limitaciones de la máquina en relación a la comprensión y la consciencia, como se discute en la Sección 2.4 y los capítulos 10 y 11.

5.2. Productividad Potencial de las Decisiones

La acción racional es la consecuencia de la toma de decisiones, pero la pregunta es si las decisiones se toman libremente o están predeterminadas. Ésta es la

oposición tradicional entre el determinismo y el libre albedrío. Este último presupone que hay un grado de indeterminación ya sea en el entorno o en la mente del agente que toma la decisión, o en ambos.

La indeterminación corresponde al contenido o cantidad de información en el entorno. La información se mide en la teoría matemática de la comunicación de Shannon con la *entropía* [37]. Shannon razonó que un mensaje es informativo en la medida que el evento que se reporta es improbable. Los mensajes que comunican hechos que ocurren necesariamente tienen un contenido de información nulo. Por ejemplo, que mañana será un día como cualquier otro. Sin embargo, si se comunica que mañana se desatará una pandemia, y el mensaje es cierto, su contenido informativo es muy significativo, así como su importancia. Shannon propuso modelar este fenómeno relacionando de manera inversa la probabilidad de ocurrencia del evento con la longitud del mensaje que lo comunica. Eventos que ocurren con toda certeza tienen probabilidad de 1 y no se requiere comunicarlos, o se comunican con "mensajes de longitud cero". La longitud del mensaje se incrementa conforme disminuye la probabilidad de ocurrencia del evento. Específicamente, si la probabilidad del evento x_i es $p(x_i)$, la longitud en bits del mensaje que lo comunica es $-log_2(p(x_i))$. Esta expresión denota consecuentemente el contenido informativo del mensaje.

Tomando en cuenta estas consideraciones Shannon definió la cantidad de información como la longitud promedio de los mensajes que ocurren en el entorno, al que se designa aquí como *el volumen de control*. A este parámetro lo designó como *entropía* o s, donde $s = -\sum_{i=1}^{n} p(x_i) \times log_2(p(x_i))$. Ésta es una fórmula de valor esperado. Si la entropía es baja los eventos que ocurren y se comunican en el volumen de control son muy predecibles; conforme se incrementa la entropía los eventos son menos probables y más impredecibles; y si la

entropía es muy alta el entorno es caótico. Consecuentemente, la entropía informacional refleja la incertidumbre de los eventos que se requieren comunicar y, consecuentemente, la indeterminación del entorno. Un corolario de esta formulación es que entornos muy predecibles o predeterminados contienen muy poca información; entornos cotidianos son predecibles en cierta medida pero tienen un grado de indeterminación y más información; y los entornos caóticos son impredecibles e indeterminados, y contienen mucha información.

La comunicación se lleva a cabo mediante señales que llevan o portan mensajes; las señales son fenómenos físicos pero la información y la comunicación pertenecen al plano de interpretación o de contenido. La entropía es proporcional a la energía que se invierte en la comunicación. Por lo mismo, la entropía refleja el esfuerzo total que la comunidad invierte en comunicarse. Estas consideraciones son explícitas o se siguen directamente de la presentación original de Shannon. En esta discusión se extienden a la relación entre la comunicación, la toma de decisiones y el cambio de conducta.

La comunicación permite a los individuos de una sociedad –la familia, la escuela, la oficina, la institución en la que se trabaja, la iglesia, la comunidad lingüística, etc.– beneficiarse del conocimiento y la experiencia de otros. Las acciones comunicativas tienen la intención de cambiar el conocimiento, las creencias, los deseos, los sentimientos, las intenciones, y de manera más fundamental, la conducta que los interlocutores llevarían a cabo sin la información que se les provee. La comunicación presupone que es posible cambiar la conducta. Si los interlocutores se desvían del curso normal de la acción gracias a la información que se les hace llegar, es porque tienen la opción. Consecuentemente, la comunicación es una precondición para la toma de decisiones en entornos sociales.

La toma de decisiones efectiva refleja la indeterminación no sólo del entorno físico sino también del social: si el mundo o la sociedad son muy rígidos no se puede cambiar la conducta, las decisiones no se pueden concretar, la comunicación no se fomenta y la entropía es baja. Por el contrario, cambios productivos de la conducta debidos a la comunicación reflejan una toma de decisiones efectiva y un entorno menos determinado, con el incremento correspondiente de la entropía. Sin embargo, un esfuerzo comunicativo considerable con alta entropía pero que se traduce en pocos cambios productivos de la conducta refleja que la toma de decisiones no es efectiva y que el entorno es demasiado impredecible o caótico. Estas consideraciones sugieren que la entropía no es sólo una medida de la indeterminación del entorno físico sino también del plano de interpretación o de contenido, y que hay un rango de entropía en el que la toma de decisiones es efectiva. Esta conjetura se denomina aquí *La Productividad Potencial de las Decisiones*.

La relación entre la entropía de los entornos de comunicación y la productividad potencial de las decisiones, cuyo valor se designa aquí como τ, se ilustra con los siguientes escenarios, cada uno de los cuales define un volumen de control particular:

- Línea de producción industrial: los obreros en una fábrica no se comunican, y si lo hacen la comunicación no afecta la dinámica del entorno. La entropía y τ son muy bajas o cero. No hay toma de decisiones.

- La vida cotidiana: la gente se comunica normalmente para cambiar las creencias y conductas de los otros; estos son los objetivos de los actos del habla en la conversación. La entropía es baja o moderada y τ tiene un valor aceptable.

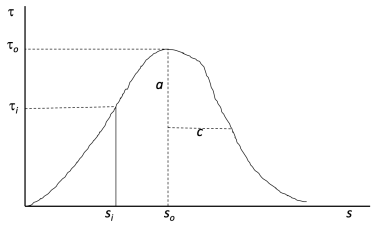


Figura 5.2: Productividad Potencial de las Decisiones

- Ambientes creativos: la gente se comunica de manera efectiva y toma decisiones que de llevarse a cabo tienen consecuencias importantes. La entropía y τ son óptimas. Hay toma de decisiones efectiva con un impacto potencial significativo.

- Situaciones de crisis: un terremoto o una pandemia. La gente se comunica significativamente y la entropía es muy alta pero el entorno es caótico y τ tiene un valor muy bajo. Puede haber una actividad de toma de decisiones muy intensa, pero éstas no se llevan a cabo y no logran sus objetivos, o los logran de manera muy limitada.

La caracterización más simple de la productividad potencial de las decisiones para diferentes entornos es una función gaussiana ψ cuyo dominio es la entropía y cuyo rango es $\tau = \psi(s) = ae^{-(s-s_0)^2/2c^2}$ donde $0 <= s < \infty$, a es el valor de τ para la entropía óptima s_o y c es la desviación estándar. Cada uno de los cuatro escenarios arriba corresponde a un volumen de control particular con entropía s_i y productividad τ_i, donde s_i se desplaza de izquierda a derecha en la Figura 5.2.

La productividad potencial de las decisiones se puede ver como un parámetro del costo-beneficio que aporta comunicarse en un volumen de control. Ésta es una variable ecológica que indica el grado en que la comunicación provee una ventaja adaptativa y esta conducta se refuerza o se fomenta. En el caso límite,

cuando $\tau = 0$, la comunicación tiene muy poca utilidad y los agentes dependen de ellos mismos, y cuando el valor de τ es muy alto la comunicación se reduce al ruido social. Esta conjetura se puede evaluar empíricamente. La comunicación y la toma de decisiones se pueden llevar a cabo por una amplia variedad de especies que producen sonidos y acciones con intención comunicativa. Estos "actos del habla" se codifican como mensajes, aunque la mayoría de las especies tienen una visión del mundo muy predeterminada, sus habilidades para enfrentar cambios en el entorno son muy limitadas, y *su* entropía y *su* τ son muy bajas. La productividad potencial de las decisiones se refleja en el espacio de la acción en la Figura 5.1 y se puede pensar como la abstracción de la dimensión vertical que subsume a la complejidad y variabilidad del entorno. Es decir, el eje vertical de la acción se puede sustituir por τ donde cada especie tiene una productividad potencial de las decisiones específica. La Figura 5.1 es especulativa y muestra una intuición subjetiva pero las coordenadas de cada especie se podrían determinar identificando su posición en el eje de la acción y evaluando empíricamente su τ mediante la observación de sus mensajes, el cálculo de su longitud promedio y los cambios de conducta productivos a los que dan lugar. Desde la perspectiva ecológica y evolutiva la productividad potencial de las decisiones es una medida de qué tanto importa el contenido para la especie. Aquellas que no se comunican o se comunican poco se limitan a actuar de manera reactiva a las señales sensadas en el entorno y no hay razón para suponer que sostienen un plano de contenido o tienen una mente; asimismo, un esfuerzo de comunicación significativo que no vaya acompañado de cambios conductuales productivos no se refuerza en la naturaleza.

Capítulo 6

Computación e Indeterminación

6.1. Determinismo de la Máquina de Turing

La máquina computacional que es causal y esencial en la toma de decisiones debe también tener un nivel de indeterminación; sin embargo, las computaciones realizadas por la Máquina de Turing están completamente predeterminadas. El propio Turing estableció en el artículo *Maquinaria Computacional e Inteligencia* [36] que las predicciones que realizan las computadoras digitales son más precisas o superan en la práctica al determinismo propuesto por Laplace (Turing, 1950, s.5). De hecho, el Demonio de Laplace es una MT que calcula todos los estados pasados y futuros del universo dadas las leyes acabadas de la física y las condiciones iniciales de un estado particular y, además, lo hace instantáneamente. Consecuentemente, si todo está predeterminado, la toma de decisiones, tanto por los organismos naturales como por las máquinas, es una ilusión.

La convención de interpretación más básica de la teoría de la computabilidad es que cada MT computa una función particular; asimismo, la enumeración

de las MTs corresponde al conjunto de las funciones computables. Una función asocia un elemento del domino a cuando más uno del codominio y esta relación se establece en la definición de la función. Los objetos matemáticos son inmutables y la relación funcional es fija. Los algoritmos son procedimientos mecánicos que producen el valor dado el argumento, pero no alteran a la función que computan. Los algoritmos se pueden conceptualizar como definiciones intensionales de funciones pero el mismo conocimiento se puede expresar de manera extensional, por ejemplo, con tablas, y los algoritmos simplemente hacen explícito el conocimiento implícito.[1] Por estas razones y bajo las convenciones de interpretación estándar, las computaciones realizadas por MTs están predeterminadas necesariamente. Por lo mismo, la noción de entropía es ajena a la teoría de la Máquina de Turing y de la Computabilidad.

La predeterminación de la MT se puede ver claramente considerando que el conocimiento del agente acerca del dominio problemático se representa por funciones matemáticas. Por ejemplo, el conocimiento del juego de ajedrez se puede conceptualizar como una función cuyo dominio y codominio son el conjunto de posiciones o tableros que pueden ocurrir en cualquier juego y el conjunto de todas las movidas posibles, respectivamente; y la relación funcional asocia cada posición con la mejor movida para el jugador en turno. La posición inicial está en el dominio para blancas pero no para negras. Como se tiene que elegir una movida necesariamente el proceso computacional está completamente predeterminado por la definición de la función.

La función "ajedrez" se ilustra diagramáticamente como una tabla en la Figura 6.1. Las columnas corresponden a las posiciones p_i, los renglones a las movidas m_j y las celdas marcadas representan la relación funcional. Por ejemplo, si

[1] La relación entre las tablas y sus algoritmos respectivos se discute en la Sección 6.6.

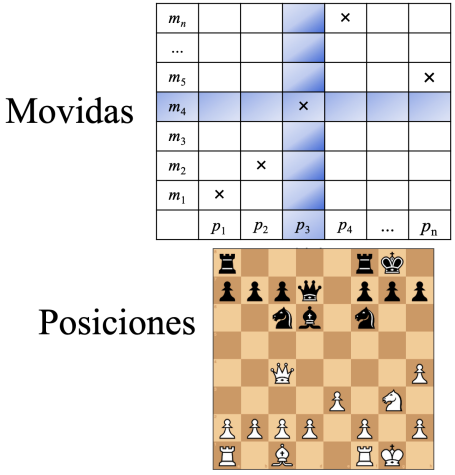


Figura 6.1: El ajedrez como una función de posiciones a movidas

la posición es p_3, el jugador en turno seleccionará necesariamente la movida m_4 por inspección directa de la tabla. La función está bien definida si cada columna tiene a lo más una celda marcada. La función es parcial si hay columnas completamente en blanco y en las posiciones correspondientes no se puede realizar ninguna movida: si el jugador en turno está en jaque el juego está perdido y, en caso contrario, el rey está ahogado y el juego es un empate.

El algoritmo codifica lo que en la tabla se estipularía de manera explícita y correrlo en una computadora estándar hace explícito el conocimiento que ya se tiene de manera implícita. La función "completa" la conoce un jugador omnisciente, como el Demonio de Laplace, mientras que los jugadores humanos tienen funciones particulares diferentes, que dependen de su constitución genética y su experiencia, pero están predeterminadas igualmente.

Esta visión de la mente se ilustra para el ajedrez en la Figura 6.2 donde el concepto de ajedrez se representa como un trazo de memoria que condiciona al

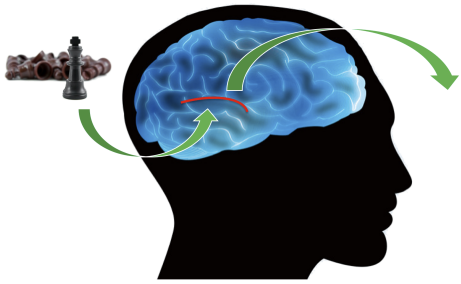


Figura 6.2: El concepto de ajedrez en la mente determinista

jugador a hacer la movida asociada a la posición como respuesta a la inspección del tablero.

6.2. Cómputo Relacional Indeterminado

Estas consideraciones nos llevan a preguntarnos si las decisiones se toman libremente o están predeterminadas, o dicho de manera más directa, si existe el libre albedrío. La respuesta que demos es esencial a nuestra noción de racionalidad y de manera más fundamental a nuestra concepción de la naturaleza humana. La posición que se tome tiene también implicaciones éticas profundas ya que si las decisiones están predeterminadas no se tiene en última instancia responsabilidad sobre sus consecuencias.

La pregunta es también relevante para nuestra comprensión de la tecnología computacional y los alcances de la Inteligencia Artificial (IA) y la robótica. Los programas de IA y los robots toman decisiones desde hace mucho tiempo –y lo harán de manera más frecuente en el futuro– y tienen cierto grado de autonomía, ya que lo hacen independientemente de sus usuarios humanos o de otros agentes y, consecuentemente, pueden desplegar conductas racionales. Sin

embargo, esta forma de racionalidad es limitada: todavía no hay programas de IA o robots que tengan intereses o valores y, de manera más general, una personalidad y una lógica afectiva a las que puedan referir sus decisiones.

La pregunta se puede plantear en relación a cualquier decisión, incluyendo las decisiones que se toman en los juegos racionales. Por ejemplo, Garry Kasparov, el campeón mundial de ajedrez de 1985 al 2000, ¿ejerce su libre albedrío cada vez que hace una movida? Si su mente y el entorno están completamente determinados todas sus movidas, es decir sus decisiones, están también predeterminadas. Kasparov puede no saberlo y creer que está tomando sus decisiones libremente, pero en este supuesto actúa por condicionamiento de su genética, su educación y la experiencia que ha adquirido a lo largo de su vida. Asimismo en el ajedrez el entorno está completamente determinado, excepto por las movidas que haga el oponente; pero si éstas están también predeterminadas todo está determinado.

Además de los juegos racionales hay que considerar a los juegos de azar. En éstos la toma de decisiones se lleva a cabo con conocimiento incompleto e incertidumbre y la pregunta es si un jugador de dominó o de póker, por ejemplo, ejerce su libre albedrío cada vez que hace una movida. En estas decisiones hay que considerar primero la productividad potencial de las decisiones. Si el entorno físico está completamente determinado es claro que las decisiones tomadas por los jugadores no tienen ningún impacto en el resultado del juego. Sin embargo, si el entorno tiene un grado de indeterminación el azar es real y su cuantificación se puede considerar en la decisión, pero si la indeterminación es muy elevada las decisiones serán igualmente intrascendentes. Asimismo, si las decisiones las toma una MT, que está predeterminada, la toma de decisiones no existe.

El determinismo absoluto de la MT se puede cuestionar desde la perspectiva de los llamados autómatas no-determinísticos.[2] Hay efectivamente autómatas no determinísticos que tienen un automata determinístico equivalente, es decir que acepta al mismo lenguaje, como es el caso de los autómatas que aceptan lenguajes regulares.[3] En este caso los autómatas no determinísticos se pueden pensar como mecanismos que permiten expresar abstracciones de manera simple y directa, pero el cómputo se realiza de manera muy eficiente por el autómata determínistico equivalente. Sin embargo, hay autómatas que son no-determinísticos genuinamente, tales como los que aceptan lenguajes ambiguos y los programas que exploran el espacio de un problema heurísticamente, el cual puede contener varias soluciones, como es el caso de los programas de ajedrez desarrollados en el programa de la racionalidad limitada. En este caso, dado que el argumento de entrada puede tener más de una solución, el objeto que se computa es una relación matemática y no una función. Sin embargo, una relación se puede computar por una o varias MTs trabajando en serie o en paralelo, y esta clase de no-determismo se simula con máquinas deterministas.

El determinismo de la MT se puede también poner en duda desde la perspectiva del cómputo estocástico. Esta estrategia consiste en partir el espacio del problema en regiones promisorias y visitarlas o saltar a ellas aleatoriamente utilizando números aleatorios o *random* hasta encontrar la solución. La estrategia involucra identificar dichas regiones, definir heurísticas para buscar en las mismas, y saltar a otra región cuando la búsqueda no sea exitosa o se haga muy costosa. La estrategia la propuso el propio Turing (Turing, 1950, s.7) quien la ilustró con el problema de encontrar un número entre 50 y 200 que sea igual a la suma de los cuadrados de los dígitos que lo constituyen. Este problema no tiene solución

[2] Ver, por ejemplo [38, 39]

[3] *idem*

pero ilustra claramente el contraste entre la estrategia determinista, que consiste en iterar desde el primero hasta el último número y verificar si cumple con la condición, y la estocástica, que consiste en escoger un número en el espacio del problema de forma aleatoria y verificar la propiedad en cuestión. La estrategia estocástica puede no terminar, a menos que se imponga un límite al número de pruebas, o se registren todos los números según se vayan inspeccionando, pero con un costo muy alto en memoria y tiempo de proceso. Turing llamó a estas estrategias o métodos como el *sistemático* y el *de aprendizaje*, y sugirió que el segundo se puede considerar como la búsqueda de una forma de conducta y, dado que es posible que muchos problemas tengan un número muy significativo de soluciones, el método de aprendizaje parece mejor que el sistemático (Turing, 1950, s.7). Una vez más cada instancia del problema tiene varias soluciones y el objeto que se computa es una relación y no una función, pero el proceso se simula con computadoras digitales del tipo ordinario y es, en última instancia, determinístico.

Para que el método estocástico sea genuino se requiere que haya un grado de indeterminación en la máquina computacional. Turing también fue explícito en este punto y sugirió una variante de la computadora digital que incluyera un elemento aleatorio que se pudiera simular por procesos determinísticos. Su propuesta fue escoger el siguiente dígito en la expansión del número π cada vez que se requiriera un nuevo número aleatorio (Turing, 1950, s.4). Aunque la secuencia está predeterminada estos dígitos no se conocen de antemano –se requieren calcular explícitamente– y se pueden utilizar para modelar procesos estocásticos, pero realmente son pseudo-aleatorios, y el cómputo es determinístico.

Números aleatorios genuinos –en oposición a simulados– se pueden producir sensando una propiedad del entorno que no se pueda predecir; en la prác-

tica se han utilizado diversas estrategias y el número aleatorio se utiliza como un argumento invisible del algoritmo estocástico; aunque dicho argumento no se conozca el objeto computacional es la función y el proceso es a final de cuentas determinístico. El entorno puede tener un nivel de indeterminación que incide en el cómputo debido a la interacción. Sin embargo, los ambientes en que se utilizan computadoras y robots tienen en general entropías muy bajas, como en las líneas de producción industrial y los ambientes de oficina, y la indeterminación es en general baja.

Para superar la limitación del determinismo absoluto de la Máquina de Turing se describe aquí una teoría de la computación con su validación empírica, al menos a nivel de prueba de concepto [40]. A continuación se detalla la maquinaria conceptual correspondiente y en el Capítulo 7 se presenta una aplicación que incide de manera directa en la noción de racionalidad que se propone en este libro [41]. Las implicaciones para el concepto de computación se elaboran ampliamente en el Capítulo 8.

Como punto de partida se extiende aquí la convención de interpretación básica de la noción de computabilidad y se define a la relación matemática –en vez de la función– como el objeto básico de computación. Este modelo se denomina aquí *Computación Relacional Indeterminada* (CRI) como sigue:

Sean los conjuntos $A = \{a_1, ..., a_n\}$ y $V = \{v_1, ..., n_m\}$, con cardinalidades n y m, el dominio y el codominio de una relación finita $r : A \rightarrow V$.[4] Para simplificar la notación se define asimismo una función $R : A \times V \rightarrow \{0, 1\}$ por cada relación r –la relación en minúsculas y la función en mayúsculas– tal

[4]Se mantiene la convención de llamar a los objetos del dominio y del codominio *los argumentos* y los *valores*, respectivamente.

que $R(a_i, v_j) = 1$ o *verdadero* si el argumento a_i se relaciona con el valor v_j en r, y $R(a_i, v_j) = 0$ o *falso* en caso contrario.

En este formalismo *evaluar una relación* se conceptualiza como seleccionar aleatoriamente uno de entre todos los valores asociados al argumento dado. De la misma forma que la expresión "$f(a_i) = v_j$" expresa que el valor de la función f asignado al argumento a_i es v_j, en el cómputo relacional indeterminado la expresión "$r(a_i) = v_j$" establece que el valor de la relación r asociado al argumento a_i es un objeto arbitrario v_j que se selecciona aleatoriamente –con una distribución de probabilidad apropiada– entre los valores para los cuales $R(a_i, v_j)$ es verdad.

6.3. Entropía Computacional

De manera análoga a la entropía de Shannon, que se define como el valor esperado de la longitud de los mensajes en el entorno, la entropía computacional se define aquí como el valor esperado del número de valores asociados a un argumento de la relación [42, 43]. Sea μ_i el número de valores asociados al argumento a_i en la relación R, n la cardinalidad de su dominio y $\nu_i = 1/\mu_i$; en caso que la relación sea parcial y $\mu_i = 0$ para el argument a_i entonces $\nu_i = 1$. La entropía computacional e de la relación R es $e(R) = -1/n \sum_{i=1}^{n} log_2(\nu_i)$.

La función asocia a lo más un valor para cada uno de sus argumentos y su entropía es cero. Las funciones parciales no definen valores para todos sus argumentos, pero éste es un hecho completamente determinado y la entropía de funciones parciales es también cero. El valor de la entropía crece conforme al incremento de valores diferentes asociados a los argumentos y la entropía es máxima cuando se asignan todos los valores posibles a todos los argumentos, en cuyo caso la información está completamente confundida y el cómputo es vacío.

La distinción entre computación no-entrópica y entrópica corresponde a la oposición entre representaciones "locales" *versus* "distribuidas". De acuerdo con Hinton [44] la Máquina de Turing mantiene representaciones locales en las que la relación entre las *unidades de forma* –los símbolos propiamente– y las *unidades de contenido* –lo que significan– es *uno a uno*, mientras que en las representaciones distribuidas esta relación es *muchos a muchos*.

La distinción se puede apreciar considerando que cada función representa a un concepto, como se ilustra arriba con la función ajedrez. Supongamos que hay una función adicional para el juego de damas que se define de manera análoga. Si los elementos de memoria en que se expresan las dos funciones son mutuamente excluyentes, como es el caso en la MT, las representaciones son locales; pero si dichas funciones comparten unidades de memoria la representación es distribuida. La ventaja de esta última es la economía de la memoria además de que las unidades compartidas permiten hacer asociaciones directas entre los dos juegos; el costo es que habrá indeterminación en qué tanto una u otra unidad de memoria contribuye a uno u otro juego; por esta razón las representaciones genuinamente distribuidas son entrópicas.

6.4. Operaciones Relacionales

El cómputo relacional indeterminado tiene tres operaciones básicas: *abstracción*, *inclusión* y *reducción*. Sean r_f y r_a dos relaciones arbitrarias de A a V, y f_a una función con el mismo dominio y codominio. Una función es también una relación por lo que en las siguientes definiciones las relaciones pueden ser funciones. Las operaciones se definen como sigue:

- Abstracción: $\lambda(r_f, r_a) = q$, tal que $Q(a_i, v_j) = R_f(a_i, v_j) \vee R_a(a_i, v_j)$ para todo $a_i \in A$ y $v_j \in V$ –i.e., $\lambda(r_f, r_a) = r_f \cup r_a$.

- Inclusión: $\eta(r_a, r_f)$ es verdadero si $R_a(a_i, v_j) \rightarrow R_f(a_i, v_j)$ para todo $a_i \in A$ y $v_j \in V$ (i.e., implicación material) y falso en caso contrario.

- Reducción: $\beta(f_a, r_f) = f_v$ tal que si $\eta(f_a, r_f)$ es verdad $f_v(a_i) = r_f(a_i)$ para todo a_i, donde la distribución aleatoria se centra en f_a. Si $\eta(f_a, r_f)$ no se satisface $\beta(f_a, r_f)$ no está definida.

La operación de abstracción construye una relación a partir de dos relaciones. La operación λ genera a q incluyendo o"agregando" a la relación r_a en r_f mediante la disyunción inclusiva de cada uno de los valores asociados de cada argumento, para todos los argumentos. Si la relación q se construye a partir de la disyunción de un conjunto finito de funciones –es decir, si r_a es siempre una función– el conjunto de funciones que construyen a q se designa C_q.

La operación de inclusión verifica si todos los valores de un argumento de R_a están incluidos en el conjunto de valores del mismo argumento de R_f, tal que $R_a \subseteq R_f$. La operación de inclusión sólo es falsa en caso de que $R_a(a_i, v_j) = 1$ y $R_f(a_i, v_j) = 0$ –o alternativamente si $R_a(a_i, v_j) > R_f(a_i, v_j)$– para al menos un par (a_i, v_j).

El conjunto de funciones que satisface la operación de inclusión o, alternativamente, el conjunto de funciones *incluidas* en la relación q se denomina aquí como I_q. Se puede visualizar fácilmente que $C_q \subseteq I_q$. Por ejemplo, sea $f_1 = \{(a_1, v_1), (a_2, v_2)\}$, $f_2 = \{(a_1, v_2), (a_2, v_1)\}$ y $q = \lambda(f_1, f_2)$. La inclusión $\eta(f_i, q)$ se satisface no sólo por f_1 y f_2 sino también por $f_3 = \{(a_1, v_1), (a_2, v_1)\}$ y $f_4 = \{(a_1, v_2), (a_2, v_2)\}$; es decir $\eta(f_3, q) = 1$ y $\eta(f_4, q) = 1$. Por esta razón las operaciones de abstracción e inclusión son productivas. Esta propiedad es análo-

ga a la capacidad de generalización de los algoritmos de aprendizaje de máquina supervisados que admiten no sólo a las muestras que se utilizan en la fase de entrenamiento sino también a aquellas suficientemente similares pero no utilizadas en el entrenamiento.

Reducción es la operación de aplicación funcional o evaluación propiamente. Si la función f_a está incluida en la relación r_f, la reducción genera una nueva función tal que el valor de cada uno de sus argumentos se selecciona aleatoriamente del conjunto de valores del argumento correspondiente de r_f. La operación de reducción selecciona una función contenida en r_f en base a una función *pista*, *índice* o *cue*. Mientras menor sea la entropía de r_f más improbable que ésta incluya a f_a, pero si éste es el caso, mayor será la similitud entre la función índice f_a y la función retribuida f_v; en el límite, si la entropía es cero y $\eta(f_a, R_f) = 1$, $\beta(f_a, r_f) = f_a$ –i.e., $f_v = f_a$. Sin embargo, β es una operación constructiva tal que la función f_a es la pista o llave para recuperar o construir una función a partir de r_f, tal que v_j se selecciona $\{v_j|(a_i, v_j) \in r_f\}$ para cada a_i utilizando una distribución aleatoria apropiada, centrada sobre $f_a(a_i)$. Si f_a no está incluida en r_f el valor de la operación β no está definido. Por otra parte, mientras mayor sea la entropía de r_f más probable que ésta incluya a f_a, pero la similitud entre f_v y f_a será menor.

6.5. El Compromiso de la Entropía

La inclusión de la entropía en la teoría de la computación da lugar a un nuevo compromiso o *trade-off* al que se designa aquí *El Compromiso de la Entropía* –*The Entropy Trade-Off* – como sigue: Si la entropía de una máquina computacional es muy baja las computaciones son predeterminadas en la misma me-

dida; en el caso límite en que la entropía es cero la computación esta completamente determinada; en el otro extremo, si la entropía es muy alta la información se confunde y las computaciones no son factibles; sin embargo, existe un intervalo de valores de la entropía en que la indeterminación es moderada y los agentes computacionales pueden realizar conductas diversas, pero las computaciones son factibles. La Máquina de Turing corresponde al caso en que la entropía es cero y las computaciones están completamente determinadas, pero esta máquina no toma en cuenta a la entropía y tiene que asumir costos computacionales muy elevados. El compromiso de la entropía corresponde a la productividad potencial de las decisiones de los diversos entornos.

La interacción del agente computacional con el entorno introduce un grado de indeterminación en el proceso de cómputo, como se ilustra en la Figura 6.3. La información del entorno, como la cara de la moneda que resulta de echar un volado, lanzar un dado o hacer la sopa en el dominó, se alimenta al volumen de control del agente, y desde su punto de vista esta información es genuinamente aleatoria. Los volúmenes de control se pueden embeber en volúmenes más amplios, desde los más específicos hasta el universo entero, y pueden contener información aleatoria.

El determinismo, como el de Laplace, se opone a las visiones indeterminadas del mundo en que rechaza que haya o pueda haber eventos genuinamente espontáneos dentro de un volumen de control; para las visiones indeterminadas, por el contrario, estos eventos ocurren todo el tiempo. En la presente propuesta esta oposición no es un absoluto y la medida en que se puedan tomar decisiones libremente, que a su vez sean efectivas, requiere que haya un nivel de entropía en la mente y en el entorno.

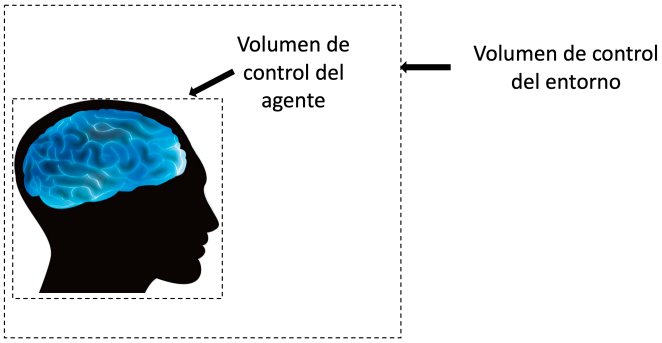


Figura 6.3: Indeterminación debida al agente y al entorno

6.6. Tablas y Computabilidad

El conjunto F de todas las funciones totales y parciales cuyos dominio y codominio son finitos se puede poner en una lista.[5] Sean $A = \{a_1, ..., a_n\}$ y $V = \{v_1, ..., n_m\}$, con cardinalidades n y m, el dominio y el codominio de un conjunto de funciones $F_{n,m}$ tales que $n, m \geq 1$. Es decir, $F_{n,m}$ es el conjunto de todas las funciones totales y parciales que se pueden formar con n argumentos y m valores. Los argumentos y los valores representan a objetos individuales arbitrarios, pero se abstrae de su extensión y su cualidad, y lo único que se requiere es que se ordenen en una lista.

Toda función $f_k \in F_{n,m}$ se puede representar en una tabla de n columnas y m renglones en la que cada columna tiene a lo más una celda marcada, donde las columnas, los renglones y las celdas marcadas representan a los argumentos, los valores y la relación funcional, respectivamente.

[5]Este conjunto es diferente del conjunto de las funciones totales y parciales para dominios y codominios infinitos, el cual no es numerable. Esto se muestra mediante el argumento antidiagonal de Cantor –e.j., (Boolos & Jeffrey, 1989, cap. 2).

f_k	1	2	4	7
v_7				X
v_6				
v_5				
v_4			X	
v_3				
v_2		X		
v_1	X			
	a_1	a_2	a_3	a_4

Figura 6.4: Representación Diagramática de las Funciones Discretas Finitas

En la Figura 6.4 se ilustra una tabla que expresa una función cuyo dominio y codominio son los conjuntos $A = \{a_1, a_2, a_3, a_4\}$ y $V = \{v_1, v_2, v_3, v_4, v_5, v_6, v_7\}$ con cardinalidades $n = 4$ y $m = 7$, respectivamente. La función tiene un índice f_k que se indica en el renglón superior, en este caso 1247. Éste es un número en base $m + 1$ con n dígitos que se forma con el subíndice j del valor v_j para el argumento a_i en caso que la función esté definida para dicho argumento y con un 0 en caso contrario, para todos los argumentos. En particular, el índice formado por una cadena de n 0s corresponde a la función vacía que no asigna valor a ninguno de los argumentos. El índice, así como los nombres de los argumentos y los valores, son metadatos que no se consideran parte de la tabla propiamente.

La figura ilustra claramente que f_k va de "0000" a "7777" en base ocho –es decir, de 0 a 4095 en base 10. El número de función, incluyendo a la función vacía, es $N_k = (f_k)_{10} + 1$, donde (f_k) es un número en base $m + 1$ y $(f_k)_{10}$ denota a f_k en base 10. En el ejemplo, $N_k = (f_k)_{10} + 1 = 1247_{10} + 1 = 680$.[6]

[6]Esta representación permite determinar de manera directa el conjunto de funciones I_q incluidas en una relación q cuyos dominio y codominio tienen cardinalidades n y m, como sigue:

1. Sea Q el conjunto vacío;

2. Para toda i tal que $0 \leq i \leq (m+1)^n$ si $\eta(f_i, q) = 1$ agregar f_i a Q.

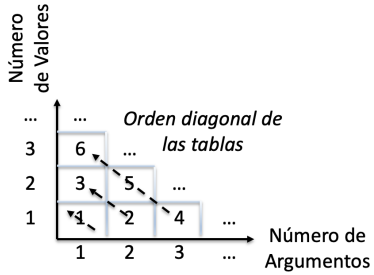


Figura 6.5: Orden de Tablas para Funciones con Dominios y Codominios Finitos

El número de funciones totales y parciales que se pueden formar con n argumentos y m valores es $(m+1)^n$. Esto se puede apreciar directamente considerando que el número de funciones es simplemente el número de combinaciones que se pueden formar con n argumentos, cada uno de los cuales se puede asociar a $(m + 1)$ valores. En la tabla de la Figura 6.4 el número total de funciones es $(7 + 1)^4 = 4096$.

Pasamos ahora a ordenar el conjunto de funciones $F_{n,m}$ –o el conjunto de funciones que se pueden incluir en una tabla de $n \times m$. Todos los pares ordenados (n, m) se pueden ordenar en una lista. Consecuentemente, todos los conjuntos $F_{n,m}$ se pueden ordenar en una lista. Cada conjunto $F_{n,m}$ tiene un índice t. En la Figura 6.5 se muestra un ordenamiento diagonal particular de todas las tablas de tamaño $n \times m$ para $m, n \geq 1$, donde cada tabla se etiqueta con su índice t. En particular $t = 1$ es el índice de las funciones $F_{1,1}$ con un argumento y un valor a lo más. Este conjunto tiene dos funciones: f_0 la cual no asigna el valor al argumento y f_1 la cual sí lo asigna. Para determinar el índice t como función de n y m:

Q es el conjunto I_q.

1. Se puede apreciar por inspección directa en la Figura 6.5 que:
   - Cada diagonal j tiene j tablas;
   - El índice t de cada tabla se incrementa sobre la diagonal de abajo hacia arriba y de derecha izquierda;
   - La tabla $n \times m$ se encuentra en la diagonal $diag(n, m) = n + m - 1$.
2. Se puede apreciar por razonamiento diagramático que:
   - El índice mayor sobre la diagonal j se define recursivamente como sigue: $max(0) = 0$ y $max(j) = max(j - 1) + j$;
   - $t = max(diag(n, m) - 1) + m$.

Por su parte, el cómputo del número de funciones de cada tabla requiere iterar desde la tabla 1 hasta la tabla $t - 1$. Para este efecto es necesario determinar el valor de n y m como función del número de tabla i, como sigue:

1. Sea $diag(i)$ la diagonal de la tabla $i = (n(i), m(i))$;
2. Se puede apreciar por razonamiento diagramático que:
   - La función $diag(i)$ se computa por minimización como sigue: $diag(0) = 0$ y $diag(i) = j$ tal que sea $j = 0$; incrementar j hasta que $max(j-1) < i \leq max(j)$;
   - $n(i) = diag(i) - m(i) + 1$;
   - $m(i) = i - max(diag(i) - 1)$.

El número absoluto N_f de la función f_k cuyos dominio y codominio tienen cardinalidades n y m es:

Definición 1 Número de una función para $n, m \geq 1$

$$N_f = N_k \; si \; t = 1;$$

$$y \; N_f = \sum_{i=1}^{t-1} (m(i) + 1)^{n(i)} + N_k \; si \; t > 1$$

Esta definición establece que todas las funciones con dominio y codominio finitos totales y parciales se pueden enumerar sin importar qué tan grandes sean n y m. Adicionalmente, cada función tiene un índice de longitud n en base $m+1$. Consecuentemente, para toda función en la enumeración es posible conocer su valor para todos sus argumentos por inspección directa de la tabla o del índice.

Esta enumeración es redundante. En particular todo conjunto de funciones F_{n,m_j} incluye todas las funciones en el conjunto F_{n,m_i} si $j > i$. Consecuentemente toda función tiene un conjunto numerable de índices, pero la enumeración es completa e incluye a todas las funciones con cardinalidades $n, m \geq 1$.

Asimismo, bajo las condiciones de interpretación estándar, que establecen que la cinta de memoria es infinita y que las computaciones se realizan inmediatamente, todas las funciones totales y parciales con dominio y codominio finitos se pueden computar por una Máquina de Turing; específicamente por la máquina que computa la Definición 1. Alternativamente, la función que enumera a todos las funciones totales y parciales con dominios y codominios finitos es recursiva.

La enumeración de las Máquinas de Turing debe incluir al menos una máquina para cada una de las funciones en la enumeración. Puede ser que haya más de una máquina que compute a la misma función y la enumeración puede ser redundante. Sin embargo, a diferencia de la tabla –que se evalúa por inspección directa– en el caso algorítmico no se puede saber el valor de la función para el argumento dado hasta que la máquina haya parado; y no se puede saber de an-

temano si una MT se detendrá para un argumento arbitrario por el problema del paro.

Es decir, hay dos vías para evaluar una función en la enumeración: (1) por inspección de la representación extensional, que es completa ya que se puede saber el valor de la función para todos los argumentos, para todas las funciones; y (2) mediante del cómputo de un algoritmo, que es incompleta debido al problema del paro.

La limitación del problema del paro se podría subsanar si fuera posible determinar el índice de la función que computa una MT arbitraria a partir de su descripción o viceversa, pero esto es poco plausible. Sin embargo, no es que no se pueda conocer el valor de la función para todos los argumentos, ya que esto se incluye en la enumeración, sino que no se puede obtener este conocimiento por la vía algorítmica. En este sentido la representación extensional de la tabla es más poderosa que la algorítmica o intensional, para las funciones en la enumeración.

La Tesis de Church establece que el conjunto de funciones computables es el conjunto de funciones que se pueden computar por la Máquina de Turing. Otras formulaciones de la tesis son que toda función computable tiene un algoritmo o que toda función computable es recursiva. Esta tesis se estipula normalmente estableciendo que el conjunto de funciones computables es el mismo que el conjunto de funciones que computa una u otra MT, y la evidencia que la apoya se deriva de la reducción recíproca o la correspondencia entre la MT y otros modelos computacionales de igual generalidad, como la Teoría de las Funciones Recursivas, entre otras formalizaciones suficientemente generales [16]. Esta tesis se probaría –y constituiría como un teorema– si se estableciera la correspondencia entre las MTs y el conjunto de todas las funciones.

La enumeración dada por la Definición 1 provee el subconjunto de funciones con dominios y codominios finitos. Consecuentemente la Tesis de Church implica que hay un algoritmo o una MT no trivial para cada función en la enumeración. En caso de que hubiera al menos una función en la enumeración para la que no hubiera un algoritmo no trivial, habría al menos una función para la que no habría un algoritmo y la Tesis de Church quedaría refutada.

Un algoritmo trivial consiste en definir una MT casuística, que asocie cada valor a su argumento correspondiente, reduciendo la definición intensional a la extensional. Por su parte, un algoritmo no trivial es un procedimiento genérico que dado el argumento produce el valor. La teoría de las funciones recursivas y el Cálculo-λ se ilustran normalmente mostrando las construcciones de las funciones básicas con un grado significativo de estructura, como la suma, la resta, la multiplicación, la división, la exponencial, la logarítmica, etc., pero además de estas funciones existen una infinidad de funciones con muy poca o nula estructura, como se puede apreciar analizando las funciones que pueden aparecer en las tablas. A la luz de la enumeración de la Definición 1 surge la duda de si realmente todas las funciones tienen un algoritmo no trivial.

En la teoría de la computabilidad, bajo las condiciones de interpretación estándar, se establece que todo modelo de computación que incluye al conjunto de funciones computadas por una MT es *Turing Completo*. La enumeración de las MTs difiere de la enumeración que establece la Definición 1 en que mientras que cada MT acepta un conjunto numerable de argumentos –es decir que el dominio es infinito, pero sus miembros se pueden poner en una lista o indezar por un número natural– el conjunto de argumentos de cada tabla o relación en el CRI es finito. En este sentido, el conjunto enumerado por la Definición 1 no es Turing Completo. En particular la función en la Definición 1 no se incluye

en la enumeración. Es también necesario considerar el caso en que el domino o el codominio no son numerables y, sin embargo, son computables, tales como los números reales. Se sabe, de acuerdo a los argumentos básicos de Cantor, que el conjunto de los reales no es numerable, ya que siempre es posible encontrar un número real que no está en una lista a partir de los reales que la constituyen, por medio de un argumento antidiagonal. Sin embargo, todo real puede estar en una lista u otra, y el dominio y el codominio de una función computable se constituyen cada uno por los elementos de una lista particular, entre el conjunto de listas posibles.

Asimismo, dado que n y m pueden ser tan grandes como se quiera, el conjunto de funciones en la enumeración se aproxima al conjunto de todas las funciones; por lo mismo, la distinción entre Turing Completo e Incompleto para este conjunto se diluye en gran medida. Siempre habrá una tabla para cualquier dominio y codominio que se requiera para efectuar una computación particular. Este no es un problema práctico ya que todas las computadoras digitales tienen registros aritméticos y de memoria finitos, y las funciones que se pueden evaluar por computadoras digitales tienen dominios y rangos finitos; consecuentemente, se incluyen en la enumeración dada por la Definición 1.

Sin embargo, sí constituye un problema teórico ya que habría que aclarar en qué sentido un número que no se puede representar porque es demasiado grande es computable, o cuál es la diferencia entre las funciones cuyo dominio se pueda ampliar tanto como se quiera y aquellas cuyo dominio es infinito.

Capítulo 7

Memoria Asociativa

Pasamos ahora a ilustrar el Cómputo Relacional Indeterminado (CRI). Para este efecto se presenta un sistema de memoria asociativa. La arquitectura de un mecanismo computacional incluye a la unidad de proceso propiamente, a la unidad de memoria, al sistema de control y a uno o varios canales de comunicación para relacionar dichos módulos. La Máquina de Turing se enfoca al cómputo de algoritmos y requiere una memoria para alojar la tabla de estados, además de la cinta en la que se expresan las configuraciones de entrada y salida, y en la que se manipulan los símbolos, pero el enfoque es el algoritmo, codificado en el control de estados, y la memoria juega un papel pasivo; aunque los recursos computacionales incluyen la capacidad de proceso y la memoria requerida, la MT no tiene una teoría de la memoria propiamente. Asimismo, las operaciones que manipulan tanto a los símbolos en la cinta como en la tabla de estados están completamente determinados y la máquina no es entrópica. Esta conceptualización de las memorias digitales se opone frontalmente a las memorias naturales, las cuales no sólo son repositorios pasivos de información sino módulos de proceso genuinos que conllevan una carga significativa del procesamiento de

la información. En este capítulo se presenta un modelo de memoria asociativa que utiliza el CRI y es entrópico. La memoria asociativa es también central en la arquitectura cognitiva y los procesos de memoria son esenciales a las formas de racionalidad más acabadas.

La memoria humana y de animales no humanos con un sistema nervioso suficientemente desarrollados es asociativa [45]. Una palabra, una imagen o un olor, a la que nos referimos como la llave o "la cue", puede iniciar una cadena de recuerdos que se asocian en base a sus contenidos o significados. Contrasta radicalmente con la memoria de las computadoras digitales, la cual se constituye por un conjunto de localidades, los llamados registros de memoria, que se accesan por su dirección. Las memorias digitales son como cómodas o cajoneras en cuyos cajones se almacenan pasivamente los contenidos y cada cajón se identifica con una referencia espacial –e.j., "el tercero de arriba a abajo"– o posiblemente con una etiqueta, y su contenido es independiente de las informaciones almacenadas en otras localidades. Asimismo, los contenidos se pueden referir lingüísticamente. Por estas razones, la memoria digital es *local* –en oposición a *distribuida*– y *declarativa* o *simbólica* –en oposición a *implícita* o *subsimbólica*. La memoria natural se opone también a la digital en que en esta última el recuerdo es una reproducción del registro original, como una fotografía, mientras que el recuerdo natural es una reconstrucción. La naturaleza constructiva de la memoria se conoce en el plano científico al menos desde los trabajos del psicólogo británico Frederic Bartlett [46], quien realizó una serie de estudios de la memoria de largo plazo en los que se presentaba una historia en forma textual o gráfica a los sujetos y se examinaban sus recuerdos en el corto, mediano y largo plazo, incluso hasta decenas de años. Los experimentos mostraron que los elementos contingentes de la historia se iban diluyendo con el tiempo pero su esencia se preservaba, y el

recuerdo era una reconstrucción muy abstracta de la historia original, que además estaba sujeta al contexto cultural y social. Estos estudios dejaron claro que la memoria humana es asociativa y constructiva. Asimismo, la memoria natural no es sólo un recipiente de información pasivo sino que juega un papel central en el pensamiento y la imaginación, y es un módulo central de la arquitectura cognitiva [47].

La creación de modelos de memoria asociativa en el paradigma de la computación simbólica se ha investigado principalmente en el contexto de las redes semánticas [48] y de las arquitecturas cognitivas orientadas a modelar las funciones superiores en el cerebro mediante sistemas de producción [23], pero éste es todavía un reto abierto de investigación. Esta limitación fue una de las motivaciones originales del programa de procesamiento distribuido y paralelo, incluyendo las redes neuronales y los sistemas conexionistas propuesto por Rumelhart y colaboradores a finales de los setenta y principios de los ochenta del siglo pasado [2], el cual cuestionó de manera explícita la capacidad de la Máquina de Turing de abordar el estudio de la memoria asociativa entre varias otras funciones cognitivas superiores.[1] El estudio de la memoria asociativa ha sido una temática central en las redes neuronales artificiales desde la *Lernmatrix* [49], el *Correlograph* [50] y el Asociador Lineal [51]; y desde muy temprano se investigó su capacidad de almacenamiento [52]. Esta línea de investigación tuvo un impulso muy significativo con la aparición de las memorias asociativas de Hopfield [53] y posteriormente con las memorias asociativas bidireccionales [54]. Los modelos iniciales eran muy limitados y se continuó su investigación con las llamadas memorias asociativas "sin pesos" [55, 56]. Posteriormente las memorias asociativas morfológicas [57, 58] y las memorias asociativas implicativas difusas [59, 60] pu-

[1] Ver la introducción de *Parallel Distributed Processing* [2].

sieron las bases para aplicaciones prácticas. Más recientemente, la definición de arquitecturas utilizando redes convolucionales profundas así como formulaciones más generales del modelo de Hopfield han tenido resultados satisfactorios con volúmenes de datos muy significativos [61].

En este paradigma los "recuerdos", llamados "patrones", se representan por matrices numéricas, y la memoria propiamente es una matriz que se construye en base a un conjunto de patrones individuales por medio de operaciones matriciales. Los patrones se representan en conjunto en dicha matriz pero no se pueden diferenciar individualmente, por lo que estas memorias son distribuidas y subsimbólicas. La recuperación se modela multiplicando la matriz que representa a la cue por la matriz que representa a la memoria, y el resultado de esta operación se toma como el argumento de una función, llamada de activación, la cual produce un patrón "de salida". Si éste es el mismo que la cue, la operación es exitosa y se recupera el recuerdo; en caso contrario, el patrón resultante se considera similar a la cue y se ingresa al proceso de manera recurrente. Este proceso se repite hasta que se reproduce la cue original, la cual corresponde a un patrón almacenado previamente. Las memorias de Hopfield tienen una función de energía la cual se ilustra normalmente como un valle con un pozo o sima por cada patrón almacenado; el proceso de recuperación se ilustra pensando al patrón a recuperar como un objeto que se lanza desde fuera del valle y "cae" sobre la ladera de una de las colinas y se desliza hacia el fondo en cada ciclo de iteración, como una piedra que rueda sobre la pendiente. La piedra puede caer sobre una ladera que no le corresponde y saltar a otra hasta encontrar la correcta. Cuando la cue llega al fondo el proceso "converge" al patrón representado por el pozo. Estos dispositivos son particularmente útiles cuando la cue se constituye por información incompleta, como una palabra pronunciada en un ambiente ruidoso,

o como un objeto visual que se encuentra ocluido parcialmente. En estos casos se requieren varios ciclos de iteración para retribuir al objeto, si éste se encuentra en la memoria. Sin embargo, como la cue converge a un patrón previamente almacenado, este tipo de memoria es también reproductiva en oposición a la memoria natural que es constructiva. Más aún, como las redes neuronales artificiales no pueden mantener representaciones estructuradas y, en consecuencia, tampoco información simbólica [62], las memorias asociativas creadas en este paradigma son más bien funciones de transferencia que mapean patrones de entrada a patrones de salida con fines de clasificación y predicción, entre otras tareas análogas, como se elabora más adelante en el Capítulo 8.

Una consecuencia adicional es que si la cue no se encuentra en la memoria se retribuye el patrón más parecido, y la memoria de Hopfield, al menos en el modelo básico, es incapaz de rechazar una cue. En esto se oponen frontalmente a la memoria natural que rechaza a la cue si el objeto que se busca no se ha incluido previamente. Por ejemplo, se sabe inmediatamente que no se conoce una palabra o una persona que no se nos ha presentado con anterioridad. Esta limitación se podría subsanar parcialmente por medio de una heurística que rechazara a la cue si no se lograra converger a un patrón almacenado en un número dado de iteraciones; sin embargo, la respuesta podría ser incorrecta debido al problema del paro, ya que no se podría saber si el patrón realmente no está contenido o simplemente hace falta buscar más. Adicionalmente, esta forma de negación requeriría un gran esfuerzo de cómputo y decir "no" tomaría mucho tiempo, lo cual se opone a la negación natural que es inmediata. Esta estrategia es una forma de negación por falla o débil, que surge de la incapacidad de probar el hecho, en este caso que el patrón se encuentra en la memoria. Consecuentemente, las memorias asociativas desarrolladas con redes neuronales asumen de forma

implícita la hipótesis del mundo cerrado. Ésta es una limitación significativa en relación a la memoria natural, a los lenguajes naturales y lógicos, y los sistemas de representación del conocimiento capaces de expresar la negación genuina o dura. La capacidad de saber que algo no se sabe de manera eficiente es muy útil para el individuo y la especie, y su carencia es una desventaja significativa para la supervivencia.

En resumen, la memoria natural se opone a la memoria digital estándar y a las memorias desarrolladas en el paradigma de las redes neuronales; mientras que la primera es asociativa, distribuida, constructiva, pero declarativa y capaz de rechazar cues directamente; la segunda es local, declarativa, reproductiva, pero no asociativa; y la tercera es asociativa, distribuida y subsimbólica, pero es incapaz de rechazar cues y retribuye el patrón más parecido, asume la hipótesis del mundo cerrado y, al igual que la memoria digital simbólica, es reproductiva.

En este capítulo se presenta una memoria asociativa basada en el Cómputo Relacional Indeterminado (CRI) que es distribuida, pero que al mismo tiempo es declarativa, como los modelos simbólicos; constructiva, en oposición a las memorias digitales, tanto simbólicas como de redes neuronales; y capaz de rechazar directamente cues de objetos que no están contenidos; consecuentemente, permite expresar la negación natural y asumir que el conocimiento es incompleto. El modelo se describe conforme a su presentación original [41].

7.1. Computación con Tablas

La implementación del modo de computación relacional indeterminado en el formato de tablas se denomina aquí *Computación con Tablas*. Este modo permite la definición de registros de memoria asociativos, cada uno de los cuales

contiene la representación distribuida de una unidad básica de contenido o un concepto básico.

Sea $[R_k]^t$ el contenido del registro R_k en el tiempo t y $\leftarrow$ el operador de asignación tal que $R_k \leftarrow R_j$ asigna $[R_j]^t$ a $[R_k]^{t+1}$, donde j y k pueden ser el mismo. Este operador corresponde a la asignación de valores entre registros de los lenguajes de programación imperativos. La máquina incluye también el operador condicional *si* que relaciona una condición *pred* con las operaciones op_1 y op_2 – i.e., *si pred* entonces op_1 alternativamente op_2, donde la cláusula *alternativamente* es opcional.

La inicialización de un registro R tal que se asigna el valor 0 o 1 a todas sus celdas se denota $R \leftarrow 0$ o $R \leftarrow 1$ respectivamente, y $R \leftarrow f_i$ denota que la función f_i se asigna o escribe en el registro R, donde el valor v_j del argumento a_i se expresa como un 1 en el renglón j de la columna i, para todos los valores de todos los argumentos. El sistema incluye también los operadores λ, η y β para computar las operaciones entre registros correspondientes, así como un generador de números aleatorios para implementar la operación de reducción. Éstas son todas las operaciones de computación por tablas.

Sea R_k un *Registro de Memoria Asociativa* (RMA) de cardinalidad $n \times m$ cuyo contenido es la representación distribuida de un objeto de la clase K, y O_k el conjunto de objetos de la clase K. Sea F_O un conjunto de funciones de cardinalidad $n \times m$ tal que $f_i \in F_O$ representa al objeto $o_i \in O_k$. Cada función f_i representa una instancia concreta de la clase K constituida por n características, cada una de las cuales tiene uno de m valores posibles. Si la función f_i es parcial hay características que no tienen valor.

Sea R_k un RMA y $R_{k-i/o}$ un registro auxiliar de entrada y salida, ambos de cardinalidad $n \times m$. La representación distribuida del concepto K se crea mediante el algoritmo *Registrar*(f_i, R_k) como sigue:

- *Registrar*(f_i, R_k):
  1. $R_{k-i/o} \leftarrow f_i$
  2. $R_k \leftarrow \lambda(R_k, R_{k-i/o})$
  3. $R_{k-i/o} \leftarrow 0$

El reconocimiento de un objeto $o_i \in O$ representado por la función $f_i \in F_O$ de la clase K, cuya representación distribuida se almacena en R_k, se lleva a cabo por el algoritmo *Reconocer* como sigue:

- *Reconocer*(f_i, R_k):
  1. $R_{k-i/o} \leftarrow f_i$
  2. Si $\eta(R_{k-i/o}, R_k)$ entonces $(R_{k-i/o} \leftarrow 1)$

El recuerdo o recuperación de un objeto almacenado en la memoria se lleva a cabo por la operación de reducción. El argumento de esa función es la función *cue* del recuerdo. Si ésta se incluye en el registro de memoria la reducción produce un objeto de la clase de manera aleatoria. Esta operación se lleva a cabo por el algoritmo *Recuperar* como sigue:

- *Recuperar*(f_i, R_k):
  1. $R_{k-i/o} \leftarrow f_i$
  2. Si $\eta(R_{k-i/o}, R_k)$ entonces $R_{k-i/o} \leftarrow \beta(R_{k-i/o}, R_k)$ alternativamente $R_{k-i/o} \leftarrow 0$

La configuración estándar de entrada de esta máquina especifica que el registro auxiliar contiene la función a registrar o reconocer, o la cue de una operación de recuperación, en el estado inicial de la operación de memoria correspondiente. Esta condición se garantiza por el paso (1) en los tres algoritmos. La configuración estándar de salida especifica el contenido del registro auxiliar al término de las tres operaciones. Éste es cero cuando concluye la operación de registro exitosamente y cuando la cue es rechazada en la operación de recuperación –es decir cuando la cue no está en la memoria; por su parte, si las operaciones de reconocimiento y recuperación son exitosas el registro auxiliar se llena de unos o contiene la función que representa al objeto retribuido, respectivamente. Las convenciones de interpretación estipulan que el contenido de un registro asociativo es un concepto abstracto, y el contenido del registro auxiliar es el concepto de una entidad concreta o el concepto de un objeto individual. La función de entrada f_i se puede generalizar a una relación R_i directamente en las tres operaciones, y el objeto registrado o reconocido, o la cue en una operación de recuperación, puede ser también un concepto abstracto. Las definiciones de la abstracción e inclusión se aplican directamente a conceptos abstractos. La reducción para este caso se define como sigue: si la relación R_a se incluye en la relación R_f, la reducción $\beta(R_a, R_f)$ es la relación R_v tal que cada valor $v_i \in R_f(a_i)$ se considera aleatoriamente para incluirse en $R_v(a_i)$ de acuerdo a alguna distribución apropiada, para todos los argumentos a_i. El compromiso de la entropía se puede apreciar en la operación de recuperación. Si la entropía del registro asociativo es muy baja, la función –o el concepto– retribuido será muy similar a la cue, pero ésta deberá ser muy precisa para que se acepte. En el límite, si la entropía es cero, la cue, la función almacenada y la función recuperada son la misma. En el otro extremo, si la entropía es muy alta, la cue se aceptará muy fácilmente pero la función re-

cuperada puede ser muy diferente. En el límite, si todos las celdas del registro tienen el valor 1 la cue se aceptará siempre pero la función que se recupera será completamente aleatoria y su semejanza a la cue puede ser nula. Sin embargo, si la entropía tiene un valor moderado, habrá una buena posibilidad de que la cue se acepte y que la función que se recupere sea suficientemente similar. Los procedimientos de memoria *Registrar*, *Reconocer* y *Recuperar* son muy sencillos y se designan aquí como *Algoritmos Mínimos*. El poder de las representaciones distribuidas viene de su computación local pero simultánea y masiva en todas las celdas de cada registro y en todos los registros que constituyan a la memoria asociativa.

7.2. Arquitectura de la Memoria

Las operaciones de registro, reconocimiento y recuperación se implementan en una arquitectura paralela, la cual se ilustra en la Figura 7.1. Ésta consta del módulo *Análisis* que transforma las representaciones "retinotópicas" –que se generan directamente a partir de las observaciones y se registran en el *Buffer Modal*– a un conjunto finito de características con sus respectivos valores. Estas representaciones se manipulan en los registros de memoria asociativa a través de un bus, y el resultado de dichas operaciones se presenta al módulo *Síntesis*, cuya función es transformarlas en una representación concreta, en una modalidad específica, a partir de la cual se genera la acción. De forma más puntual, el módulo de análisis presenta a la memoria cada observación sensada en el medio externo como una función finita f_i que representa al objeto o_i. Asimismo, las funciones recuperadas de la memoria se presentan al módulo de síntesis y éste las convierte

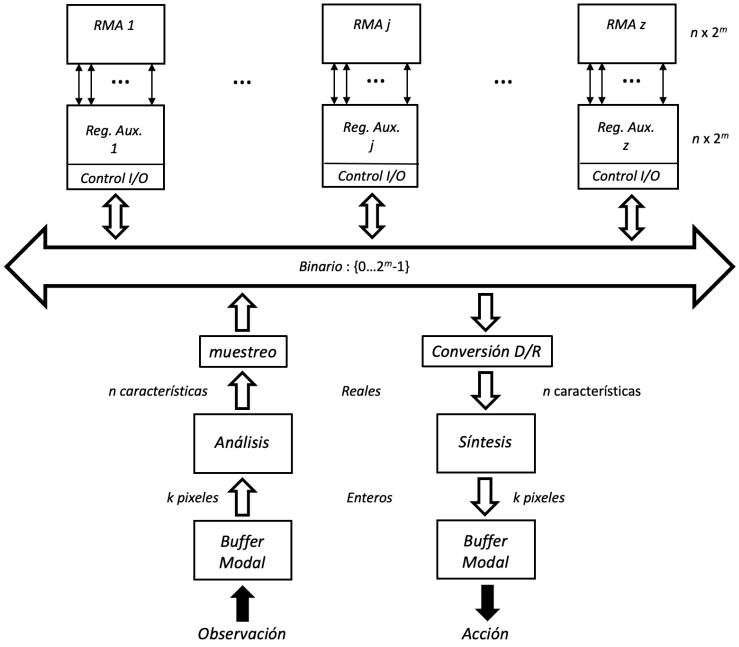


Figura 7.1: Arquitectura de la memoria asociativa

en un conjunto de características concretas que se expresan en un buffer modal de salida.

El módulo de análisis tiene una entrada por cada pixel del buffer modal, cuyo valor es, por ejemplo, el nivel de gris del pixel correspondiente. En la figura se ilustra un buffer con k pixeles, cada uno de los cuales tiene un número finito de valores discretos –por ejemplo 256 tonos de gris que van de 0 a 255. La salida es un vector o función con n argumentos o caractererísticas con sus respectivos valores. Los valores de los argumentos se muestrean en 2^m niveles –que corresponden a los niveles de los RMAs– por el módulo *Muestreo*, para su representación y manipulación en las operaciones de registro, reconocimiento y recuperación. La información se ingresa al sistema de memoria vía un bus con

n pistas –una por cada argumento a_i– y cada pista consta a su vez de m líneas en las que se representa el valor v_j del argumento a_i en forma binaria tal que $0 \leq v_j \leq 2^m - 1$. El valor se representa con m bits. La memoria cuenta a su vez con un número arbitrario –pero finito– de registros de memoria asociativa de cardinalidad $n \times 2^m$ con sus respectivos registros auxiliares de entrada y salida, donde cada registro almacena la representación distribuida de un objeto individual. Los registros de memoria están conectados al bus a través de sus respectivos registros auxiliares, así como por un control de entrada y salida que mapea el valor de la característica representada de forma binaria en el bus al renglón correspondiente en el registro auxiliar para las operaciones de entrada y viceversa para las de salida. El número binario k en el bus selecciona el renglón $k + 1$ del registro auxiliar para tomar en cuenta a la función vacía. La salida de los registros de memoria se presenta al bus como una función de n argumentos, con sus respectivos valores binarios en el rango de muestreo. Esta información alimenta al módulo de síntesis como una función real –el módulo *Conversión D/R* transforma los valores de cada argumento en el bus de notación binaria a real. La salida de dicho módulo es una función que representa a la imagen de salida en la modalidad específica. Por ejemplo, la imagen recuperada de un registro de memoria asociativa como un conjunto de características abstractas se representa de manera concreta en el buffer de salida. Adicionalmente, el sistema cuenta con una unidad de control –no mostrada en la figura– que envía a los registros de memoria asociativa su estado –activo o inhibido– y la operación a realizar, y cada registro envía una señal al control indicando el estatus final de la operación –exitosa o no– así como su entropía. La operaciones centrales de los algoritmos de registro, reconocimiento y recuperación se llevan a cabo directamente entre los registros de memoria asociativa y auxiliares en dos o tres pasos

de cómputo –i.e., las operaciones λ, η y β más las asignaciones correspondientes. Las asignaciones $R_{k-i/o} \leftarrow f_i$ y $R_{k-i/o} \leftarrow 0$ se llevan a cabo asimismo en un sólo paso de computación. Además de los procesos centrales se requiere llevar a cabo los protocolos de entrada y salida, como sigue:

- *Protocolo de entrada*:

    1. Sensar el objeto o_i del medio externo y escribirlo en el buffer modal;
    2. Producir su representación f_i con valores reales mediante el módulo de análisis;
    3. Producir su representación binaria mediante el módulo de muestreo y escribirla en el bus;
    4. Escribir la función en forma diagramática en el registro auxiliar $R_{k-i/o}$.

- *Protocolo de salida*:

    1. Escribir el valor de $R_{k-i/o}$ en el bus en forma binaria;
    2. Convertir el contenido del bus a números reales con el módulo de conversión D/R;
    3. Convertir la representación abstracta que se obtiene de la memoria en una representación concreta en el buffer de salida mediante el módulo de síntesis.

Con estos protocolos, las operaciones de memoria se llevan a cabo directamente, como sigue:

- *Registrar*(f_i, R_k): Se activa el registro R_k y se inhiben el resto de los registros de memoria; se ejecutan el protocolo de entrada y la operación de registro propiamente.

- *Reconocer*(f_i, R_k): Se activan todos los registros; se ejecuta el protocolo de entrada; se lleva a cabo la prueba de reconocimiento; los registros envían al control el estatus de la operación (exitosa o no); en caso de que el reconocimiento haya sido exitoso para más de un registro, el control selecciona el registro con menos entropía.

- *Recuperar*(f_i, R_k): Se activan todos los registros; se ejecuta el protocolo de entrada; se lleva a cabo la operación de recuperación; los registros envían al control el estatus de la operación; en caso de que la operación haya sido exitosa para más de un registro, el control selecciona el registro con menor entropía; se inhiben todos los registros excepto el seleccionado por el control; se lleva a cabo el protocolo de salida; el objeto recuperado queda en el buffer de salida.

7.3. Análisis y Síntesis

El módulo de análisis tiene por función transformar la información de carácter concreto que se sensa del medio externo y se expresa en el buffer de entrada –donde las características representan señales externas– a una representación abstracta que identifica a las entidades de una clase. Ambas se expresan como funciones aunque en el primer caso los argumentos tienen una interpretación espacial y corresponden a la configuración del entorno –como los pixeles en una imagen digital– mientras que en el segundo las características son independientes de la modalidad. El módulo de análisis se implementó con una red neuronal con tres capas convolucionales, que se conoce como "codificador" o *encoder*[63], como se ilustra en la Figura 7.2. El módulo de síntesis se implementó como una red neuronal transpuesta de dos capas, llamada "decodifica-

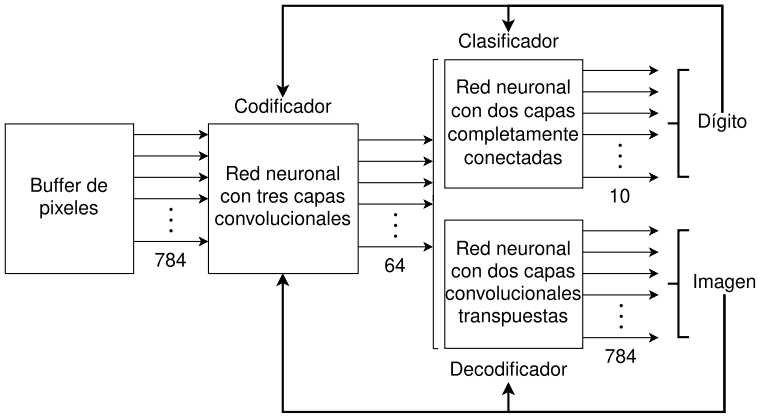


Figura 7.2: Entrenamiento del módulo de análisis

dor" o *decoder*, la cual mapea conjuntos de características a sus imágenes respectivas. El codificador y el decodificador tomados en conjunto constituyen un "autocodificador" o *autoencoder*[64, 65]. La fase de entrenamiento se llevó a cabo añadiendo un clasificador constituido por una red neuronal completamente conectada (RNCC) de dos capas, la cual mapea conjuntos de características a sus dígitos correspondientes. El diagrama muestra el caso en que el encoder tiene 784 características de entrada que corresponden al buffer de tamaño 28×28 pixeles, cada uno tomando un valor entre 256 niveles de gris, y su salida es una función de 64 argumentos con valores reales. El codificador, el decodificador y el clasificador se entrenaron conjuntamente de manera supervisada utilizando propagación hacia atrás. El clasificador se retira una vez que el autocodificador se entrena, y el codificador y el decodificador se utilizan como los módulos de análisis y síntesis directamente. La salida de los RMAs se alimenta al módulo de síntesis cuya salida se escribe directamente en el buffer modal respectivo.

7.4. Una memoria visual para dígitos manuscritos

La computación con tablas y la memoria asociativa se utilizaron para almacenar, reconocer y retribuir dígitos manuscritos del 0 al 9. Para este efecto se utilizó la base de datos MNIST.[2] Cada dígito en este recurso se expresa como un buffer o arreglo de 28×28 pixeles con 256 niveles de gris. Este conjunto de datos cuenta con 70, 000 instancias y es balanceado; es decir, tiene aproximadamente 7, 000 muestras de cada dígito. Este recurso es ampliamente conocido y utilizado para aprender a utilizar y aplicar la tecnología de redes neuronales profundas, y es posible clasificar los dígitos con una precisión y cobertura muy altas. Existen también versiones de MNIST contaminadas con ruido [66] y se ha intentado hacer la clasificación ocluyendo los dígitos, aunque los resultados en esta última tarea no son tan satisfactorios; sin embargo, el uso de otras técnicas, como los llamados modelos composicionales, permite mejorar el reconocimiento [67]. Los dígitos se pueden reproducir con autocodificadores así como con las llamadas redes generativas adversariales; sin embargo, todas estas metodologías siempre generan un objeto a pesar de que no sea posible distinguir el patrón de entrada, ya sea por el nivel de ruido o el tamaño de la oclusión; como ya se ha dicho, las redes neuronales aproximan la entrada al objeto más parecido de la clase, ya que no tiene una forma natural de rechazar la entrada, y el objeto generado es más bien un prototipo de la clase que una reconstrucción de la entrada. Adicionalmente, si la información de entrada es pobre, la precisión del sistema se degrada y es posible que se generen objetos de clases incorrectas.

MNIST se ha utilizado también para evaluar el desempeño de la memoria de Hofpfiel utilizado como clasificador en relación a un clasificador estándar construido con redes neuronales [68], y también para modelar memorias asociativas

[2] http://yann.lecun.com/exdb/mnist/

con una orientación biológica [69]; sin embargo, este conjunto de datos no se ha utilizado para modelar memorias asociativas con fines de almacenamiento y de recuperación de objetos, en particular cuando la cue se ocluye de manera significativa. En general, la clasificación de este conjunto de datos con un enfoque constructivo y no sólo reproductivo tampoco se ha estudiado con este conjunto de datos. El corpus se dividió en tres conjuntos disjuntos:

- Corpus de entrenamiento (*TrainCorpus*): Para el entrenamiento de las redes convolucional y transpuesta de los módulos de análisis y síntesis (57 %).
- Corpus de recuerdo (*RemCorpus*): Para la creación o llenado de los registros de memoria (33 %).
- Corpus de prueba (*TestCorpus*): Para probar el sistema de memoria de manera integral (10 %).

Para efectos de eliminar contingencias posibles debido a la selección de los datos se llevó a cabo un procedimiento de validación cruzada; el corpus se dividió en 10 partes y los experimentos se llevaron a cabo el mismo número de veces, rotando las particiones en cada iteración. En total se llevaron a cabo cinco experimentos como sigue:

1. Experimento 1: Se definió un sistema de memoria asociativa incluyendo un RMA para almacenar la representación distribuida de cada uno de los diez dígitos. Se determinó la precisión y la cobertura de los RMAs de forma individual y del sistema integrado. Se identificó el tamaño de los registros con un desempeño satisfactorio.

2. Experimento 2: Se investigó si los RMAs pueden contener representaciones distribuidas de objetos de diferentes clases. Para este efecto se incluyó un RMA para almacenar la representación distribuida de dos dígitos traslapados. Se determinó la precisión y la cobertura de cada RMA así como del sistema integrado.

3. Experimento 3: Se determinó la precisión y la cobertura de RMAs con niveles diferentes de entropía, para registros con parámetros operativos satisfactorios, los cuales se identificaron en el experimento 1.

4. Experimento 4: Se retribuyeron objetos de la memoria a diferentes niveles de entropía y se sintetizaron sus imágenes visuales; para este efecto se utilizaron los mismos RMAs del experimento 3. Se evaluó cualitativamente la similitud entre la cue y los objetos recuperados a diferentes niveles de entropía.

5. Experimento 5: Se investigó la recuperación de dígitos a partir de cues de objetos ocluidos significativamente, y se evaluó la calidad de las imágenes generadas; se determinó la precisión y la cobertura de esta operación.

El buffer de pixeles se representa por una función de 784 argumentos (i.e., 28×28 pixeles) cuyos valores son niveles de gris correspondientes y la salida del módulo perceptual es una función f_k con dominio $\{a_1, ..., a_{64}\}$ donde cada argumento a_i tiene un valor real, como se ilustra en la Figura 7.2. Intuitivamente, el proceso de análisis corresponde a "ver" el dígito, procesarlo de abajo hacia arriba y escribir la salida en el registro auxiliar. El módulo de análisis presenta los dígitos en el buffer de pixeles al registro auxiliar de la memoria asociativa en las operaciones de registro, reconocimiento y recuperación, de acuerdo con

la configuración estándar de entrada. Al término de estas operaciones el registro auxiliar presenta la información en la configuración estándar de salida. Los registros auxiliar y de memoria asociativa se construyeron para 2^m niveles y, consecuentemente, su cardinalidad es 64×2^m. El parámetro m determina la granularidad del registro, y cada experimento se llevó a cabo con granularidades de 2^m para $0 <= m <= 9$. El contenido de los registros de memoria asociativa se creó con el algoritmo *Registrar*. Para este efecto se utilizó el corpus *RemCorpus* y las evaluaciones del reconocimiento y la recuperación se llevaron a cabo con el *TestCorpus*. Es decir, el corpus *TrainCorpus* sólo se utilizó para entrenar a los módulos de análisis y síntesis, y los datos utilizados en los experimentos de memoria son independientes de los utilizados para entrenar a las redes convolucionales. Asimismo, las pruebas de reconocimiento y recuperación se llevaron a cabo de forma independiente.[3] Para efectos del reconocimiento, cada instancia debe aceptarse por el registro de memoria asociativa correspondiente y rechazarse por el resto; sin embargo, es posible que un dígito se reconozca por más de un registro, con el subsecuente decremento de la precisión. También es posible que un dígito se rechace por el registro que le corresponde, con el consecuente decremento de la cobertura. Si una imagen se rechaza por todos los registros no se reconoce como un dígito.

7.4.1. Experimento 1

El propósito de este experimento fue determinar las características de RMAs de tamaño 64×2^m para $0 \leq m \leq 9$:

[3]Los datos y el código para replicar los experimentos, incluyendo los resultados detallados así como el *hardware* utilizado, están disponibles en `https://github.com/LA-Pineda/Associative-Memory-Experiments`.

1. Registrar la totalidad del corpus *RemCorpus* en sus registros correspondientes mediante la operación *Registrar*;

2. Probar el desempeño del reconocimiento utilizando todas las instancias del *TestCorpus* con la operación *Reconocer*;

3. Calcular el promedio de la precisión y cobertura del reconocimiento, así como la entropía de los RMAs.

4. Seleccionar un objeto único para ser recuperado por la operación *Recuperar*; determinar la precisión y cobertura de la operación de reconocimiento para el sistema de forma integral, una vez que se decide cuál es el objeto seleccionado.

En la Figura 7.3 (a) se muestran la precisión, la cobertura y la entropía de los diez RMAs. La precisión para el RMA con un sólo renglón es del 10 % y corresponde a la proporción de los datos de prueba para cada dígito, y la cobertura es del 100 % ya que toda la información se confunde en el mismo renglón y todo se acepta. La precisión crece con el número de renglones de los RMAs y llega a un valor satisfactorio a partir de los 32 renglones. La cobertura, por su parte, permanece muy alta hasta que la granularidad de la tabla es muy fina donde empieza a decrecer moderadamente. La entropía aumenta casi de forma lineal con el tamaño del RMA, empezando desde cero para un sólo renglón donde todas las características tienen a lo más un sólo valor, que corresponde al único nivel de muestreo.

La precisión, cobertura y entropía promedio del reconocimiento del sistema integrado se muestra en la Figura 7.3 (b). La precisión es similar al caso anterior pero la cobertura se reduce significativamente en los RMSs con un valor pequeño de m –la precisión y la cobertura tienen un valor prácticamente igual para

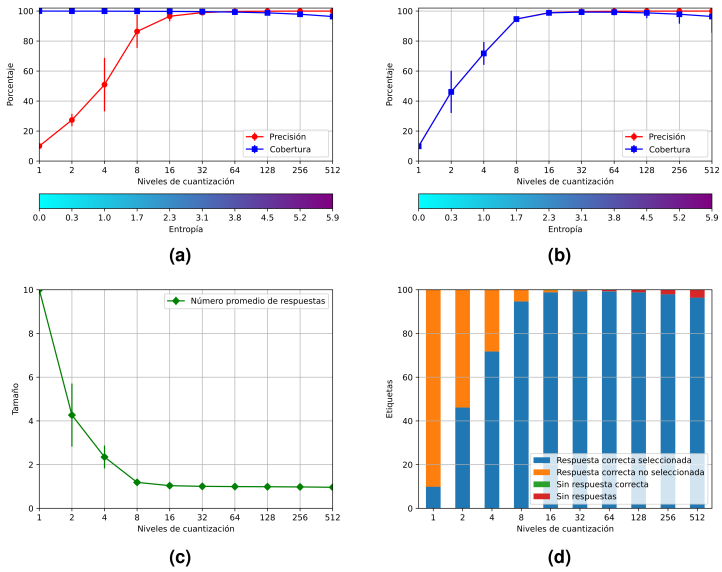


Figura 7.3: Resultados del Experimento 1

$m \leq 4$. La cobertura disminuye porque para m muy bajas hay muchos falsos positivos y varios de los registros aceptan a la misma cue, por lo que la selección del dígito correcto no se determina completamente. Para resolver este problema se exploró seleccionar el registro con menor entropía pero no hubo mejora respecto a una selección aleatoria con distribución normal. Sin embargo, a partir de $m = 5$, es decir, 32 niveles, la cue se acepta por un sólo RMA o por ninguno, en cuyo caso se rechaza. Este resultado se muestra claramente en la Figura 7.3 (c) donde se ve claramente que los diez RMAs aceptan la cue si hay un sólo nivel de muestreo, pero se reducen a prácticamente uno cuando hay 8 o 16 niveles, y a uno estrictamente a partir de 32 niveles. Este efecto se ilustra adicionalmente en la Figura 7.3 (d) en la que se ilustra el promedio de las cuatro respuestas po-

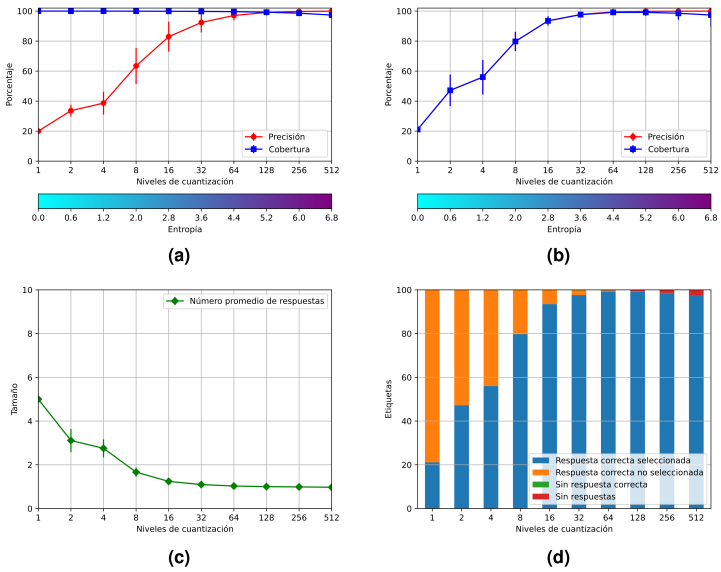


Figura 7.4: Resultados del Experimento 2

sibles de un RMA: i) se escoge la respuesta correcta; ii) no se escoge la respuesta correcta; iii) se escoge la respuesta incorrecta; y iv) no hay respuesta.

7.4.2. Experimento 2

En este experimento cada RMA contiene la representación distribuida de dos dígitos diferentes: “0” & “1”; “2” & “3”; “4” &“5”; “6” & “7”; y “8” &“9”. El procedimiento es el mismo que en el experimento 1. Los resultados son análogos, excepto que la entropía de los registros que contienen dos dígitos se incrementa respecto a los registros que contienen uno sólo, como se muestra en la Figura 7.4. Este experimento muestra que es posible crear memorias asociativas con las representaciones distribuidas “traslapadas” con más de un objeto, con un desempeño satisfactorio.

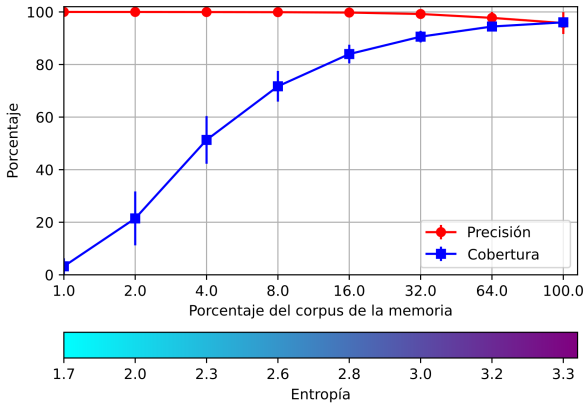


Figura 7.5: Resultados del Experimento 3

7.4.3. Experimento 3

El propósito de este experimento fue investigar el desempeño de RMAs con características de operación satisfactorias en relación a su entropía o contenido de información. El experimento 1 mostró que los registros de tamaño 64×32 y 64×64 satisfacen este requerimiento; como su desempeño es prácticamente el mismo se escogió el menor por consideraciones de economía.

Los RMAs se llenaron con niveles de información diferentes del corpus *RemCorpus* –1 %, 2 %, 4 %, 8 %, 16 %, 32 %, 64 % y 100 %– como se muestra en la Figura 7.5. La entropía se incrementa de acuerdo a la cantidad de datos en la memoria conforme a lo esperado. La precisión es muy alta para valores muy bajos de entropía y disminuye ligeramente conforme al incremento de este parámetro, pero permanece bastante alta cuando se incluye todo el *RemCorpus*. La cobertura, por su parte, es muy baja para valores muy bajos de entropía pero se incrementa rápidamente conforme los RMAs se llenan con más datos.

Figura 7.6: Similitud entre la cue y el objeto recuperado como función de la entropía

7.4.4. Experimento 4

Este experimento consistió en evaluar la similitud entre los objetos recuperados de la memoria con sus respectivas cues. En el escenario básico, si la cue corresponde específicamente al objeto recuperado, la imagen en el buffer de salida debería ser exactamente la misma que la imagen en el buffer de entrada. Sin embargo, la recuperación de memoria produce un objeto un poco diferente que la cue debido a la naturaleza aleatoria de la operación β. En el experimento se utilizó una distribución aleatoria triangular centrada en los valores de la función para todos los argumentos que representa a dicha llave, los cuales se seleccionaron entre todos los valores posibles para cada argumento del RMA. La hipótesis es que el incremento de la entropía va de la mando con la cobertura en la operación de recuperación, pero la indeterminación impacta de forma negativa en la precisión y en la similitud entre el objeto recuperado y la cue. La Figura 7.6 ilustra una instancia de cada tipo de dígito retribuida a niveles diferentes de entropía, donde cada columna muestra la imagen recuperada y reconstruida con la misma cue. En esta figura se muestran cues que se aceptan a todos los niveles.

El primer renglón contiene la cue para la operación de recuperación; el segundo la imagen producida directamente por el autocodificador. Esta imagen debería ser la misma que se produciría si la operación β se especificara como la función identidad; sin embargo, en dicho caso las operaciones de reconocimiento y recuperación no se distinguirían, y la recuperación sería reproductiva en vez de constructiva. La imagen generada y la cue no son exactamente la misma, pero esto se debe a que el decodificador computa tan sólo una aproximación de la función inversa del codificador. El resto de las imágenes de arriba hacia abajo corresponden a los objetos recuperados en los nueve niveles de entropía considerados del *RemCorpus* (la imagen codificada corresponde a $e = 0$). Los RMAs se utilizan dentro de su rango operativo y la calidad de las imágenes es en general muy buena en todos los niveles, aunque se puede apreciar que empieza a decrecer cuando la entropía sube alrededor de 3.0 para algunos de los dígitos. La Figura 7.7 ilustra una instancia de cada dígito que se rechaza por la operación η, por lo que la imagen de salida no se produce y la celda respectiva se muestra en blanco. La gráfica de la izquierda muestra instancias que se rechazan a niveles bajos de entropía pero que se aceptan a niveles altos, en cuyo caso se genera la imagen de salida. En la figura se puede apreciar que aunque la calidad de los dígitos es satisfactoria a niveles moderados de entropía, empieza a decrecer a niveles altos. La figura de la derecha, por su parte, muestra instancias que se rechazan a todos los niveles de entropía. Éstos son casos extremos de sus respectivas clases, y son confusos incluso para los seres humanos, especialmente si se muestran fuera de contexto. Por ejemplo, el "8" se parece a la letra griega γ, y el '9" se podría interpretar como la letra *a*. Sin embargo, el autocodificador genera símbolos de salida en ambos casos, es decir el 7 y el 0 respectivamente, aunque esto es claramente un error. Ésta es una limitación significativa de los clasificadores

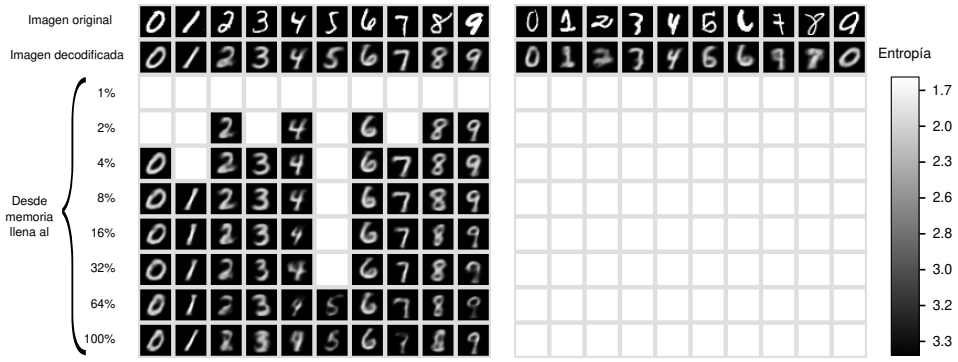


Figura 7.7: Recuperación de memoria con rechazo

que no tienen una noción clara de rechazo y que siempre aproximan al objeto más parecido, de acuerdo a una medida implícita y abstracta de similitud. Dado que la imagen de entrada no se puede apreciar plenamente la salida es más bien una imágen genérica o prototípica de la clase de objetos que se asigna a la cue, y consecuentemente no hay una forma clara de evaluar qué tan similares son los objetos de entrada y salida. Más aún, el objeto generado puede ser un falso positivo de la clase de salida.

7.4.5. Experimento 5

El propósito de este experimento fue ilustrar el desempeño del sistema de memoria en el reconocimiento, la retribución y la generación de objetos ocluidos significativamente. La recuperación de patrones con información incompleta, como ruido u oclusiones, ha sido una motivación muy importante en el desarrollo de memorias asociativas en el paradigma de las redes neuronales desde las propuestas iniciales. Aquí se aborda esta problemática desde la perspectiva del presente modelo. En este experimento se utilizó el *RemCorpus* para rellenar los RMAs y se entrenó el autocodificador como en los experimentos anteriores,

pero los objetos del *TestCorpus* se ocluyeron significativamente con rectángulos opacos cubriendo el 50 % de la parte superior e inferior de los dígitos –como se muestra en las figuras 7.8 y 7.9 respectivamente; y con barras verticales y horizontales de 4 pixeles de ancho en el buffer de entrada –figuras 7.10 y 7.11. En la columna de la izquierda de dichas gráficas se muestra la precisión, la cobertura y los niveles de entropía correspondientes. Las celdas en blanco muestran que el dígito respectivo se rechaza al igual que en la Figura 7.7.

La expectativa es que la información incompleta en la entrada aumente la ambigüedad de las imágenes y que las cues se acepten más fácilmente, con el consecuente incremento de la cobertura, pero que la precisión disminuya de acuerdo al tamaño de la oclusión, aunque si la información en la entrada es muy pobre las cues se rechazan directamente. Este efecto se aprecia en las gráficas y en las imágenes con bajas entropías en las figuras 7.8, 7.9, 7.10 y 7.11. Sin embargo, el autocodificador siempre genera una imagen de salida, a pesar de que la imagen original sea prácticamente ilegible, como se aprecia en el segundo renglón en todas las imágenes de la columna derecha.

El algoritmo *Reconocer* –la operación η– es muy estricto, ya que para que un dígito se acepte se deben aceptar el 100 % de sus características abstractas y sólo se requiere una característica de la entrada que no esté en el RMA para que el dígito se rechace. Por esta razón la cobertura del sistema es muy baja si la cue es muy pobre. La operación de reconocimiento se puede relajar si se permite que algunas características de la cue fallen la prueba de inclusión. El número de características en la configuración de los experimentos es 64 y los efectos de relajar 1, 2 y 3 características –1.6 %, 3.1 % y 4.7 %– se muestran en el segundo, tercero y cuarto nivel de las figuras 7.8 a 7.11. La relajación aumenta la cobertura pero impacta en la precisión, que disminuye, como se puede apreciar en las figuras, y

la recuperación de memoria puede producir más de una hipótesis de interpretación, algunas de las cuales son en este caso incorrectas. Este efecto se aprecia de forma más clara cuando se relajan tres características.

La funcionalidad básica de la memoria es conservadora ya que las cues se rechazan estrictamente: si las imágenes no pertenecen a la clase se deben rechazar. Sin embargo, cuando la entrada es muy pobre ya sea por ruido o por disfunciones de los órganos o dispositivos sensoriales, no se sabe de antemano la forma del objeto ni a qué clase pertenece, y relajar el reconocimiento y la retribución de memoria puede ser una conducta productiva. La relajación aumenta la flexibilidad de la memoria y puede dar lugar a un buen compromiso para la interpretación de imágenes incompletas. Como se puede apreciar las imágenes se rechazan a bajos niveles de entropía pero se pueden retribuir imágenes útiles cuando la entropía es moderada, aunque hay ambigüedad y se producen imágenes incorrectas. En general la recuperación de memoria obedece al compromiso de la entropía. Si la entropía es muy baja la memoria puede ser muy inflexible pero si sube demasiado los objetos recuperados de la memoria pueden ser muy arbitrarios. Sin embargo, a niveles moderados de entropía la memoria puede ser flexible y admitir diferentes hipótesis de interpretación que se puede resolver en un contexto de interpretación más amplio.

7.4.6. Configuración experimental

La simulación se programó con Python 3.8 sobre la distribución the Anaconda. Las redes neuronales se implementaron con TensorFlow 2.3.0 y la mayoría de las gráficas con Matplotlib. Para correr los experimentos se utilizó una máquina Alienware Aurora R5 con un procesador Intel Core i7-6700, 16 GBytes de RAM y una tarjeta gráfica o GPU NVIDIA GeForce GTX 1080.

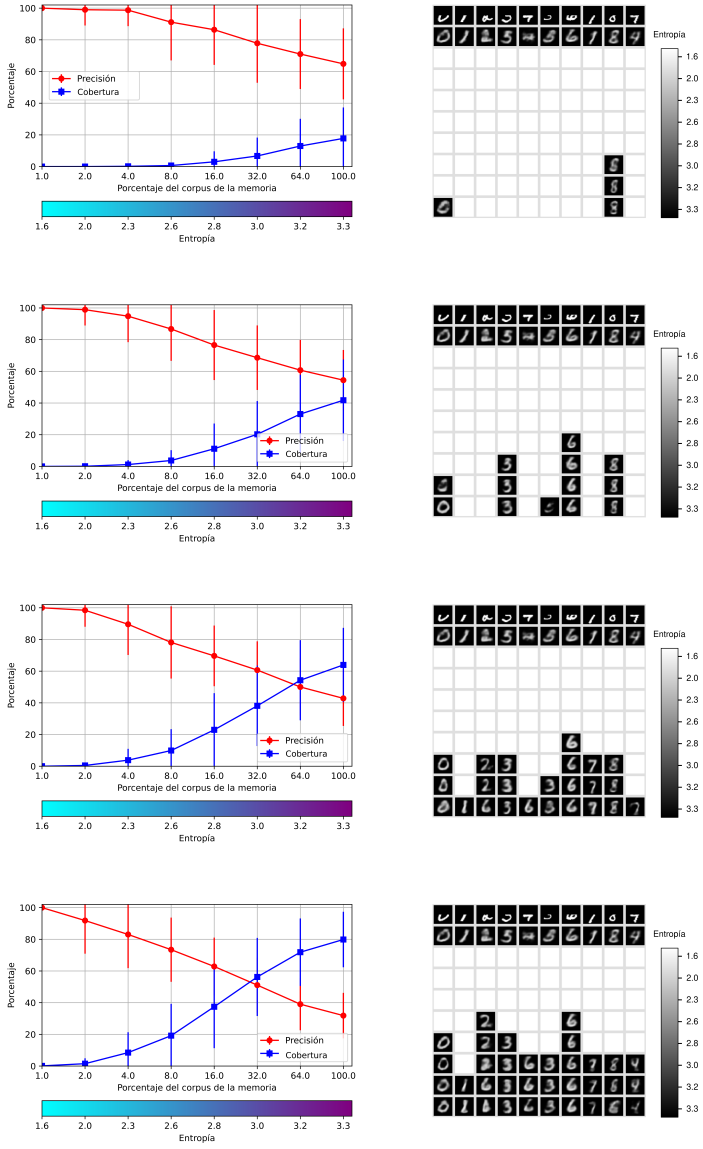


Figura 7.8: Recuperación de memoria con oclusiones del 50 % en la parte superior, con relajación de 1, 2 y 3 de las 64 características.

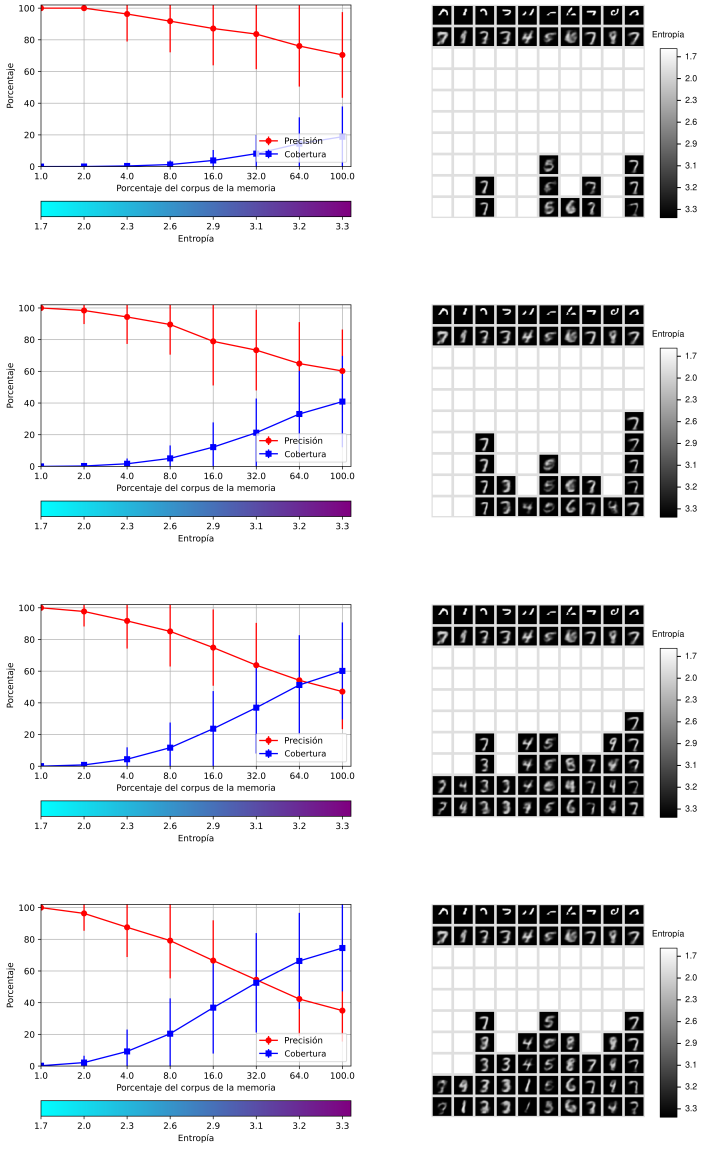


Figura 7.9: Recuperación de memoria con oclusiones del 50 % en la parte inferior, con relajación de 1, 2 y 3 de las 64 características.

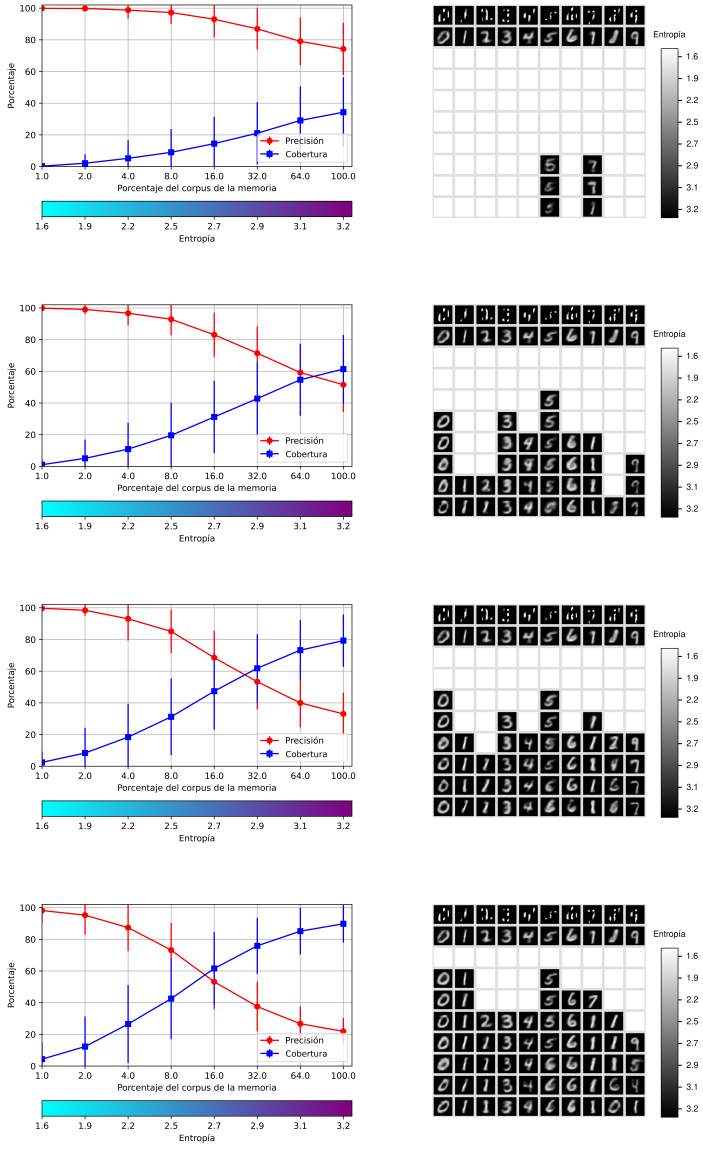


Figura 7.10: Recuperación de memoria con oclusiones por barras verticales de 4 pixeles de ancho, con relajación de 1, 2 y 3 de las 64 características.

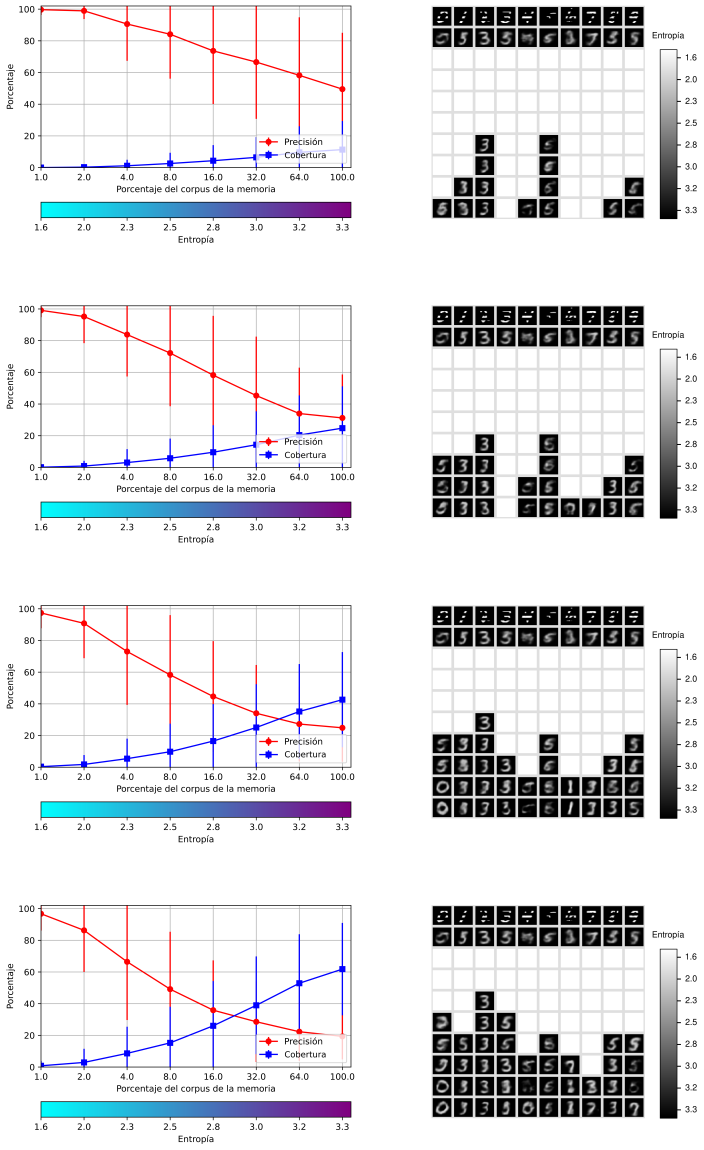


Figura 7.11: Recuperación de memoria con oclusiones por barras horizontales de 4 pixeles de ancho, con relajación de 1, 2 y 3 de las 64 características.

7.5. Propiedades Generales

En este capítulo se presenta un sistema de memoria asociativa y distribuida, pero a la vez declarativa y constructiva, con la capacidad de rechazar directamente cues de objetos que no están en la memoria. Las instancias individuales de los objetos representados se caracterizan en tres niveles: i) como representaciones concretas en modalidades específicas, que se sensan o se generan directamente o, de manera alternativa como funciones de pixeles a valores; ii) como representaciones abstractas, independientes de la modalidad o amodales, en el espacio de características; estas representaciones se relacionan uno a uno con sus representaciones concretas respectivas en el primer nivel; y iii) como representaciones distribuidas consistentes de la abstracción disjuntiva de las instancias, posiblemente muy numerosas, de los objetos expresados en el segundo nivel. El primer nivel es declarativo y simbólico; el segundo es todavía declarativo pero independiente de la modalidad y puede consistir en una representación integrada de objetos presentados en modalidades diferentes; y el tercer nivel es una estructura distribuida y sub-simbólica que mantiene la abstracción de un conjunto de objetos del segundo nivel.

Las operaciones de registro y reconocimiento utilizan a la disyunción lógica y a la implicación material; se llevan a cabo por manipulación directa entre celdas correspondientes en las tablas; y la información se toma y se deposita en el bus también por manipulación directa. Adicionalmente la capacidad de rechazar cues provee la funcionalidad básica para habilitar la negación fuerte, y da soporte a la expresión de la negación en el lenguaje natural, en los lenguajes lógicos y en los sistemas de representación que asumen que el conocimiento es incompleto, como se discute ampliamente en el capítulo 4. Este aspecto lógico y directo refuerza el carácter declarativo del sistema. La propiedad asociativa de-

pende del papel dual jugado por las representaciones intermedias que expresan contenido pero al mismo tiempo selecciona sus RMAs correspondientes en las operaciones de reconocimiento y recuperación. La operación de registro, por su parte, es análoga a la fase de entrenamiento del aprendizaje de máquina supervisado, y presupone un mecanismo de atención que seleccione el RMA en que se debe incluir la información. Este problema se debe abordar en relación a la arquitectura cognitiva y los mecanismos de atención.

Los módulos de análisis y síntesis que mapean representaciones concretas a abstractas y viceversa son computaciones seriales desde la perspectiva conceptual –independientemente de que las micro-operaciones internas se puedan llevar a cabo en paralelo por unidades de procesamiento gráfico o GPUs, pero las operaciones que manipulan los símbolos que se almacenan en celdas correspondientes de las tablas involucradas, las cuales se efectúan en un número muy reducido de pasos computacionales, se pueden llevar a cabo de manera paralela si se dispone del *hardware* apropiado. Estas operaciones involucran la activación simultánea de todos los registros de memoria asociativa, y este paralelismo no sólo es algorítmico o implementacional sino que se lleva a cabo en el nivel de sistema computacional o funcional en el sentido de [70], como se elabora adelante en el Capítulo 8. En la presente simulación los mecanismos de análisis y síntesis se implementan mediante redes neuronales profundas estándar, pero esto es una contingencia que depende del estado del arte de la tecnología computacional. Sin embargo, desde la perspectiva conceptual, es posible que esta funcionalidad se pueda implementar con otros modos de computación, disponibles actualmente o que se puedan descubrir en el futuro, que mapeen representaciones concretas al espacio de características abstractas y viceversa.

La funcionalidad de la memoria declarativa propiamente se puede también distinguir de los procesos de entrada y de salida en términos del nivel de indeterminación de los objetos de computación. El análisis y la síntesis computan una función cuyo dominio y codominio son conjuntos de funciones, y estos procesos están completamente determinados, ya que proveen el mismo valor para cada argumento necesariamente. Consecuentemente la entropía de estas computaciones es cero. Sin embargo, las representaciones que se mantienen en los registros de memoria asociativa tienen un grado de indeterminación, el cual se mide con la entropía computacional. La entropía es un parámetro del desempeño del sistema. En primer lugar mide el rango operacional de los RMAs, como se muestra en los experimentos 1 y 2. Si la entropía es muy baja, la precisión y la cobertura son también bajas para el sistema integrado, pero si la entropía es muy alta, la cobertura disminuye ligeramente. Sin embargo, hay un nivel de entropía en el que tanto la precisión como la cobertura son muy buenas. Los experimentos muestran que los RMAs de tamaño 64×32 y 64×64, con entropías de 3.1 y 3.8 respectivamente, tienen características de operación satisfactorias. El registro más pequeño de 64×32 se escogió para los demás experimentos por consideraciones básicas de economía.

El experimento 2 muestra que un sólo RMA puede mantener la representación distribuida de dos o más objetos de clases diferentes. El costo es que la entropía se incrementa, y se requieren registros de mayor capacidad con grandes cantidades de información para la construcción de memorias operativas. Sin embargo, esta funcionalidad es esencial para la construcción de abstracciones más profundas y posiblemente para la definición de conceptos compuestos. El experimento 3 aborda la pregunta de cuál es la cantidad de información y el nivel de entropía que se requiere para mantener un nivel efectivo de reconocimiento

y retribución de información dado un RMA operacional. El resultado es que la precisión en el reconocimiento es en general muy alta sin importar mucho la cantidad de información que tiene el registro de memoria. Consecuentemente, cuando una cue o descripción se acepta es bastante seguro que el objeto descrito corresponde al concepto o clase representada en el RMA. Sin embargo, la cobertura en el reconocimiento es muy baja para niveles bajos de información y entropía, pero es alta para niveles moderados de entropía. Una vez más, hay un rango de entropía donde el valor de este parámetro no es ni muy bajo ni muy alto en el que la información es rica y el sistema de memoria es efectivo. El experimento 4 se enfoca a la pregunta de qué tan similares son los objetos recuperados de la memoria en relación a la cue o descriptor. Los resultados muestran que la similitud es muy buena en general aunque baja un poco para niveles altos de entropía siempre y cuando los RMAs operen dentro de su rango funcional. Este experimento muestra también el rechazo cuando los objetos no están en la memoria o cuando la entropía es demasiado baja. En el caso básico, cuando la entropía es cero, el objeto retribuido es el mismo y las operaciones de retribución y reconocimiento de memoria no se distinguen. Esta situación corresponde a las memorias de acceso aleatorio o RAM de las computadoras digitales, que son reproductivas. Sin embargo, las memorias naturales son constructivas en el sentido de que la operación de recuperación produce un objeto genuinamente novedoso. Ésta es la razón para definir al operador β utilizando una distribución aleatoria. Siempre que la cue o el descriptor se acepta, la operación de recuperación selecciona un objeto cuya representación está entre las funciones incluidas en la relación representada por la tabla. El objeto retribuido se puede o no haber incluido por una operación de registro de memoria explícita, pero siempre es el producto de una construcción en la memoria. El experimento muestra que

cuando la entropía es baja la cobertura es baja pero la precisión es muy alta; que la cobertura sube conforme al incremento de la entropía aunque a costa de la precisión; que hay cues que no se retribuyen a bajas entropías pero sí se aceptan a entropías más altas –y consecuentemente que el recuerdo depende tanto de la cue como de la cantidad de información almacenada– y que hay cues que nunca se aceptan. Este resultado sugiere asimismo que la recuperación de memoria va de la "copia fotográfica" a los objetos reconstruidos, a los objetos imaginados y finalmente al ruido.

La naturaleza constructiva de la recuperación de memoria se opone al carácter asociativo y reproductivo del autocodificador. En la práctica este último genera una imagen ligeramente distorsionada de la imagen de entrada, pero esto se debe a que el codificador computa tan sólo una aproximación de la inversa del codificador, y a que la imagen de salida se asocia a la imagen más parecida en el espacio de imágenes incluidas implícitamente en el codificador, pero esta operación no se debe confundir con el carácter constructivo de la memoria, que se debe a la definición aleatoria de la operación β y al nivel de entropía del registro de memoria. Las asociaciones resultan de computaciones determinísticas con entropía cero mientras que las construcciones genuinas son producto de computaciones cuya entropía es mayor que cero. La diferencia entre la propiedad constructiva de la memoria y la asociativa del autocodificador ilustra la oposición tradicional entre las teorías de aprendizaje *asociacionistas*, –como el conductismo clásico– y las *constructivistas* –como el constructivismo propuesto por Barttlet e incluso por Piaget.

El experimento cinco abordó el registro, reconocimiento y recuperación de memoria con información incompleta. Esta funcionalidad ha sido una de las motivaciones más importantes en el estudio de la memoria asociativa desde las

investigaciones de Bartlett acerca del recuerdo, y también para el desarrollo de memorias asociativas en el paradigma de las redes neuronales desde las propuestas iniciales y, en particular, desde la propuesta original de Hopfield. Las cues pobres e incompletas ponen un reto significativo a este tipo de sistemas y resalta el peso de la entropía en las operaciones de reconocimiento y recuperación. Los resultados muestran de manera mucho más clara que a bajas entropías la cobertura es baja pero la precisión muy alta; en estos niveles es difícil aceptar cues y puede haber falsos negativos, pero cuando se aceptan se reconocen correctamente. Esta relación se invierte conforme se incrementa la entropía. Hay un rango de entropía con valores moderados donde el compromiso entre la cobertura y la precisión es muy bueno; sin embargo, si la entropía sube demasiado la cobertura sube a niveles muy altos y las cues se aceptan muy fácilmente pero puede haber muchas interpretaciones incorrectas y la precisión baja significativamente. La información incompleta permite apreciar de manera muy clara el compromiso de la entropía.

Este experimento resalta asimismo las diferencias entre la presente propuesta y las memorias asociativas en el paradigma de Hopfield, como sigue: i) mientras que el presente modelo generaliza los patrones almacenados y ofrece una explicación a la imaginación, estas últimas contienen patrones que se han almacenado exclusivamente, de la misma manera que las memorias RAM clásicas: ii) mientras que en el presente modelo la recuperación de memoria es constructiva y produce un objeto novedoso, las memorias de Hopfield asocian la cue al patrón más parecido; y iii) mientras que el presente modelo rechaza directamente las cues de objetos que no están en la memoria, las memorias de Hopfield buscan intensamente por un patrón similar y, consecuentemente, siempre regresan un patrón aunque éste no corresponda a la cue; o fallan, implementando una

forma de la negación por falla, y asumen implícitamente la hipótesis del mundo cerrado; además no se sabe si la respuesta es correcta debido al problema del paro y esta forma de negar es muy costosa en tiempo y esfuerzo de cómputo, lo cual se contrapone con la experiencia cotidiana: saber que no se sabe es inmediato. Que uno le componga es otra cosa.

Este último experimento resalta también que la recuperación de memoria involucra tanto la operación de abajo hacia arriba –*bottom-up*– que mapea la cue a sus características abstractas, como su contribución de arriba hacia abajo –*top-down*– que depende del contenido y la entropía del registro de memoria. Desde esta perspectiva el presente sistema de memoria tiene una orientación bayesiana, aunque no utiliza probabilidades de manera explícita. En esta analogía el módulo de análisis computa una "plausibildad" o *likelihood*; el contenido almacenado en la memoria es un apriori; y la operación de recuperación retribuye, de forma constructiva, el mejor compromiso entre la evidencia externa provista por el módulo de análisis y el conocimiento almacenado en la memoria. El módulo de síntesis computa una plausibilidad que se pondera con el apriori para recuperar una forma de la memoria. Esta perspectiva ilustra claramente el impacto de la memoria en la interpretación perceptual y en la generación de la salida: mientras que la memoria permite el uso de la experiencia previa, que se aprende, las arquitecturas cognitivas que carecen de una memoria declarativa dependen completamente de las habilidades perceptuales, caracterizadas por el módulo de análisis, y las habilidades motoras, caracterizadas por el módulo de síntesis, que se entrenan.

El estudio de los mecanismos de memoria para almacenar representaciones de objetos individuales es central a la cognición y a la construcción de dispositivos computacionales, tales como los sistemas de información y los sistemas

robóticos. Los datos producto de la sensación se presentan a la mente en el formato de frecuencias naturales en el que la información espacial y temporal está disponible directamente, pero se representa como una abstracción muy profunda e independiente de la modalidad. Estas representaciones se pueden recuperar directamente por la percepción en la producción de interpretaciones, por el pensamiento para la toma de decisiones y la planeación, y por la motricidad para el despliegue de las habilidades motoras.

Los sistemas o mecanismos de memoria asociativa habilitan a la memoria de largo plazo, tanto episódica como semántica, y se accesan cuando se requiere por la memoria de trabajo. Estos dispositivos se pueden utilizar para la construcción de recuerdos y conceptos compuestos que se guarden también en RMAs o de conceptos con un mayor nivel de estructura que se basen en RMAs básicos. Los modelos de memoria asociativa pueden ser esenciales a la construcción de lexicones o diccionarios mentales, memorias enciclopédicas, memorias de modalidades específicas, como caras, formas o voces prototípicas, tanto en estudios cognitivos como en la construcción de aplicaciones computacionales. El sistema de memoria asociativa que se presenta en este capítulo se puede ver como una prueba de concepto, y queda abierta la posibilidad de definir memorias para dominios realistas y aplicaciones computacionales prácticas.

Capítulo 8

Concepto de Computación

8.1. Cognición y Representación

La Máquina de Turing permitió por primera vez en la historia crear y manipular representaciones genéricas firmemente enraizadas en la tecnología y consecuentemente en el materialismo científico. Las cadenas de símbolos en la cinta de la máquina son representaciones que se manipulan mecánicamente y se interpretan por los seres humanos. Las representaciones formales o maquinales contrastan con las representaciones naturales –que hipotéticamente sostienen los seres humanos y los animales no humanos con cerebros suficientemente desarrollados– ya que en estas últimas no se sabe cuál es el medio, la naturaleza de los objetos que las constituyen y las operaciones que los manipulan, ni cómo son accesibles al proceso de interpretación. Suponemos intuitivamente que las representaciones mentales se construyen en la memoria a partir de la percepción, son el objeto del pensamiento y son causales de la acción intencional, pero su formato y sus procesos asociados no son sujetos de inspección directa.

El estudio de la percepción, el pensamiento y la memoria, se remonta a los griegos;[1] pero actualmente, desde la perspectiva computacional y de la Inteligencia Artificial, son la materia de la Cognición, la cual consiste en el estudio de la mente como procesos de información o, más precisamente, como procesos computacionales.

El paradigma computacional de la cognición hizo posible hablar de representaciones naturales en oposición a perspectivas especulativas o introspectivas que no eran científicas, y también al conductismo psicológico, para el cual las representaciones no eran objetos genuinos de investigación científica. Chomsky hizo la propuesta explícita con la publicación de *Estructuras Sintácticas* (*Syntactic Structures*) en 1957, donde la sintaxis del inglés se podía modelar con reglas de carácter computacional que producían las estructuras o representaciones sintácticas de oraciones particulares [73], así como con su refutación a la visión del lenguaje presentada en el libro *Conducta Verbal* (*Verbal Behavior*) de Skinner [74]. En esta misma línea Fodor propuso que el conocimiento se representa en el *Mentalese* [75], el cual consiste en un lenguaje con carácter proposicional y composicional, el llamado "lenguaje del pensamiento" –*The Language of Thought* (LOT)– inspirado directamente en la MT.

La noción de representación en IA se hace explícita en la Hipótesis de Representación del Conocimiento que establece que un proceso, un mecanismo o un sistema computacional, es *representacional* si sus ingredientes estructurales se pueden expresar como proposiciones de carácter lingüístico, que además son causales y esenciales a la conducta del agente [76]. Esta hipótesis subyace a la distinción tradicional en IA entre las representaciones *simbólicas versus sub-*

[1] Ver, por ejemplo, *Del Alma* y *Del Sentido y lo Sensible* ([71] pp. 823 y pp. 873 respectivamente); para una introducción genérica ver [72].

simbólicas, que dan lugar a los sistemas respectivos. La distinción es que en los sistemas simbólicos el conocimiento se expresa de manera declarativa, por ejemplo, mediante lenguajes lógicos o funcionales, y en particular mediante el lenguaje natural, además de que es posible razonar acerca de las proposiciones expresadas de forma explícita, mientras que la conducta de los sistemas sub-simbólicos se debe a estructuras opacas, como algoritmos especializados o redes neuronales, que son causales del despliegue o ejercicio de las habilidades.

Esta distinción se relaciona también con la capacidad de hablar o razonar acerca de nuestras propias creencias y la posibilidad de considerarlas o juzgarlas como verdaderas o falsas, en oposición a las emociones, que no son sujetas del juicio sino más bien "se sienten". La distinción se relaciona también con la consciencia y la experiencia: mientras que se puede "ser consciente" del conocimiento, el ejercicio de las habilidades "se experimenta".

Sin embargo, los ingredientes estructurales de la MT no distinguen el conocimiento de las habilidades. Es posible regimentar la información para hacerla más transparente o menos opaca, utilizar lenguajes declarativos *versus* procedurales, o representaciones más concretas *versus* más abstractas, con un compromiso entre la expresividad y la eficiencia,[2] pero, a final de cuentas, todas las cadenas de símbolos se manipulan por las operaciones básicas de la máquina y en este sentido toda computación es simbólica. Alternativamente, mientras que lo simbólico es lo que se presenta en el estado latente de la máquina, como en las configuraciones de entrada y salida, o los contenidos de los registros de la máquina, lo

[2] Por ejemplo, un programa fuente en un lenguaje procedural como Fortran o Pascal es más simbólico que el mismo programa compilado, que se podría considerar sub-simbólico: mientras que el programa fuente expresa el conocimiento a un nivel de abstracción apropiado para el consumo humano pero su interpretación directa es muy ineficiente, la ejecución del programa compilado es eficiente pero el programa es opaco a la interpretación humana.

sub-simbólico corresponde a lo que ocurre durante la ejecución del programa, donde las cadenas intermedias en la cinta no son generalmente interpretables. En particular, mientras que la máquina manipula los símbolos localmente a través del escáner, el ser humano interpreta las cadenas en la cinta de manera global, y esta interpretación sólo es posible al inicio o al término de la computación. Por estas razones, la distinción de lo simbólico *versus* los sub-simbólico no está en la máquina y es más bien un producto de la interpretación humana.

Estas consideraciones nos llevan a un segundo sentido de representación: el objeto mental que resulta de interpretar las cadenas de símbolos, los estados y/o los procesos computacionales. Esta interpretación es conocimiento humano que no tiene contraparte en la máquina. *Representar* es adscribir significados a los objetos de interpretación, y de esta forma entender y ser conscientes; este sentido incluye al significado léxico o de las palabras, al significado composicional o de las oraciones; incluye también a la interpretación de las expresiones lingüísticas en relación al contexto o uso, es decir en relación a un sujeto, una situación espacial y un tiempo, e incluso a un mundo posible, como el actual o uno situado en la fantasía. Incluye también las creencias, los deseos y las intenciones; incluye el conocimiento de un dominio específico o especializado, el conocimiento del sentido común y, de forma más general, toda la perspectiva conceptual y afectiva del intérprete hacia el objeto representacional y a través de éste hacia el mundo.

La distinción entre los dos tipos de representación se puede analizar en términos de niveles de sistema. El estudio de fenómenos de la naturaleza o dispositivos complejos requiere abordarse postulando diversos niveles de abstracción o granularidad, donde cada nivel tiene su universo de objetos o de discurso particular, su conjunto de términos teóricos y sus leyes de comportamiento, y se puede analizar de manera independiente, pero cada nivel puede sustentar o re-

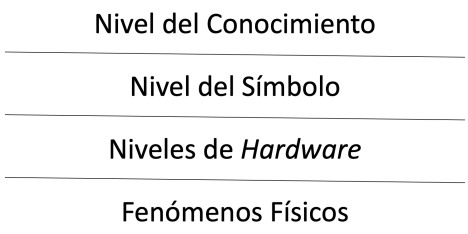


Figura 8.1: Niveles de sistema propuestos por Newell

ducirse a otros niveles. Por ejemplo, la física, la química y la biología abordan el estudio de la naturaleza, pero la física sustenta a la química y ésta a la biología y, de forma recíproca, la biología se reduce a la química y ésta a la física.

En el caso de las computadoras digitales Newell propuso los siguientes niveles, de abajo hacia arriba: i) el de los fenómenos físicos; ii) el de los dispositivos básicos, como resistencias, capacitores e inductores; iii) el de los circuitos electrónicos; iv) el de los circuitos lógicos; v) el de la transferencia de registros o de la arquitectura computacional; vi) el nivel del símbolo o del *software*; vii) y el nivel del conocimiento hasta arriba de esta jerarquía [21]. En la Figura 8.1 se ilustra una versión simplificada de la jerarquía de Newell.

La MT propiamente se define en el nivel del símbolo, donde se lleva a cabo la manipulación simbólica. El significado, por su parte, pertenece al nivel del conocimiento. Newell propuso que éste emerge del nivel del símbolo pero no es reducible, mientras que el nivel del símbolo y todos los niveles inferiores se reducen al nivel que los sostiene. Por ejemplo, el nivel del *software* se reduce al del *hardware*. También sostuvo explícitamente, como ya se ha dicho en relación a la racionalidad limitada, que el medio del nivel del conocimiento es el conocimiento mismo y su única ley de comportamiento es el principio de racionalidad. Esta propuesta complementa la hipótesis del sistema de símbolos físicos que se

refiere también al nivel del símbolo.[3] De estas afirmaciones se puede colegir que Newell y Simon sostienen que el nivel del conocimiento contiene las representaciones que resultan de interpretar a las estructuras simbólicas en el nivel del símbolo y, consecuentemente, que las computadoras digitales pueden entender y ser conscientes.

Sin embargo, Newell no distingue el proceso computacional de evaluar las estructuras simbólicas, que realiza la MT, del proceso de interpretarlas por parte de los seres humanos, y es esta última la interpretación relevante para el nivel del conocimiento. Por lo mismo no es necesario suscribir estas propuestas de Newell y Simon, y se puede simplemente sostener que el nivel del conocimiento reside en la mente humana y contiene las interpretaciones que los humanos hacemos de las estructuras simbólicas en la cinta de la máquina. En la presente reinterpretación de la jerarquía de Newell todos los niveles desde el nivel del símbolo hasta el nivel de los fenómenos físicos se implementan en las máquinas, pero el nivel del conocimiento es conocimiento humano.

Otra jerarquía de niveles de sistema es la propuesta por Marr [70]. Ésta se constituye por tres niveles que son, de arriba hacia abajo, el nivel computacional o funcional, el nivel algorítmico y el nivel implementacional. El primero se refiere a la especificación funcional o la función matemática que modela la facultad de la mente que se estudia, como la visión o el lenguaje; éste es conocimiento humano, y corresponde al nivel del conocimiento en la reinterpretación propuesta de la jerarquía de Newell. En el nivel algorítmico se definen los programas de cómputo propiamente y corresponde al nivel del símbolo. Finalmente, el nivel inferior contiene a todos los aspectos de *hardware* y *software* que sostienen al nivel algorítmico, pero que son contingentes a la computación; por ejemplo, el

[3] Ver sección 3.3.

lenguaje de programación en que se programa el algoritmo o la computadora específica en que se corre dicho programa.

La imposibilidad de que la MT sostenga el nivel del conocimiento fue señalada de forma directa y contundente por John Searle quien la ilustró con la historia del Cuarto Chino [77]. En ésta un sujeto, que no sabe chino, puede sin embargo responder a preguntas en chino siguiendo una tabla con datos e instrucciones, pero sin comprender ni ser consciente de las preguntas que se le hacen ni de las respuestas que él mismo da en chino. En términos de la reinterpretación de la jerarquía de niveles de sistema propuesta aquí, todas las operaciones que realiza la persona en el cuarto chino se dan en el nivel del símbolo o algorítmico, pero como dicho sujeto no entiende ni las preguntas ni las respuestas, no representa en el nivel del conocimiento. El punto de la historia es que se puede manipular representaciones simbólicas sin entender ni ser consciente de lo que significan, es decir, sin interpretarlas. Searle se refirió a esta forma de conceptualizar a la Inteligencia Artificial como *IA débil* en contraste con la *IA fuerte*, que sostiene que las computadoras pueden entender y ser conscientes, como es el caso de Newell en su sentido del nivel del conocimiento. Sin embargo, la distinción entre los dos sentidos de representación está ya presente implícitamente en la definición original de la Máquina de Turing. La máquina transforma la cadena de símbolos en la cinta en el estado inicial de la computación a la cadena en el estado final sin entender o estar nunca consciente del significado de dichas cadenas. Este proceso es completamente mecánico, pero de forma ortogonal a la manipulación simbólica, las cadenas se interpretan por los seres humanos en relación a un conjunto de convenciones de interpretación estándar, como lo enfatizan Boolos y Jeffrey [16]. La MT manipula unidades de forma, es decir los símbolos en la cinta, pero el intérprete humano entiende las unidades de contenido

representadas por dichas formas. Una computación involucra siempre estos dos aspectos: el dispositivo o fenómeno natural que transforma las representaciones "mecánicamente", y el agente que las interpreta, es sujeto de la experiencia y/o realiza la atribución semántica.

8.2. Nociones Alternativas de Computación

La Máquina de Turing articula o modela la intuición particular de Turing del fenómeno computacional. Esta concepción se identifica con el fenómeno computacional mismo ya que anteriormente no había un concepto de computación aceptado y compartido por el mundo de la ciencia y la tecnología, y mucho menos por la sociedad en general. La metáfora computacional de la mente no existía y no se pensaba que el pensamiento fuera un proceso computacional. Sin embargo, la noción de Turing no es la única posible, como se discutió ampliamente en relación a la Computación Relacional Indeterminada, y hay antecedentes explícitos o implícitos de propuestas alternativas.

En particular, el programa del Procesamiento Paralelo y Distribuido (PPD) o Conexionismo [2] propuso que la computación natural consiste en la interacción de un número muy grande de unidades de procesamiento, donde cada unidad realiza operaciones muy simples y se comunica directamente con sus vecinos, como en las Redes Neuronales Artificiales (RNAs). En esta propuesta el énfasis no es en el proceso de cálculo matemático sino más bien en el proceso de la mente.

Sin embargo, como ya se ha dicho, la posición conexionista ha sido objeto de un debate muy intenso. Fodor y Pylyshyn [62] subrayan que las RNAs no pueden expresar estructura sintáctica –ya que no pueden expresar símbolos– y

consecuentemente no pueden guardar información como lo hacen las memorias declarativas estándar, como las memorias de acceso aleatorio o RAM de las computadoras digitales estándar.[4]

Es importante resaltar que tanto el mentalese como los sistemas conexionistas son representacionales, en el segundo sentido del término, en oposición a las posiciones llamadas "eliminativistas" que sostienen que las representaciones mentales no existen, como lo enfatizan Fodor y Pylyshyn [62]. La diferencia es que en las representaciones simbólicas los significados se adscriben a las cadenas de símbolos en la cinta, mientras que en los sistemas conexionistas los significados se adscriben a la acción colectiva de las unidades en las redes neuronales.

Sin embargo, no es completamente claro qué significa adscribir significado a una representación distribuida. El problema es determinar al ente que se presenta a la consciencia de entre todos los representados. Por ejemplo, si la representación distribuida contiene los conceptos diez personas, hombres y mujeres, ¿cuál es el que se presenta a la mente cuando se interpreta? La alternativa es que lo que se presenta a la consciencia sea el conjunto de entes pero desprendidos de su cualidad; sin embargo, al no haber una imagen, dicha interpretación se asemeja más a una experiencia que a la consciencia. Este problema no se da en las representaciones simbólicas ya que la interpretación de una cadena de símbolos presenta a la mente al ente denotado. Si éste es concreto la interpretación se puede presentar como su imagen en alguna o varias modalidades de la percepción, y si es abstracto, lo que se presenta a la consciencia es el símbolo mismo, como la imagen textual o sonora de una palabra.

Fodor y Pylyshyn conceden que los sistemas conexionistas podrían ser una implementación del lenguaje del pensamiento, que para ellos es el objeto de la

[4]Ver Capítulo 7.

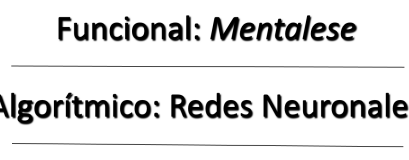


Figura 8.2: Mentalese y su implementación en RNAs

cognición propiamente, y que ésta emerge de las redes neuronales naturales de los cerebros humanos. Esta relación se ha explicado en términos de niveles de sistema, pero utilizando la jerarquía propuesta por Marr. Desde esta perspectiva el mentalese se ubica en el nivel funcional, las RNAs en el nivel algorítmico, y el sustrato neuronal propiamente en el nivel implementacional, como se ilustra en la Figura 8.2.

La reducción del mentalese a las redes neuronales ha sido sujeta de un gran esfuerzo de investigación, pero con resultados muy poco satisfactorios; el problema es que representar la estructura sintáctica en redes neuronales no es trivial y las representaciones que se han logrado modelar a la fecha son muy superficiales –e.j., [78, 79, 63]– tal y como lo anticiparon Fodor y Pylyshyn. Por su parte, los defensores del conexionismo sostienen que las RNAs son representaciones plenas que se deben ubicar en el nivel funcional [80], aunque este debate parece estar estancado desde hace mucho tiempo.

Sin embargo, independientemente de la arquitectura cognitiva ilustrada en la Figura 8.2, el conexionismo se desvió de su programa original ya que en la práctica las RNAs se especifican como funciones computables por la MT, se implementan con estructuras de datos y algoritmos estándar y se ejecutan en computadoras digitales estándar, es decir en Máquinas de Turing, por lo que las RNAs son MTs. Mas aún, se ha mostrado que toda MT se puede especifi-

car como una red neuronal recurrente [81, 82], por lo que ambos formalismos son equivalentes en términos del conjunto de funciones que pueden computar. Por lo mismo, si la limitación de la MT en relación a las RNAs es que en estas últimas la memoria es distribuida y no local, a la vez que las implementaciones prácticas de RNAs utilizan MTs, cuya memoria es local, la propiedad esencial de los sistemas distribuidos, *la distributividad*, sólo se simula pero no es real o actual cuando se implementa en computadoras digitales estándar.

Esta distinción se puede ver también en términos de niveles de sistema, utilizando la reinterpretación propuesta de la jerarquía de Newell: La interpretación de los sistemas conexionistas y las RNAs se expresa en el nivel del conocimiento, en el cual los seres humanos consideran a las RNAs como sistemas distribuidos, pero su implementación como programas de cómputo estándar se lleva a cabo en el nivel del símbolo, en el que todas las representaciones son locales. Por lo mismo, lo apropiado es conceptualizar a las RNAs como máquinas virtuales que usan a la MT para ejecutar el proceso computacional en última instancia.

Pasamos ahora al impacto en la presente discusión de la Computación Relacional Indeterminada y la arquitectura de la memoria asociativa presentada en el Capítulo 7. Los contenidos de memoria en este paradigma son representaciones distribuidas cuya interpretación directa no es transparente, como se señala arriba para las representaciones distribuidas en general. Sin embargo, la indeterminación de la interpretación se disuelve en el contexto de la arquitectura integrada del sistema de memoria. Interpretar una representación distribuida consiste en este caso en seleccionar un individuo por medio de la operación de retribución de memoria y depositar su representación concreta en el buffer de salida –o en algún otro buffer de la arquitectura cognitiva asociado a alguna modalidad de la percepción– y de esta forma hacerlo accesible al pensamiento consciente. Des-

de esta perspectiva los símbolos son vehículos que se pueden hacer públicos y permiten la presentación de contenidos tanto a la consciencia como en la comunicación, pero este aspecto se pierde cuando los contenidos se almacenan o representan de manera distribuida en los registros de memoria asociativa.

La distinción se puede ilustrar con una analogía al famoso experimento mental del gato de Schrödinger de la física cuántica. En ésta hay un gato en una caja opaca cuya vida o muerte depende de un evento cuántico con probabilidad de ocurrencia del 50 %. La descripción cuántica de este fenómeno corresponde a la superposición de la función de onda de "estar vivo" con la de "estar muerto" y si no hay un observador el gato está vivo y muerto al mismo tiempo; sin embargo, si se abre la caja el sistema se colapsa al estado "vivo" o al "muerto", que es visible al observador externo. El sistema está indeterminado mientras no haya un observador, pero en el momento en que se hace la observación se determina perfectamente. De la misma forma, la representación distribuida contiene un conjunto de entes que están indeterminados, pero en el momento que la memoria se consulta a través de una operación de recuperación, el sistema se decanta hacia un objeto, cuya representación se hace pública mediante un símbolo. La analogía no es exacta, ya que en la historia del gato de Schrödinger el objeto es un ente real mientras que la memoria asociativa contiene representaciones, pero el punto es que éstas pueden estar sobrepuestas, y que el hecho de consultarlas produce una representación concreta, con propiedades determinadas.

También hay que considerar que a diferencia de la MT, en la que los objetos de manipulación son los símbolos, el proceso de información en el CRI se lleva a cabo directamente en el espacio de características abstractas. En este modo no hay manipulación simbólica excepto la que realizan los algoritmos mínimos.

Asimismo, el pensamiento se debe a operaciones globales sobre las representaciones distribuidas y no es transparente a la consciencia.

En la computación artificial los símbolos se expresan en representaciones externas y son objeto de una manipulación muy intensa a través de algoritmos complejos. En las computadoras digitales estándar hay una correspondencia unívoca entre el símbolo almacenado en la memoria, el símbolo que se manipula en la unidad de proceso y el símbolo que se presenta para la comunicación a través de los dispositivos de entrada y salida. Sin embargo, esta correspondencia "por diseño" no tiene que existir necesariamente, y lo más probable es que en el cómputo natural no exista.

8.3. Cognición sin Representación

La observación de que las representaciones en el nivel del conocimiento no existen en las máquinas y no se pueden inspeccionar en la inteligencia natural motivó la propuesta de eliminar estos constructos teóricos en el estudio de la cognición. Brooks hizo esta postura explícita al tiempo que presentó las llamadas "arquitecturas embebidas" para modelar mecanismos bio-inspirados. Este programa se llevó a cabo con algoritmos específicos y RNAs, aunque ahora asumiendo que éstas no son representacionales. La propuesta de Brooks constituyó un nuevo tipo de formalismo de IA en relación a las dimensiones de lo simbólico y lo representacional, aumentando las posturas de la IA fuerte, la IA débil y los sistemas sub-simbólicos, como se ilustra en la Figura 8.3.

Brooks asume implícitamente la postura de Searl en relación a la computacion artificial y se suma a la IA débil, ya que su propuesta rechaza no sólo la interpretación simbólica de los procesos computacionales sino también la inter-

	Simbólico	Sub-Simbólico
Representacional	IA Fuerte	RNAs, algoritmos específicos, etc.
No-Representacional	IA Débil	Arquitecturas embebidas

Figura 8.3: Cuadrantes Computacionales

pretación representacional de las redes neuronales, en oposición directa al programa conexionista original.

Su postura se puede entender también como el rechazo del estado latente de la MT en la cual la información se presenta "de manera etérea" para el consumo humano. En los sistemas embebidos, el proceso de cómputo es un continuo en el que el agente interactúa con el mundo a través de sus mecanismos de sensación y acción.

Sin embargo, la postura de Brooks es "eliminativista" [62] –y se opone frontalmente a la de Searl– ya que tiene como consecuencia que la experiencia y la consciencia tampoco existen en la inteligencia natural, o que éstas no son objetos genuinos de la investigación científica, como en el conductismo psicológico. En este sentido se asemeja a corrientes de opinión muy fuertes en la biología y la psicología que enfatizan la importancia del cuerpo y la interacción con el medio ambiente y la ecología para la cognición. Estas visiones son también no-representacionales y, en conjunto con los sistemas embebidos, dieron lugar a la Cognición Corporal (*Embodied Cognition*) (e.j., [4]).

Una tendencia particular de este movimiento es el Enactivimso [5]; éste incorpora la tradición cibernética, que modela a la cognición con sistemas de control automático y sistemas dinámicos. Estas corrientes tomadas en conjunto die-

ron lugar a la llamada *Cognición E4*, la cual tiene una presencia muy significativa en las ciencias cognitivas en la actualidad. Tales líneas de pensamiento, especialmente el Enactivismo, hacen eco al Conexionismo y sostienen de manera explícita que el fenómeno de la mente no se puede modelar con la Máquina de Turing. Sin embargo, y a diferencia de este último, la Cognición E4 no propone una alternativa a la MT ni tampoco una noción intuitiva del fenómeno computacional. Se puede decir que para este movimiento la mente no es un proceso computacional. Por lo mismo, la Cognición E4 se suma a las disciplinas que estudian a la mente desde una perspectiva descriptiva pero fuera del paradigma computacional, como las neurociencias, la psicología, la lingüística, la filosofía de la mente, la antropología, la biología, y posiblemente otras. Dichas disciplinas han logrado describir muchos aspectos de la mente y tienen aplicaciones potenciales de alto impacto. Sin embargo, cambian la pregunta que inició la metáfora computacional de la cognición, que se enfocaba en modelar los procesos de información causales y esenciales de la mente con la Máquina de Turing.

La Cognición E4, tal como lo hiciera en su momento el Conexionismo, señala las limitaciones de la IA, y consecuentemente al programa de Turing, para modelar la percepción, el pensamiento, el lenguaje y la memoria. Esta crítica es pertinente dado el estado actual y las perspectivas de la IA tradicional. A pesar del éxito incuestionable que la IA ha logrado recientemente con el aprendizaje profundo y por refuerzo, la competencia de los programas y robots actuales para llevar a cabo tareas de la vida cotidiana es muy limitada. En particular, varios aspectos de la consciencia se han podido caracterizar dentro de las disciplinas descriptivas, en oposición a los sistemas de IA que no pueden ser conscientes. Sin embargo, el reto no es a la IA de forma particular sino de manera más fundamental al paradigma computacional y a la Cognición basada en la Máquina de

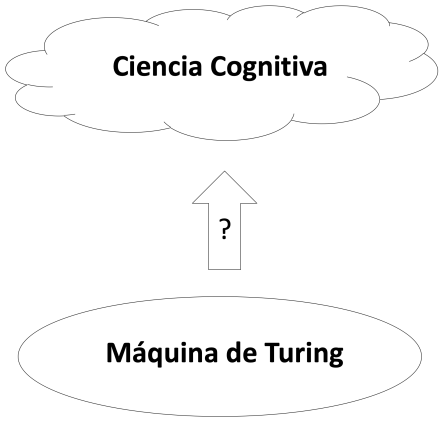


Figura 8.4: Dilema de los estudios Cognitivos

Turing. El dilema es abandonar el paradigma computacional y usar a las computadoras simplemente como herramientas de estudio, como en todas las disciplinas científicas de la actualidad, o extender el modelo de computación que articula la MT para modelar de manera más apropiada a la Cognición. Este dilema se ilustra en la Figura 8.4.

La introducción del Cómputo Relacional-Indeterminado y de un sistema de memoria asociativa de carácter declarativo enriquece los modelos computacionales de la cognición natural. En particular, dado su carácter dual declarativo y distribuido, el sistema descrito en el Capítulo 7 tiene aspectos simbólicos y sub-simbólicos, y abarca los dos cuadrantes inferiores no-representacionales de la Figura 8.3. Por otra parte, sus ingredientes estructurales no aportan nada a la pregunta de cómo surge el nivel del conocimiento ni a la dimensión cualitativa de la interpretación, la experiencia y la consciencia. Por lo mismo, el CRI se opone a los paradigmas en ambos cuadrantes del renglón superior; pero en oposición a Brooks y de manera consistente con el argumento de Searl, la postura que se propone aquí no es eliminativista y simplemente sostiene que no hay razones

Figura 8.5: El Modo de Computación

para pensar que los mecanismos computacionales son capaces de sustentar contenidos mentales, ya que no se sabe cómo surge la consciencia en la inteligencia natural y, consecuentemente, cómo se podría dotar de esta facultad a los entes artificiales.

8.4. El Modo de Computación

El conjunto de intuiciones que motivan a las máquinas o formalismos computacionales es abierto, como se ilustra por el Cómputo Relacional-Indeterminado, y no se puede descartar que surjan nuevos conceptos del fenómeno computacional. Esta diversidad se puede ver en términos de la jerarquía propuesta para los niveles de sistemas en la Sección 8.1 en la que cada máquina o formalismo tiene un nivel de sistema distintivo al cual se designa aquí como el *Modo de Computación*. Este nivel se ubica directamente abajo del nivel del conocimiento y arriba del resto de los niveles de sistema hasta el nivel físico, como se ilustra en la Figura 8.5.

El modo de computación es la estrategia, dispositivo o artefacto, ya sea artificial o un fenómeno natural, que provee el soporte material y realiza las opera-

ciones para encontrar el valor de una función o una relación para cada uno de sus argumentos.

En el caso de la MT y sus formalismos equivalentes —-el cálculo λ, la Teoría de las Funciones Recursivas, las Computaciones *Abacus* (que implementa la Arquitectura de Von Neumann), las RNAs, etc.– el modo de computación es *Computación Algorítmica*. Este modo utiliza procedimientos bien definidos o algoritmos que mapean argumentos a valores de la función que se computa por medio de manipulación simbólica. La teoría de autómatas y lenguajes formales, la teoría de la complejidad computacional y la teoría de la computabilidad y las funciones no computables, como el Problema del Paro (*Halting Problem*), se desarrollaron en relación a la computación algorítmica.

Sin embargo, hay otros modos que también usan manipulación simbólica pero difieren de la computación algorítmica en sus características fundamentales. Por ejemplo, la Computación con Tablas (CT), que utiliza algoritmos mínimos y siempre termina, por lo que su estudio no requiere todo el aparato teórico de la teoría de algoritmos, pero incluye a la entropía computacional y aumenta en este sentido al cómputo algorítmico.

Hay también modos de computación que no usan manipulación simbólica y que realizan computaciones por otros medios, tales como las computadoras analógicas y cuánticas, la holografía, e incluso sensores y transductores de tipos diversos. Estos modos no dependen de algoritmos: las computadoras analógicas eléctricas transforman señales de entrada a señales de salida de manera casi instantánea y no tiene sentido decir que estas máquinas siguen un procedimiento discreto bien definido. Las computadoras cuánticas son también analógicas y el término “algoritmo cuántico” es informal por esta razón.

El modo de computación se puede dividir en dos o más sub-niveles; por ejemplo, un programa escrito en un lenguaje procedural, como *Fortran*, *Pascal* o *C*, está en un sub-nivel inmediatamente arriba del mismo programa compilado en ensamblador, que a su vez está arriba de su codificación binaria; estos niveles constituyen diferentes niveles de abstracción para el consumo humano y conllevan un compromiso entre expresividad hacia arriba y eficiencia computacional hacia abajo. Otro ejemplo son los programas declarativos en *Prolog* o *Lisp*, cuyos intérpretes se programan en *C*, que a su vez se expresan en ensamblador y en forma binaria. Cada uno de éstos es un sub-nivel del modo simbólico. Asimismo, las RNAs se pueden pensar como un modo de computación o una máquina virtual que se ubica arriba de la MT, que es la máquina que realiza la computación propiamente.

Para los modos no simbólicos, como las computadoras cuánticas, se puede definir una interfaz simbólica a través de la cual se presentan las entradas y salidas del fenómeno natural que se utiliza como modo de computación. Asimismo, las computadoras analógicas tradicionales pueden tener una interfaz gráfica o con tablas para presentar las entradas y salidas al intérprete humano. La interfaz más el modo de computación se pueden ver como sub-niveles, el superior orientado a la comunicación y el inferior a realizar el proceso de cómputo propiamente.

Modos de computación diferentes se pueden integrar en sistemas complejos para aprovechar sus fortalezas particulares. Por ejemplo, el entrenamiento de redes neuronales profundas se puede conceptualizar como el aprendizaje de una tabla muy grande para usarse posteriormente en las aplicaciones finales; por ejemplo, para reconocimiento visual o de objetos. Esta combinación de modos puede brindar un compromiso muy bueno entre el esfuerzo algorítmico para el aprendizaje y la computación con tablas para realizar las computaciones de

manera eficiente. En el estado actual de la tecnología la CT se simula con MTs pero se puede implementar directamente con procesadores especializados o con dispositivos de hardware dedicados, para expresar y evaluar tablas muy grandes de forma paralela.

Sin embargo, las computaciones no dependen del modo solamente: para que haya una computación, las entradas y salidas al proceso se interpretan en relación a un conjunto predefinido de convenciones, y el producto de dichas interpretaciones es conocimiento humano en el nivel del conocimiento. Si hay interpretación hay computación pero también, de manera recíproca, si hay computación hay interpretación. Consecuentemente, los eventos computacionales y los actos interpretacionales son equivalentes, y constituyen dos aspectos del mismo fenómeno.

8.5. Computación Natural

La computación natural es aquella que se lleva a cabo por cerebros humanos y de animales no humanos cuyo sistema nervioso está suficientemente desarrollado. El cerebro, como todo sistema complejo, se debe estudiar en términos de una jerarquía de niveles de sistema. Su componente o dispositivo básico es la neurona natural, la cual computa una función binaria en un modelo idealizado. La neurona es un dispositivo en un nivel de sistema particular arriba de los niveles físico, químico y biológico, donde residen los fenómenos de soporte respectivos, y abajo directamente del nivel de la red neuronal, que es también un nivel de sistema distintivo. Sin embargo, debe haber varios niveles de sistema entre la red neuronal y el nivel del conocimiento propiamente. Aunque estos no se conocen se pueden postular hipotéticamente desde varias perspectivas. Una es

considerar a las regiones del cerebro como "órganos" orientados hacia funciones particulares, como la corteza pre-frontal o la cingulada-anterior; sin embargo, la relación de regiones y las funciones mentales que éstas sostienen es muy posiblemente *muchos-a-muchos* por lo que se puede proponer un nivel superior a dichos órganos constituido por las redes de organización funcional, como las redes de alerta, de orientación y de atención ejecutiva [83, 84] las cuales juegan un papel central en la regulación de la conducta. En particular esta última juega un papel central en los procesos deliberativos conscientes, como la planeación, la solución de problemas y el razonamiento, así como en el control de los sentimientos afectivos tanto positivos, como la simpatía y la empatía, como negativos, como la antipatía y la animadversión, y también da lugar a la consciencia de los contenidos y a la conducta voluntaria (Posner et al., 2019, p. 139). Por analogía con la discusión previa, si hay computación natural, debe haber un modo de computación arriba de las redes funcionales y abajo directamente del nivel del conocimiento que llamamos aquí *El Modo de la Computación Natural*. Este modo correspondería a la "máquina" que sostiene a la computación natural.

Aunque la neurona juega un papel central en la computación natural, la función o la relación que computa al nivel del dispositivo no es la función que computa el cerebro al nivel del modo de computación natural. La diferencia se puede apreciar fácilmente mediante una analogía con los sistemas electrónicos: el flip-flop –un circuito electrónico que puede mantener dos estados a su salida (prendido y apagado) que se emplea en los registros de memoria y de los procesadores en computadoras digitales estándar– computa una función binaria en el nivel de sistema de los circuitos electrónicos en la jerarquía de Newell, pero ésta es diferente que la se computa en el nivel del símbolo, que depende del programa de cómputo que se esté corriendo, aunque el dispositivo electrónico involucra-

do sea el mismo. Asimismo, un flip-flop puede formar parte de una unidad de memoria o de un registro de un procesador, y por lo mismo ser parte de un equipo de cómputo, pero también de un sistema de control automático o un radio, que no se considera como una computadora, a pesar de que el flip-flop computa la misma función en todos los circuitos electrónicos de los que forma parte. De manera análoga, una neurona natural que computa una función en el nivel del dispositivo puede formar parte de redes neuronales que eventualmente son sujetas de interpretación, y por lo mismo, participar en la computación natural, pero puede también formar parte de otras estructuras enfocadas al control, que no se interpreten directamente. Más puntualmente, ni el flip-flop ni la neurona natural son computadoras en sí mismas, ya que esta propiedad depende del circuito del que forman parte y de si éste es sujeto de interpretación.

En el mundo de los artefactos inventados por los seres humanos hubo sistemas de control automático mucho antes de que se presentara la Máquina de Turing y de que hubiera una noción de computación clara y ampliamente aceptada. De forma análoga, en la historia filogenética del cerebro, las estructuras que realizan funciones de control automático son anteriores a las estructuras que regulan los estímulos sensoriales, el sentimiento de alerta y la atención ejecutiva. La distinción es otra vez la interpretación: la maquinaria estándar se distingue de la maquinaria computacional en que mientras que las primeras se usan pero no se interpretan, las computadoras transforman representaciones de entrada a representaciones de salida para ser interpretadas y construir conocimiento humano.

El modo de computación paradigmático es simbólico ya que nos presenta a la mente la información de carácter lingüístico y proposicional. Éste es el modo del mentalese y es consistente con la Máquina de Turing[5] pero el cerebro puede sostener otros modos de computación natural. Por ejemplo, el modo "imaginístico" constituido por un procesador de "imágenes mentales". En este modo los objetos que se presentan al proceso de interpretación no son símbolos lingüísticos sino imágenes [86, 87]. Sin embargo, esta forma de "imaginación" (*imagery*) ha sido objeto de un rechazo directo por quienes sostienen que el conocimiento tiene un carácter lingüístico y proposicional y ha dado lugar al llamado "debate de las imágenes" [88]. El lado proposicionalista sostiene que la noción de imagen tiene un carácter concreto y apela a las propiedades de la entrada perceptual, pero que el conocimiento tiene un carácter completamente abstracto, consistente en las proposiciones, independientemente de su modalidad de adquisición, y que en todo caso debe reflejar el carácter del formato en el que se efectúa la computación propiamente [89]. El Modo de Computación ofrece una explicación a este dilema: el lenguaje y las imágenes tienen modos diferentes, el primero simbólico, como la MT, y el segundo distribuido, con mayor relevancia de las relaciones espaciales.

Por su parte, el Modo Relacional-Indeterminado y la arquitectura de memoria asociativa explican el fenómeno de las imágenes mentales en términos de los buffers modales en los que se depositan las representaciones concretas que se retribuyen de los registros de memoria asociativa: la imagen corresponde a la

[5]La posición de Fodor varía respecto a la relación entre el *mentalese* y la Máquina de Turing: en El Lenguaje del Pensamiento rechaza que la mente sea una MT [75], pero en su discusión del problema del cuerpo y la mente sostiene que éste se expresa en última instancia en la MT [85] y en su crítica al Conexionismo, en conjunto con Pylyshyn, acepta explíctamente que el mentalese está inspirado en la MT [62].

interpretación de los contenidos en estos buffers. Las representaciones proposicionales son simplemente las imágenes textuales o sonoras de conceptos que pueden o no tener expresión en otras modalidades de la percepción, aunque el razonamiento propiamente se realiza en el espacio de características abstractas.

La caracterización precisa de los niveles de sistema del cerebro es un problema de investigación abierto. Si se descubrieran eventualmente, un nivel privilegiado sería el modo o los modos de computación natural. Su descripción debería especificar el formato de las configuraciones de entrada y de salida así como las convenciones de interpretación. Las redes de orientación, alerta y atención jugarían un papel central para sostener a dicho nivel de sistema, y la intensidad de la experiencia y la consciencia corresponderían al grado de desarrollo de dichas redes. Por lo mismo, individuos de especies animales carentes de dichas redes no tendrían experiencia ni consciencia, y los organismos con daño o degradación cerebral, especialmente de dichas redes, tendrían una experiencia y consciencia degradada en una medida similar. En todo caso, las estructuras al nivel del Modo de Computación Natural sostienen la producción de interpretaciones y en última instancia de los significados en el nivel del conocimiento.

8.6. Computación Artificial *versus* Natural

La computación artificial basada en la Máquina de Turing supone el uso intenso de algoritmos. Éstos se pueden ver como descripciones intensionales de las funciones que computan y su propósito es simplemente hacer explícita la extensión de dichas funciones. La computación procede de manera serial y apoyada por un escáner que inspecciona a la memoria de manera local, la cual consiste en un "recipiente pasivo", de acuerdo a los ingredientes estructurales de la MT.

Las computaciones son determinísticas y se llevan a cabo con gran precisión y a alta velocidad.

Sin embargo, se puede preguntar qué tan natural es que la mente utilice algoritmos, que a final de cuentas son una invención humana, y qué tanta de la información en la mente o en el cerebro es extensional. De hecho, los seres humanos somos muy limitados para hacer cálculos algorítmicos, y cuando lo hacemos requerimos ayudas externas y nos apoyamos en la maquinaria computacional. La información del mundo se presenta a los seres vivos a través de la percepción de manera extensional y es plausible que una buena parte de ésta se guarde y procese de esta forma.

Adicionalmente, en la computación natural se requiere con mucha frecuencia realizar cómputo inverso: se trata de reconocer a un objeto a través de las propiedades sensibles, o de diagnosticar las causas de un evento a través de la descripción de sus efectos, frecuentemente con incertidumbre e indeterminación. Por lo mismo, es plausible que la computación natural utilice formatos que permitan establecer asociaciones inversas, de valores a argumentos, de manera muy eficiente. Las RNAs tienen estas propiedades hasta cierto punto por lo que reflejan a la computación natural, pero la necesidad de simularlas en la MT limita significativamente su capacidad explicativa. Por otro lado, el CRI y la computación por tablas tienen estas propiedades directamente, por lo que es plausible que modelen mejor algunos aspectos de la computación natural.

Las presentes consideraciones sugieren una distinción entre la computación artificial –tanto algorítmica como relacional-indeterminada– versus la computación natural, en seis dimensiones principales: capacidad algorítmica; estructura de memoria; estructura del proceso; si hay memoria asociativa; el grado de la entropía computacional y si es representacional, en el sentido de ser capaz

	Capacidad Algorítmica	Estructura de la Memoria	Estructura del proceso	Memoria Asociativa	Entropía	Representa-cional
Máquina de Turing	Muy alta	Local	Serial	No	No	No
Cómputo Relacional Indeterminado	Baja	Dual Local/ Distribuida	Paralelo	Si	Si	No
Computación Natural	Muy baja	Distribuida	Paralelo	Si	Si	Si

Figura 8.6: Computación Artificial *versus* Natural

de hacer interpretaciones, y tener experiencias y consciencia, como se ilustra en la Figura 8.6. Esta oposición es informal y se presenta aquí como una conjetura acerca de las dimensiones principales que diferencian a estas tres formas de computación.

Asimismo, es plausible que el cerebro utilice algoritmos sencillos pero también otros modos de computación y que establezca un compromiso entre éstos. Este compromiso puede ser el de la entropía computacional, que sostiene que una entropía moderada permite una expresividad rica, computación efectiva pero con cierta indeterminación de la conducta y posiblemente pérdida de información.

8.7. La Tesis de Church

En la teoría estándar de la computación se estipula que la Máquina de Turing es la máquina computacional más general y que toda máquina computacio-

nal es en última instancia una MT. Esta hipótesis se conoce como la Tesis de Church y en ocasiones como la Tesis de Church-Turing. La tesis establece informalmente que la MT modela apropiadamente la noción intuitiva de calculabilidad efectiva, como la que llevan a cabo los seres humanos cuando realizan cálculos matemáticos a mano siguiendo un conjunto de reglas de manera mecánica, sin apelar a la intuición, pero asumiendo que no hay limitaciones de tiempo y recursos materiales, como papel y lápiz. De manera más puntual la tesis establece que:

1. Hay una MT para cada función computable y viceversa, para todas las funciones totales y parciales. Toda MT computa una función y toda función tiene una MT o un algoritmo que mapea cada uno de sus argumentos a su valor; alternativamente, esta tesis establece que si una función de enteros positivos es recursiva entonces es computable y viceversa.

2. Todos los formalismos computacionales suficientemente generales son equivalentes a la MT.

3. No existe una MT que pueda computar la función de paro, o determinar de antemano si una MT o un algoritmo arbitrario parará para cualquiera de sus argumentos.

El punto (1) se analizó en la Sección 6.6. En este sentido la Tesis está abierta pero es sujeta a refutación ya que es posible que haya funciones que carezcan de algoritmos no triviales; siempre existe una MT que computa una función si ésta se reduce a su expresión extensional y el algoritmo encuentra el valor asignado a cada argumento por inspección directa; éste es justamente el algoritmo que se emplea en la Computación por Tablas, pero es claro que si la Tesis se sustenta

en dicho algoritmo la proposición que expresa es trivial. La fortaleza de la tesis radica en (2) y (3). Church y Turing mostraron que el conjunto de funciones computables por la MT y las que se pueden expresar en el Cálculo-λ es el mismo, que a su vez es el conjunto de las funciones recursivas, y que éste es el conjunto de todas las funciones computables, como se discute en la Sección 2.2. Por su parte, se sabe que si la máquina de paro existiera no sería una MT, pero como la MT es la máquina más poderosa que puede existir por hipótesis, el problema del paro no se puede resolver de forma absoluta e.j., [16]. Por supuesto, si se descubriera una máquina más poderosa que la MT que pudiera resolver el problema del paro, la Tesis sería refutada por esta razón.

Sin embargo, la noción de procedimiento efectivo o computabilidad de Turing se ha extrapolado y la tesis se ha interpretado estipulando que la MT puede simular a todos los mecanismos posibles, en particular en las formas que Copeland [90] llama la Tesis de Maximinalidad (*The Maximality Thesis*) que sostiene que todas las funciones que se pueden realizar o generar por máquinas se pueden computar por MTs o que cualquier cosa que se puede calcular por una máquina se puede calcular por una MT; la Tesis de Simulación (*Simulation Thesis*) que sostiene que los resultados de Turing implican que el cerebro, y de hecho, cualquier sistema o mecanismo físico o biológico se puede simular por la MT, y de manera más radical, la Tesis Física Church-Turing (*The Physical Church-Turng Thesis*) que sostiene que cualquier dispositivo computacional físico en un sentido amplio o cualquier experimento de pensamiento físico (*physical thought-experiment*) que sea diseñado por cualquier civilización en el futuro se podrá simular por una Máquina de Turing. Estas tesis, así como otras proposiciones similares, tomadas en conjunto, se designan informalmente como la versión fuerte de la Tesis de Church: la Máquina de Turing es la máqui-

na computacional más poderosa que puede existir en cualquier sentido posible. Sin embargo, la noción de qué tan poderosa es una máquina computacional es problemática en varios sentidos, empezando porque diferentes máquinas y formalismos computacionales pueden ser más poderosos o más débiles unos con respecto a otros en aspectos particulares, aunque todos computen el mismo conjunto de funciones. Por ejemplo, la computación algorítmica enfrenta un compromiso fundamental entre el poder expresivo de las representaciones y la tractabilidad o la posibilidad de realizar una computación con recursos de tiempo y memoria finitos. Este compromiso es explícito en la jerarquía de los lenguajes formales de Chomsky que incluye, de abajo hacia arriba, a los lenguajes regulares, los lenguajes con categorías gramaticales independientes del contexto, los lenguajes con categorías dependientes o sensitivas al contexto y los lenguajes sin limitaciones estructurales de ningún tipo, e.j., [91]. Los lenguajes regulares pueden expresar representaciones concretas o abstracciones muy limitadas pero su cómputo es muy eficiente, mientras que en el otro extremo se pueden expresar abstracciones muy profundas pero el costo computacional puede ser muy significativo y no estar acotado.

El modo algorítmico asume también el *compromiso de la representación del conocimiento* que establece esencialmente que las representaciones concretas se pueden computar efectivamente pero tienen limitaciones expresivas, mientras que las abstractas pueden ser muy expresivas pero no computables de manera efectiva [32, 33]. Por ejemplo, si el conocimiento se puede expresar en la lógica proposicional las consecuencias lógicas se pueden obtener muy fácilmente, pero si se requiere toda la capacidad expresiva de la lógica de primer orden o de lógicas de orden superior, como las que se requieren para razonar acerca del tiempo, los mundos posibles, las creencias, etc., el costo de la computación puede ser muy

alto. Otro ejemplo, como ya se ha dicho, es que las RNAs pueden expresar representaciones distribuidas y procesarlas de manera muy eficiente pero no pueden expresar estructura sintáctica ni almacenar información como memorias declarativas. Asimismo, hay sentidos en los que los modos no algorítmicos son o pueden ser más poderos que la MT. Por ejemplo, mientras que la MT está dirigida a hacer cálculos y es no decidible de forma general debido al problema del paro, la Computación por Tablas se orienta a la memoria y las computaciones siempre terminan. Además, mientras que el cómputo de una función y su inversa, en caso de que exista, requieren diferentes algoritmos, la CT provee el cálculo de las funciones o relaciones inversas por inspección de manera directa usando un sólo algoritmo. La MT se opone también a la Computación Relacional-Indeterminada ya que la primera es determinística mientras que la segunda es indeterminada, estocástica y entrópica. Por su parte, las computadoras analógicas son muy eficientes y las computaciones se llevan a cabo por fenómenos físicos de manera prácticamente instantánea, pero computan conjuntos específicos de funciones; además son indeterminadas y no tienen capacidades de memoria. En resumen, todo modo de computación asume algunos compromisos fundamentales que definen su capacidad explicativa y sus aplicaciones potenciales. Puede haber propiedades de modos de computación diferentes que sean comparables, como la velocidad o la capacidad de memoria de las computadoras digitales *versus* las cuánticas, pero comparar modos de computación diferentes de manera genérica es un error categórico. De manera más fundamental, la formulación estándar de la MT y la Tesis de Church es independiente de la interpretación. Es decir, la computación se concibe como una propiedad objetiva de los mecanismos. Sin embargo, la propiedad distintiva de los mecanismos computacionales en relación a la maquinaria estándar es que los primeros se diseñan justamente

para expresar y transformar representaciones para ser interpretadas por los seres humanos y el producto de dichas interpretaciones es conocimiento consciente, y la caracterización de la computación natural requiere considerar estos aspectos; por lo mismo, la máquina de la computación natural es más poderosa que la MT en este sentido. Por todas estas razones, la versión fuerte de la Tesis de Church es incoherente.

8.8. Cognición y Consciencia

El Modo de Computación y la jerarquía propuesta de niveles de sistema proveen una perspectiva novedosa de las visiones representacionales y no representacionales de la Cognición. En la primera, el nivel del conocimiento mantiene representaciones de las interpretaciones de los objetos o de los procesos en el nivel del modo de computación, pero en la última, estos objetos no existen. El estatus ontológico de los "objetos mentales" en el nivel del conocimiento es problemático en el materialismo, pero si se postula su existencia la noción con más parsimonia es considerar que sus propiedades reflejan las propiedades o procesos de los objetos correspondientes al nivel del modo de computación, para todos los modos que se usen en la computación natural. El modo algorítmico usa manipulación simbólica cuyas interpretaciones en el nivel del conocimiento se pueden conceptualizar como proposiciones en el mentalese. Otros modos simbólicos, como Computación por Tablas y el Relacional-Indeterminado, se caracterizan por su mayor uso de las relaciones espaciales y, el segundo, por la indeterminación de sus estructuras, por lo que su interpretación en el nivel del conocimiento se puede concebir como imágenes, en oposición a las proposiciones de carácter lingüístico del mentalese.

En la visión representacional se requiere sostener que los objetos en el nivel del conocimiento son interpretaciones o contenidos directamente ya que de otra forma se incurriría en una regresión de interpretaciones infinita. Se tiene que sostener asimismo que los seres humanos son conscientes por el mero hecho de tener en mente dichas representaciones, al menos cuando son el foco de atención, incluyendo el aspecto cualitativo, experencial o fenomenológico de la consciencia [92]. Se tiene que considerar adicionalmente que el fenómeno mental es causado por la actividad del cerebro y depende del sustrato físico o medio en el que el proceso se lleva a cabo –que es causal y esencial al objeto mental. Dicho medio y operaciones constituyen el modo de la computación natural. Consecuentemente, en la visión representacional la cognición y la consciencia dependen del modo de la computación natural. Por lo mismo, se requiere responder a la pregunta de cuál es el modo de computación natural que es causal a la experiencia y a la consciencia. La caracterización de este modo y de la propiedad que hace que los intérpretes adscriban significado a los símbolos o procesos en el modo de computación y por lo mismo tengan experiencias subjetivas, es equivalente a resolver el problema fuerte de la consciencia [92]. Sin embargo, estas preguntas están abiertas para la ciencia y no hay o no parece haber ninguna respuesta en el horizonte.

La visión no representacional, por su parte, se puede entender de dos maneras: 1) sostener que el llamado "nivel del conocimiento" es simplemente el proceso de interpretación continuo de los objetos en el nivel del modo de computación y 2) que el nivel del conocimiento no existe de manera absoluta en la computación artificial ni en la natural. La primera de estas dos postura admite que en la computación natural los niveles del conocimiento y del modo de computación se funden en uno sólo, que es el causal de la experiencia y la cons-

ciencia. Si dicho nivel utiliza una estructura representacional, como en el conocimiento simbólico o declarativo, dicha estructura es la representación misma; en este caso, las estructuras representacionales son estables y se pueden "inspeccionar" en "estados mentales" diferenciados. Esta estabilidad permite que dichos estados además de experimentarse sean objetos de la consciencia. Por su parte, si no hay estructuras representacionales estables, como en los esquemas sub-simbólicos, los objetos de interpretación corresponden a las entradas y salidas del proceso sub-símbolico, que no es penetrable. En este caso la interpretación es también la causa de la experiencia y la consciencia, pero los estados son mucho más breves y próximos entre sí, como los cuadros de una película, que son discretos, pero que se despliegan a alta velocidad y producen una experiencia continua, pero que por su brevedad se resisten a ser objetos de la consciencia de manera individual. Tanto en el caso simbólico como en el sub-simbólico, la interpretación permite al agente computacional estar alerta, orientarse, ser consciente y experimentar al mundo. La cualidad de la experiencia correspondería al modo de computación natural empleado en computaciones particulares, incluyendo proposiciones e imágenes; para los modos no-representacionales el proceso computacional corresponde directamente a la interpretación. Esta visión es posiblemente más parsimoniosa que la que sostiene que hay objetos mentales, ya que éstos dejan de ser necesarios, y permite que los modos representacionales y no-representacionales, o simbólicos y sub-simbólicos, co-existan coherentemente. Por último, queda la visión de que el nivel del conocimiento no existe en ningún sentido, tanto en la computacional artificial como en la natural. En este caso no hay interpretaciones y la computación se convierte en una propiedad objetiva de los mecanismos. Sin embargo, esta propiedad se podría adscribir a órganos, como el corazón o el estómago, a dispositivos de control –e.j.,

termostatos, etc.– e incluso a transductores y dispositivos de censado de todo tipo. También se podría adscribir a fenómenos físicos, químicos o biológicos que se podrían considerar también máquinas computacionales, incluyendo al universo mismo. En esta visión no-representacional radical no hay interpretaciones, ni significados, ni consciencia ni experiencia; la computación se desliga o se desprende de la interpretación, cualquier mecanismo se puede considerar una computadora y el concepto de computación se vacía y carece de contenido. En resumen, si existe el nivel del conocimiento, ya sea que se postulen objetos mentales o que se conciba como la interpretación de los objetos o procesos en el nivel del modo de computación, hay interpretación y consecuentemente hay computación. La interpretación es causal a la experiencia y a la adscripción de significados, y cada modo de computación natural da lugar a una "cualidad" o a un tipo particular de experiencia. La cualidad del sentimiento o la experiencia surge de cómo se siente computar con un modo particular, para todos los modos que emplee la computación natural. La interpretación, la experiencia y la consciencia son tres aspectos del mismo fenómeno, o la experiencia y la consciencia son la manifestación de computar/interpretar. La interpretación distingue a la computación de la maquinaria estándar. En el caso de la computación artificial el proceso computacional se distribuye en dos agentes: la máquina que sostiene al modo de computación y el agente que hace la interpretación. En la computación artificial estos son dos entes diferentes, pero en la computación natural quien sostiene el modo de computación y hace la interpretación es el mismo individuo.

Capítulo 9

Arquitectura Cognitiva

La síntesis e integración de los diversos aspectos de la racionalidad de un agente computacional situado en el entorno se concretan en la arquitectura cognitiva. Ésta se constituye por los módulos funcionales de la percepción, el pensamiento, la memoria y la motricidad o acción motora, e incluye también la conducta reactiva, como se ilustra en la Figura 9.1. El agente computacional puede razonar y tomar decisiones para anticipar al mundo, pero a su vez se encuentra inmerso en un ciclo continuo de interacción con el entorno por lo que la presente se denomina aquí como *Arquitectura Cognitiva Orientada a la Interacción –Interaction Oriented Cognitive Architecture–* o IOCA.[1] Por supuesto, en la arquitectura cognitiva natural los módulos funcionales no están claramente demarcados y sus fronteras son difusas, y IOCA es tan sólo una idealización.

El módulo del pensamiento incluye al razonamiento declarativo o deliberativo, el cual recibe las interpretaciones que produce la percepción y genera las

[1]La presente arquitectura es una generalización abstracta de la arquitectura con el mismo nombre que se propuso e implementó para el desarrollo de robots de servicio en el contexto del proyecto Golem [24].

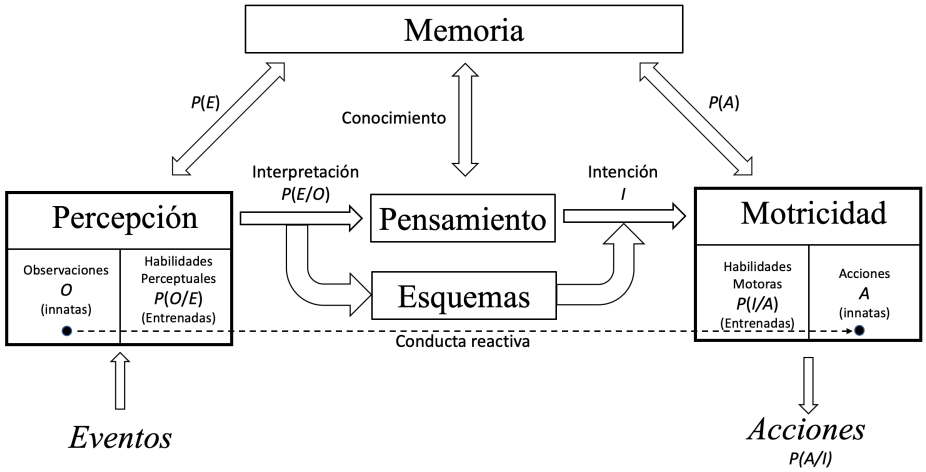


Figura 9.1: Arquitectura Cognitiva

intenciones que se realizan por el módulo de motricidad. Sin embargo, éste es un recurso costoso que se usa por demanda, y se puede "saltar" por medio de esquemas que transforman directamente la salida de la percepción a la entrada de la motricidad. El módulo de memoria es asociativo e incide bidireccionalmente sobre los módulos de la percepción, el pensamiento y la motricidad.

Los términos $P(E)$, $P(O/E)$, $P(E/O)$ en la Figura 9.1 corresponden respectivamente a la probabilidad de que un evento ocurra en el entorno; a la pericia del agente para saber que una observación O tiene por causa dicho evento o la habilidad perceptual del agente; y a la interpretación, que consiste en determinar el evento, entre muchos posibles, que causó la observación.

Los términos $P(A)$, $P(I/A)$ y $P(A/I)$, por su parte, representan respectivamente a la probabilidad de que la acción A se realice en el entorno; a la probabilidad de que la intención se satisfaga si se realiza la acción, que articula a la habilidad o pericia motora del agente; y a la probabilidad de que la acción se seleccione

para satisfacer a la intención entre todas las acciones potenciales. Estos términos constituyen a la contraparte motora de la percepción, donde la intención I que surge del pensamiento y la acción A que realiza el agente corresponden a la observación y al evento, respectivamente. La intención es la causa "invisible" de la acción, que es observable. La acción es un evento, pero causado por el agente para modificar al mundo. El significado e impacto de estos términos se elabora en la Sección 9.2.

9.1. Arquitectura Computacional

Desde la perspectiva computacional, IOCA es una extensión de la arquitectura de la memoria asociativa en la Figura 7.1 con sus respectivas unidades, las cuales incluyen los *RMA* propiamente, los registros auxiliares y los controles de entrada y salida. La comunicación entre los diversos módulos se lleva a cabo mediante un bus central en el que las habilidades perceptuales y motoras depositan y toman información respectivamente, como se ilustra en la Figura 9.2.

El bus tiene n tracks y su contenido se interpreta como una función de n argumentos con sus respectivos valores binarios para efectos de las operaciones de memoria y la interacción con las habilidades perceptuales y motoras. La información circula a través del bus como una forma abstracta, independiente de las modalidades de la percepción y la acción, y sólo se materializa de forma simbólica en los buffers de entrada y salida.

La arquitectura permite incluir un número arbitrario de esquemas que computan funciones específicas, en cuyo caso el contenido del bus representa a los argumentos y valores de funciones o relaciones al inicio y al término de la computación, respectivamente. Puede haber asimismo un número arbitrario de méto-

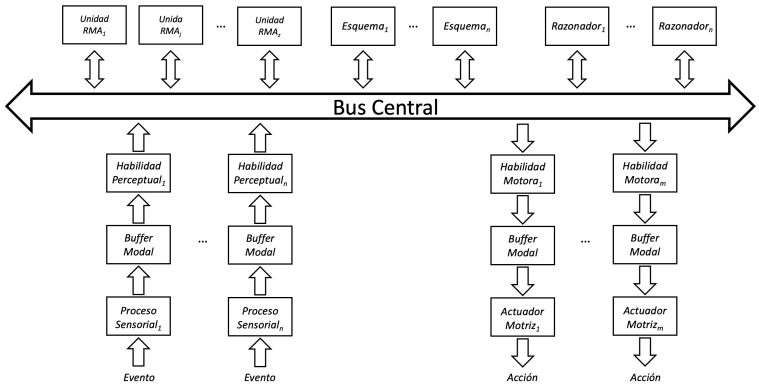


Figura 9.2: Arquitectura Computacional

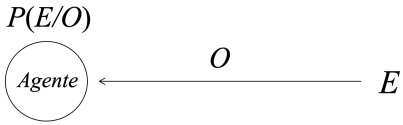
dos de pensamiento o *razonadores* cuyas entradas y salidas se representan de la misma forma.

Consecuentemente, el contenido del bus tiene una interpretación dual, como una función completa –con sus argumentos y sus valores respectivos– o como argumentos y valores específicos, de acuerdo al módulo de la arquitectura que lo consume o lo genera.

Los módulos de la arquitectura se especifican de manera abstracta, independientemente de formatos representacionales y/o estrategias de cómputo, en el sentido de otras arquitecturas como SOAR [22] y ACT-R [23], y la presente arquitectura cognitiva constituye una teoría al nivel computacional o funcional en el sentido de Marr [70].

9.2. Principios de Interpretación y Acción

Un aspecto fundamental del ciclo de inferencia y la interacción es la capacidad de los agentes de diagnosticar correctamente lo que pasa en el entorno o, en



$$E^* = ArgMax_E P(E/O) = P(O/E) \times P(E)/P(O)$$

Figura 9.3: Modelo Bayesiano

un sentido más amplio, de interpretar sus observaciones del mundo. La interpretación es el insumo central a la toma de decisiones y en última instancia a la acción. Los elementos de la interpretación se muestran claramente en el llamado Modelo Bayesiano, basado en el teorema de Bayes, el cual se ilustra en la Figura 9.3. El agente computacional hace la observación O que a su vez es el efecto del evento E que ocurre en el mundo, o dicho de otro modo E es la causa de O. Para realizar esta inferencia hay que tener en cuenta que la observación se puede deber a diversas causas o eventos, y el problema para el agente es determinar el evento más probable E^* dado O. Esto es esencial ya que la decisión y la acción se realizan en relación a las causas, y sólo de manera paliativa en relación a los efectos o las observaciones.

El teorema de Bayes plantea el problema en términos de probabilidades donde el conocimiento deseado E^* corresponde al evento que maximiza la probabilidad $P(E/O)$, es decir la probabilidad que el evento ocurra dada la observación. Este término depende de que la observación O se dé como efecto del evento E, es decir $P(O/E)$ y de la probabilidad $P(E)$ de que el evento E ocurra, la cual se conoce como la probabilidad a priori del evento. El operador $ArgMax_E$ indica que se escogerá el evento E que maximice $P(E/O)$ y, consecuentemente, el término de $P(O/E) \times P(E)/P(O)$ –es decir, el evento que hace la observación más probable en relación a todos los demás eventos.

El término $P(O)$ indica la probabilidad total de que ocurra la observación; sin embargo, dado que lo que se busca es la probabilidad de los diferentes eventos potenciales E para la misma observación O, $P(O)$ es el mismo para todos los casos y se puede descartar. Esta simplificación se puede valorar mejor si se considera que $P(O) = P(O/E) \times P(E) + P(O/\overline{E}) \times P(\overline{E})$ donde el primer y segundo sumandos corresponden, respectivamente, a las observaciones que se clasifican correctamente como instancia del evento y a los falsos positivos, que en conjunto constituyen a la totalidad de las observaciones; con esta simplificación $P(E/O)$ no es ya una probabilidad sino simplemente el peso del evento seleccionado en relación a los demás eventos potenciales, es decir: $ArgMax_E P(E/O) = P(O/E) \times P(E)$.

En esta expresión el término $ArgMax_E P(E/O)$ corresponde a la interpretación, $P(O/E)$ es la información que entrega la habilidad perceptual y $P(E)$ al conocimiento que se tiene en la memoria y que es relevante al evento. La observación O no figura como un término independiente y se puede conceptualizar como una interpretación preliminar que surge del aparato sensorial y de la forma en que éste integra las señales sensadas en el mundo, pero la habilidad perceptual le da una perspectiva adicional que incluye al evento que la causa. Esta perspectiva puede ser innata pero también depender de la experiencia profundamente asimilada.

Por ejemplo, el propósito del diagnóstico médico es determinar la enfermedad a partir de observar los síntomas para decidir un tratamiento. En este contexto las causas o los eventos son las enfermedades y las observaciones son los síntomas. El término $P(O/E)$ indica la probabilidad o plausibilidad (*Likelihood*) de que el síntoma se presente dado que se padece la enfermedad: la habilidad perceptual que el médico adquiere a través de sus estudios y la práctica profesional,

y que le confiere su calidad de experto. En el caso de los sistemas computacionales expertos, determinar este parámetro es un objetivo central del aprendizaje de máquina. La observación O se puede pensar como el producto de la sensación innata –cualquier persona puede sentir el calor de un cuerpo– pero la relación entre la observación y la enfermedad es la habilidad que subyace a la intuición del médico. Por su parte, el término $P(E)$ se puede recabar objetivamente mediante estadísticas y censos.

Supongamos un escenario en el que hay cuatro enfermedades –tifoidea, hepatitis, cáncer y sida– y que el paciente presenta fiebre. Si se conocen las estadísticas de ocurrencia de estas cuatro enfermedades y el médico sabe en qué medida se tiene fiebre cuando se tiene cada una de las enfermedades, el diagnóstico consiste simplemente en escoger la enfermedad que maximiza el producto $P(\textit{fiebre}/x) \times P(x)$ donde x es la enfermedad. Hay que considerar que las enfermedades tienen normalmente varios síntomas –por ejemplo, fiebre, dolor de cabeza, diarrea y palidez del rostro– aunque en diferentes grados y, si se asume que los síntomas son independientes entre sí, la expresión bayesiana es $ArgMax_E P(E/O_{1,\ldots,n}) = P(O_1/E) \times \ldots \times P(O_n/E) \times P(E)$. El mejor diagnóstico tomando en cuenta los cuatro síntomas es la enfermedad que maximice el producto $P(\textit{fiebre}/x) \times P(\textit{dolor de cabeza}/x) \times P(\textit{diarrea}/x) \times P(\textit{palidez del rostro}/x) \times P(x)$. El estudio del diagnóstico o del razonamiento causal en términos probabilísticos se ha abordado de manera muy intensa en IA –por ejemplo, [6, 7].

El modelo Bayesiano ilustra muy claramente el compromiso que hay entre la habilidad perceptual y el conocimiento en entornos que se pueden cuantificar con probabilidades, pero estas dos fuentes de información admiten diversos tipos de modelos y la ponderación de la que surge la hipótesis de interpretación, es decir el producto de la percepción, se puede pensar de manera abstracta. Por

ejemplo, en el caso del médico la habilidad perceptual $P(O/E)$ se puede modelar con una red neuronal, y el término $P(E)$ se puede expresar de forma simbólica en una base de conocimiento como la descrita en el Capítulo 4. En ésta se pueden incluir conceptos propios de la enfermedad y los síntomas, pero también quién la descubrió, cuál es su vehículo, en que condiciones geográficas o socioeconómicas se puede dar, cuál es la mejor medicina para tratarla, cuáles son las contraindicaciones, etc. y la hipótesis de interpretación o el diagnóstico se puede generar a partir de ponderar el producto de la red neuronal con el conocimiento que se tiene en la memoria, aunque cómo hacer esta integración en la práctica es un problema abierto de investigación.

Otro ejemplo es cómo ven una posición en el tablero los ajedrecistas novatos *versus* los expertos. Mientras que los novatos ven las piezas de manera objetiva o neutra, como producto de la facultad visual innata, los expertos utilizan su habilidad ajedrecística. La interpretación del novato es simplemente la observación O mientras que la del experto surge de su habilidad perceptual $P(O/E)$. Ésta se adquiere con los años de experiencia. Por su parte el término $P(E)$ corresponde a los conceptos de ajedrez explícitamente almacenados en la memoria, como la teoría del juego, las aperturas clásicas, los campeones mundiales que las jugaron, los conceptos estratégicos, etc., y también a los implícitos, como el producto de Minimax, pero son públicos y se pueden expresar a través del lenguaje. La hipótesis de interpretación surge de ponderar el producto de la habilidad perceptual, que permite "ver la jugada", con el conocimiento conceptual almacenado en la memoria, como la movida del libro, y del análisis producto de la inferencia deliberativa. El agente tiene que confiar en sus observaciones innatas y en sus habilidades perceptuales adquiridas a través de la experiencia, ya que de otra forma la especie no hubiera evolucionado o se hubiera extinguido. Sin embargo, esta

fuente de información no es completamente confiable ya que puede haber ruido en el ambiente o las funciones sensoriales pueden estar disminuidas, ya sea por enfermedades o por accidentes; y si cambia el entorno –por un cataclismo o una pandemia, o por causas inducidas por otros agentes, como el cambio climático inducido por la acción humana– o el agente se mueve a un entorno novedoso, sus percepciones se verán afectada, sus interpretaciones serán defectuosas y sus acciones no serán las adecuadas.

La percepción enriquece a las observaciones primitivas dándoles un contexto y una perspectiva desde el punto de vista del agente, y el conocimiento enriquece la interpretación al tomar en cuenta su experiencia más profunda. El componente sensorial pesará más en la interpretación si el entorno inmediato es lo relevante, como cuando se evita pisar un charco; habrá contextos donde lo sensorial y el conocimiento tienen pesos balanceados como en el ajedrez donde pesa tanto ver la jugada como analizar sus consecuencias; y otros donde el conocimiento prevalece ampliamente sobre lo sensorial, como en el diagnóstico de problemas políticos, económicos y sociales; o en la educación donde hay que considerar la formación integral de la persona. Lo que hay que subrayar aquí es que en todo caso la interpretación –el producto de la percepción– surge de ponderar la información proveniente del entorno –que se habilita por la habilidad perceptual– con el conocimiento que se tiene en la memoria.

La hipótesis que la gente razona utilizando el teorema de Bayes se ha estudiado empíricamente y los experimentos iniciales sugirieron que éste no es el caso [8, 93]. En particular, mostraron que la gente ignora las probabilidades a priori y prefiere razonar mediante heurísticas, que aunque puedan ser confiables en varios contextos pueden traducirse en conductas irracionales en otros. A este resultado se le denominó *the base-rate neglect fallacy* (e.j., [8]) –la falacia de las tasas

base– y fue ampliamente aceptado en psicología cognitiva durante las últimas tres décadas del siglo pasado; aunque también se observó que las probabilidades a priori interactúan con información de carácter específico y se consideran si son relevantes para el problema particular [94]. Investigaciones más recientes sugieren que los experimentos que motivaron dicha falacia asumían de manera implícita que la gente, por ejemplo los médicos, usan el teorema de Bayes en el formato estándar de la teoría de probabilidad. Es decir, que la probabilidad posterior o inversa $P(E/O)$ se calcula con base en la probabilidad a priori, la plausibilidad y la probabilidad total de las observación, que implica calcular la razón de falsos positivos (comúnmente llamados en el inglés original *base rate*, *hit rate* y *false alarm rate*, respectivamente), lo que a su vez presupone que se conoce el teorema de Bayes y que se puede utilizar de manera operativa por los seres humanos, aunque de manera implícita o inconsciente; pero la mayoría de los médicos, como la mayoría de la gente, no estudian la teoría de la probabilidad y no están familiarizados con el teorema ni con la notación y, consecuentemente, su desempeño es muy pobre. Sin embargo, si el médico sabe cuántas personas de su localidad presentaron tanto la enfermedad como los síntomas y aquellos que presentaron los síntomas pero no la enfermedad, puede calcular la probabilidad a posteriori utilizando una formulación válida del teorema de Bayes en términos de frecuencias en vez de utilizar las probabilidades directamente, a pesar de que no sepa el teorema explícitamente [95, 12] y la hipótesis que las probabilidades a priori se desechan no se sostiene [11]. Esta presentación de la información se ha denominado como *formato de frecuencias naturales* [95] y en su discusión se arguye de manera muy clara que el mismo objeto matemático –en este caso del teorema de Bayes– se puede representar e implementar de formas diversas; que la elección particular tiene un impacto muy directo en el costo computacional

en términos de tiempo y memoria; y que la gente puede efectivamente razonar con el teorema de Bayes siempre y cuando la información se presente como frecuencias naturales.

Por otra parte, desde la perspectiva ecológica la información que se colecta o proviene del entorno debe presentarse en un formato que se pueda utilizar directamente por la mente (e.j., [12]); pero también el formato debe ser el que se usa por la memoria y el pensamiento. Por ejemplo, la información que se expresa en un buffer visual refleja directamente la estructura espacial, pero su transformación a una forma amodal y su registro y almacenamiento en la memoria asociativa refleja el formato de frecuencias naturales. En dicho sistema el producto de la red convolucional de entrada implementa la habilidad perceptual y genera el término $P(O/E)$, pero expresado como una función y no como un número real. Por otra parte, el contenido del registro de memoria asociativa es la abstracción de las diversas instancias del objeto almacenado –también en un formato natural– y corresponde a la probabilidad a priori $P(E)$, y la operación de reconocer o retribuir un objeto de la memoria "pondera" la información que proviene del entorno con la información almacenada en la memoria, y maximiza de manera implícita el producto de ambas fuentes de información. La operación de memoria produce directamente la interpretación –reconoce un objeto o retribuye un objeto de la memoria– y corresponde a la probabilidad posterior $P(E/O)$, aunque las probabilidades numéricas nunca se hayan presentado y las operaciones aritméticas involucradas en el cómputo del teorema nunca se hayan realizado explícitamente.

Percibir es interpretar. La intuición básica del teorema de Bayes, que consiste en ponderar la información que proviene del entorno con conocimiento, se lleva a cabo por las operaciones de la memoria, pero no es necesario, y tal vez no

es posible, representar las probabilidades y hacer los cálculos aritméticos explícitamente. La mente hace el cómputo de otro modo. El propósito de la computación es interpretar la observación, como cuando un robot requiere identificar el evento más plausible que ocurre en el mundo para actuar en consecuencia, o el médico requiere determinar cuál es la enfermedad que produce los síntomas para decidir cuál es el mejor tratamiento, pero ni el robot ni el médico requieren necesariamente representar al mundo mediante probabilidades ni hacer cálculos de manera explícita si cuentan con mecanismos causales alternativos; por ejemplo, si se emplea un modo de computación analógico y se recurre a la memoria.

Desde la perspectiva de las neurociencias la hipótesis del cerebro Bayesiano [96] sostiene que la percepción, la acción, el aprendizaje y la memoria se llevan a cabo de acuerdo a este mismo principio. La maximización implícita que se requiere en la producción de la interpretación puede involucrar un ciclo de operaciones de memoria y comparar las interpretaciones tentativas con las siguientes observaciones, hasta que la interpretación sea estable; este proceso es dinámico pero en esencia pondera también los estímulos externos con el estado del cerebro.

Estas consideraciones en conjunto sugieren que la estrategia de interpretación óptima para cualquier agente, de cualquier especie, natural o artificial, es seleccionar la hipótesis de cómo es el mundo que resulte de ponderar la información aportada por sus habilidades perceptuales con su conocimiento acerca de los eventos que pueden ocurrir en el entorno. Esta es una proposición de carácter general a la que nos referimos aquí como *Principio de Interpretación*. En el caso básico, cuando no se acumula la experiencia, la habilidad perceptual se reduce a la observación directamente; por su parte, el conocimiento incluye a

la información conceptual explícita e implícita almacenada en la memoria, y al producto de la inferencia deliberativa, que es conocimiento implícito.

La motricidad se puede pensar como la percepción pero la direccionalidad se revierte. En este caso la intención I del agente proviene de su mente y la acción A se realiza para satisfacer dicha intención. En esta analogía $P(I/A)$ representa la medida en que la intención I se satisface si se realiza la acción A; este término se entrena a través de la experiencia y representa a la pericia motora del agente. La habilidad motora es la contraparte de la perceptual. Ésta se despliega, por ejemplo, cuando se anda en bicicleta, se juega billar o se toca el piano. Al igual que en la percepción se distingue de las acciones básicas o innatas: cualquiera puede tomar un bastón de billar e impulsar la bola y cualquiera puede presionar las teclas del piano, pero sólo el billarista experto puede hacer la carambola de tres bandas y el pianista producir música cuando toca el piano.

Las habilidades perceptuales y motoras pueden corresponder a la misma modalidad, como cuando se golpea una pelota con efecto; pero es común que sean diferentes y se desplieguen de manera coordinada, como el billarista y el pianista que perciben mediante la visión, y golpean la bola o presionan las teclas mediante la modalidad motora.

Las habilidades se caracterizan también porque el conocimiento causal que se despliega en su desempeño no es transparente al análisis. Si se le pregunta al médico cómo interpretó el síntoma o al ajedrecista cómo vió la jugada ganadora, dirán que no saben, que nada más "la vieron"; o al ciclista o al piloto del auto de carreras cómo manejan sus vehículos, darán en todo caso una explicación superficial y genérica, pero no podrán decir por qué movieron el volante tantos grados o pisaron el freno con cierta fuerza, ya que esta información está profundamente

embebida en la percepción y la acción motora, pero no es conocimiento explícito.

Asimismo, aunque en algunos dominios de experiencia pueda prevalecer la habilidad perceptual sobre la motora o viceversa, el experto demanda normalmente ambas de forma balanceada: el billarista ve la mesa, valora la posición de las bolas y ejecuta el golpe de manera precisa, y el pianista lee el pentagrama y toca las notas con gran fluidez. Las habilidades hacen al experto y pensar durante su despliegue puede afectar negativamente el desempeño.

Por su parte, el término $P(A)$ se refiere a la probabilidad de que la acción se realice en el entorno. Éste se puede conceptualizar como el componente ecológico acerca de la acción, análogo a la productividad potencial de las decisiones, pero para satisfacer la intención particular. Si este conocimiento no se toma en cuenta hay una gran incertidumbre acerca de si la acción seleccionada para satisfacer la intención se puede realizar. Este es el caso, por ejemplo, cuando la acción se selecciona directamente por la habilidad motora sin consultar a la memoria. Disminuir dicha incertidumbre es esencial ya que no sirve de nada tener las mejores intenciones si las acciones que las satisfacen no se pueden realizar debido a restricciones externas al agente. Este término también puede incluir el costo de realizar la acción; si éste es muy elevado la acción no se podrá realizar por falta de recursos, aunque potencialmente satisfaga la acción.

Finalmente, el término $A^* = ArgMax_A P(A/I)$ denota la acción A que se selecciona para satisfacer a la intención I y corresponde al diagnóstico. De manera análoga a la interpretación que entrega la percepción, la acción que satisface la intención se escoge ponderando la pericia del agente con la probabilidad de que la acción se puede realizar en el entrono, es decir, maximizando el producto $P(I/A) \times P(A)$ para todas las acciones relevantes. La pericia aumenta con

la experiencia y la diferencia entre el novato y el experto es que las acciones del primero dependen principalmente de su dotación innata.

Al igual que en la percepción, puede haber diversas estrategias para representar la habilidad motora y el conocimiento acerca de la acción, y se puede estipular que la mejor estrategia para actuar es seleccionar la acción que resulte de ponderar el grado en que las acciones potenciales satisfacen la intención con la capacidad y/o el costo de realizarlas. Ésta es también una proposición de carácter general a la que nos referimos aquí como *Principio de Acción*.

Como en la percepción, en que la interpretación completa surge de una operación de recuperación de memoria –en la que el descriptor que surge de la habilidad perceptual es la cue– la acción que satisface la intención surge de una operación de recuperación de un registro de memoria motora cuya cue representa a la intención y el objeto recuperado es la entrada a la habilidad motora, implementando de esta forma el principio Bayesiano.

Es también posible que el descriptor que produce la habilidad perceptual sea la entrada a la habilidad motora directamente, saltando la memoria y los esquemas o razonadores, aunque en este caso la interpretación y la acción no se beneficien del conocimiento, y la conducta se empobrezca en la medida en que el conocimiento de la probabilidad de que la acción se realice en el entorno sea relevante.

9.3. Pensamiento Esquemático

El pensamiento esquemático consiste en mapear directamente la interpretación que surge de la percepción al módulo de motricidad a través de un esquema saltando el módulo del pensamiento explícito. Los esquemas pueden ser inna-

tos aunque se pueden emplear de manera contingente por analogía a conceptos aprendidos, o se pueden desarrollar a través de la experiencias, como las habilidades perceptuales y motoras, y manifestarse como hábitos conductuales. El módulo *Esquemas* en la Figura 9.1 ilustra esta ruta.

La conducta cotidiana, como caminar, comer, bañarse e incluso conversar, se puede considerar como esquemática ya que se lleva a cabo de manera automática, sin mediación del pensamiento, siempre y cuando las expectativas del agente sean consistentes con sus observaciones en el mundo. Sin embargo, cuando ocurren eventos espontáneos no previstos, la ejecución del esquema se interrumpe y es necesario enfrascarse en un proceso de pensamiento explícito. La conducta esquemática se puede posponer para atender al evento o se puede continuar de manera concurrente con el pensamiento, pero la atención se focalisa en el proceso de deliberación consciente.

Un ejemplo muy ilustrativo que involucra la operación de un esquema es la heurística de rastreo o seguimiento –*gaze heuristics*– que se usa para seguir la trayectoria de un objeto, como cuando se cacha una pelota; ésta consiste en fijar la mirada en la pelota y correr hacia donde se espera que caiga, ajustando la velocidad de forma tal que el ángulo de la mirada en relación al suelo sea constante [13]. El desempeño de esta conducta involucra un proceso continuo de interpretación y acción, y la esencia de la heurística es calcular a la velocidad del agente como función del ángulo. Este problema se puede plantear como el cálculo de la trayectoria, como en los problemas de balística; pero la racionalidad ecológica –y el sentido común– sugiere que esto va más allá de las capacidades de cálculo de los seres humanos y de otros animales, como los perros cuando se les lanza la pelota. Pero, al no haber alternativa aparente, se sostiene que sí es posible calcular el ángulo y hacer el ajuste de la velocidad por medios naturales, ya que estos

cálculos son mucho más simples [13]. Sin embargo, esto sigue siendo problemático y no es claro que estos cálculos se puedan hacer mentalmente en tiempo real y sin la ayuda de representaciones externas.

Una alternativa es postular que el ángulo no se computa explícitamente y que el problema se resuelve interpretando la imagen visual directamente como una función en el espacio de características abstractas. En esta formulación la imagen en el buffer visual de entrada se traduce por la habilidad perceptual a una representación amodal que codifica el ángulo implícitamente, y el descriptor resultante se deposita en el bus de la arquitectura computacional ilustrada en la Figura 9.2. Este proceso corresponde a computar el término $P(O/E)$ que caracteriza la experiencia en la formulación Bayesiana. El proceso de percepción se completa realizando una operación de recuperación de memoria en el que el descriptor que entrega la habilidad perceptual es la cue y el descriptor retribuido representa la interpretación final, maximizando el producto de $P(E)$ y $P(O/E)$, de acuerdo con el principio de interpretación.

La interpretación es a su vez la entrada al esquema, el cual computa una función cuyos argumentos y valores son los ángulos y las velocidades respectivamente. La aplicación del esquema es análoga a la operación de recuperación de memoria, en la que el ángulo "retribuye" a la velocidad. Mientras que la operación de memoria β computa una función cuyo dominio y rango son objetos del mismo tipo, el esquema computa una función cuyo dominio y rango son de tipos diferentes. La arquitectura no asume un modo de computación específico y el esquema se puede implementar con una MT, como una red neuronal, como computación por tablas o algún otro modo de computación, con la única restricción que los argumentos de entrada y salida se representen en la configuración estándar de forma coherente con la estructura del bus.

El esquema debe contar con un control de entrada para validar que el argumento en el bus pertenezca al dominio de la función, de manera análoga al cómputo de la operación de reconocimiento de memoria η, y lo presente al modo de cómputo en el formato en que éste se defina. El control de salida deberá a su vez transformar el valor de la función al formato de características y depositarlo en el bus.

La salida del esquema se puede considerar como la plausibilidad –una probabilidad condicional– con que se retribuye el conocimiento de otro registro de memoria, donde el descriptor resultante corresponde al término $P(A)$, de acuerdo con el principio de acción. Este último descriptor es a su vez la entrada a la habilidad motora –es decir, $P(I/A)$– que lo mapea al buffer de la modalidad motora –o a un descriptor concreto– que es a su vez la entrada a los actuadores motrices que generan la acción propiamente. De forma más general, la información en los registros de memoria y proceso conectados al bus se pueden considerar como conocimiento a priori y la información que se deposita en el bus como plausibilidades o probabilidades posteriores en el espacio de características; y el pensamiento esquemático se puede conceptualizar como la composición del principio Bayesiano, desde la interpretación perceptual hasta la selección de la mejor acción que satisface la intención.

En el caso que el esquema se defina de forma extensional, como computación con tablas o alguna forma de razonamiento diagramático, el cómputo se llevaría a cabo por inferencias directas y, al igual que en las operaciones de memoria, se implementaría con algoritmos mínimos; en este caso se evitaría la necesidad de efectuar cálculos aritméticos de manera explícita.

Por otra parte, ejecutar un esquema no implica que se toma una decisión necesariamente. Determinar el cambio de velocidad para mantener el ángulo es

un proceso automático y no se puede decir realmente que el jugador de beisbol o un perro cachando la pelota calculan el ángulo y toman la decisión de aumentar, mantener o disminuir la velocidad. La conducta esquemática es directa y se opone a la toma de decisiones genuina en que el pensamiento permite anticipar al mundo en el corto, mediano y largo plazo, y toma en cuenta el conocimiento, las creencias, los intereses y los valores del agente. Por esta razón el pensamiento simbólico o declarativo se distingue de la conducta esquemática.

9.4. Pensamiento y Toma de Decisiones

El pensamiento explícito o deliberativo corresponde al módulo *Pensamiento* en la Figura 9.1; esta facultad extiende a la memoria asociativa y a los esquemas, y se lleva a cabo con módulos "Razonadores" conectados al bus, como se ilustra en la Figura 9.2, cada uno de los cuales incluye un método de razonamiento general.

Un ejemplo paradigmático puede ser el algoritmo Minimax en el ajedrez computacional. La entrada y la salida son descriptores que representan respectivamente la posición en el tablero y la movida que hace el jugador en turno. La arquitectura computacional liga a este módulo con el entorno por medio de las habilidades perceptuales y motoras, y los respectivos canales de entrada y salida.

Minimax, en su formulación tradicional, subsume al diagnóstico, la toma de decisión y la planeación, e incluso la memoria, en un bloque monolítico, como se discute en el Capítulo 3; sin embargo, los estudios de ajedrez humano realizados extensamente por el ajedrecista y psicólogo holandés Adriaan de Groot (1914-2006), que se han confirmado por estudios posteriores, sugieren que el desempeño del ajedrecista depende en gran medida de la percepción y la memoria visual. De Groot realizó una amplia gama de experimentos incluyendo

principiantes hasta grandes maestros y concluyó que la elección de la siguiente movida involucra cuatro fases a las que llamó: 1) de orientación; 2) de exploración; 3) de investigación; y 4) de prueba o verificación. En la primera se evalúa la posición de manera general y se deciden intuitivamente las líneas de acción relevantes; en la segunda se exploran algunas ramas del espacio de búsqueda con una profundidad moderada; en la tercera se selecciona la movida más probable; y en la cuarta se verifica que la decisión es correcta. De éstas la primera es la que distingue más claramente al experto del novato y determina la calidad de su juego. Sus resultados mostraron que tanto los novatos como los expertos evalúan aproximadamente el mismo número de opciones y con profundidad similar; aunque también se ha mostrado posteriormente que los expertos lo hacen con mayor rapidez. A manera ilustrativa, se dice que Kasparov sólo analizaba 10 posiciones en 3 minutos en promedio.

Estas cuatro etapas se pueden referir grosso modo a la arquitectura computacional, donde la primera se puede conceptualizar como una operación de recuperación de memoria en la que la posición en el tablero es el cue; la segunda como una inferencia abductiva o de diagnóstico; la tercera como la toma de decisión propiamente; y la cuarta correspondería al módulo de planeación o análisis; las tres últimas inferencias se implementarían en sus respectivos módulos de razonamiento, aunque con recursos computacionales muy limitados.

Sin embargo, recientemente se propuso una alternativa radical a los juegos racionales. Ésta consiste en el programa AlphaZero [18] que derrotó a los campeones mundiales humanos y computacionales de ajedrez, go y shogi.[2] Este programa utiliza redes neuronales profundas y aprendizaje por refuerzo.

[2] AlphaZero derrotó contundentemente a Stockfish, que había dominado ampliamente el ajedrez computacional desde 2014, como se muestra en la Figura 3.4.

El conocimiento dado a AlphaZero consiste sólo en las reglas del juego y se entrena jugando contra sí mismo. La red neuronal una vez entrenada produce para cada posición o estado s y movida m la probabilidad de ganar $P(m/s)$ si el jugador en turno realiza dicha movida en dicha posición o estado del tablero. Sin embargo, AlphaZero no realiza la movida directamente sino evalúa las consecuencias de las mejores movidas, pero no mediante heurísticas, sino con un método estocástico llamado Búsqueda Montecarlo, que selecciona las movidas aleatoriamente y premia las que llevan a posiciones más favorables o ganan el juego y las penaliza en caso contrario. AlphaZero analiza en el orden de 80×10^3 movidas por segundo; esta cifra es mucho mayor que la capacidad de análisis humana, pero mejora en tres órdenes de magnitud a la búsqueda realizada por Stockfish, que es del orden de 35×10^6, que mejora a su vez a la realizada por DeepBlue en la versión que venció a Kasparov, que era de 200×10^9.

Estas consideraciones se mapean directamente a las arquitecturas cognitiva y computacional en las Figuras 9.1 y 9.2. La red neuronal profunda corresponde directamente al término $P(O/E)$, que caracteriza a la habilidad perceptual, y produce el descriptor que consume el módulo de búsqueda Montecarlo, sin recurrir a la memoria. Por su parte, el proceso de búsqueda corresponde a la fase del pensamiento y aporta el término $P(E)$, aunque codificado de manera implícita en la especificación del algoritmo.

Este proceso de percepción-razonamiento corresponde parcialmente al que propuso de Groot, pero de manera empobrecida ya que no se emplea a la memoria y la interpretación es simplemente una plausibilidad; el módulo de la búsqueda Montecarlo subsume a las tres etapas deliberativas propuestas por de Groot, pero utiliza recursos computacionales de carácter algorítmico que no están disponibles en la computación natural. Sin embargo, AlphaZero sugiere, del mis-

mo modo que de Groot, que en los juegos racionales pesa más la habilidad que el pensamiento explícito, y cambia la concepción tradicional del ajedrecista como el pensador por excelencia.

AlphaZero ilustra, sin embargo, sólo algunos aspectos de la arquitectura cognitiva y la computacional, y un uso más pleno del pensamiento se da en el ciclo de inferencia de la vida cotidiana, que involucra la formulación de diagnósticos, la toma de decisiones y la planeación explícita, como se discute ampliamente en la Sección 3.4 y se ilustra en la Figura 3.6.[3] Dicha secuencia de inferencias, apoyadas por la recuperación de información de la memoria, se puede conceptualizar como el uso de tres razonadores conectados al bus, cuya funcionalidad se demanda en serie, donde la entrada global consiste en las interpretaciones producidas por la percepción y la de salida en las intenciones que genera el pensamiento como un todo.

La inclusión de la memoria asociativa como columna vertebral de la arquitectura cognitiva contrasta con los modelos estándar de estas inferencias, tanto simbólicos como sub-simbólicos. Los modelos tradicionales asumen explícita o implícitamente que los posibles diagnósticos y/o decisiones están dadas de antemano, y las inferencias se enfocan a ponderar las alternativas y seleccionar las de más valor, de acuerdo a alguna función que también se estipula externamente, como en la teoría de juegos, la optimización y la investigación de operaciones. Sin embargo, desde el punto de vista de los agentes autónomos situados en el

[3]La arquitectura IOCA y el presente ciclo de inferencia se implementaron en estas líneas generales en el robot de servicio Golem-III; en éste la comunicación, la memoria y la inferencia utilizan un formato simbólico [24]. El robot se construyó con computadoras digitales estándar y la única fuente de entropía que se considera viene del entorno, pero como los ambientes de demostración y competencia de estos dispositivos son muy predecibles, la entropía es muy baja y la conducta del robot, incluyendo la toma de decisiones, está significativamente predeterminada.

mundo el proceso central de estas inferencias es la síntesis de los diagnósticos y de las decisiones potenciales. Antes que evaluar las alternativas hay que ponerlas sobre la mesa. En los modelos tradicionales se asume que la mesa está servida, pero para los agentes naturales éste no es el caso.

La presente propuesta ofrece una respuesta a cómo se puede llevar a cabo este proceso: los símbolos provienen de operaciones de retribución de información en la memoria asociativa, aunque también podrían provenir de los módulos de razonamiento heurístico y del pensamiento deliberativo. Ante una contingencia como el charco que está en la puerta de la casa, los diagnósticos potenciales se pueden hacer disponibles en el bus central, independientemente que se materialicen o no como símbolos a través de los módulo de síntesis y se presenten en el buffer de salida. De manera análoga, las decisiones potenciales pueden provenir de la memoria o de otros módulos de pensamiento, y se pueden realizar explicita pero no necesariamente, como símbolos.

En los modelos simbólicos el repertorio de símbolos está dado de antemano y se proveen al sistema por medios externos mientras que en los sistemas subsimbólicos no hay símbolos. De manera más general, los modelos clásicos de IA no dicen nada a la génesis del símbolo. Por esta razón la memoria asociativa es indispensable al pensamiento deliberativo.

Realizar operaciones de memoria, computar esquemas y efectuar procesos de razonamiento, incluyendo el cómputo de las habilidades perceptuales y motoras, se pueden ver en IOCA como procesos Bayesianos, pero implementados sin necesidad de efectuar cálculos aritméticos de manera explícita. En esta conceptualización el contenido del bus al inicio de una operación corresponde con el término $P(O/E)$, el módulo que se utiliza en la operación codifica el a priori $P(E)$ y el valor que se deposita en el bus como resultado de la operación es la

interpretación $P(E/O)$. En operaciones en serie éste último es a su vez la entrada $P(O/E)$ de la siguiente operación y el descriptor final que alimenta la habilidad motora se puede considerar como el $P(E)$ de todo el proceso inferencial y de memoria. La habilidad perceptual provee la plausibilidad inicial a partir del buffer modal de entrada y la habilidad motora computa la interpretación final que se deposita en el buffer modal de salida. Desde una perspectiva adicional, el proceso se puede conceptualizar como una composición funcional Bayesiana en el espacio de características.

9.5. Memoria *versus* Habilidades

Las memorias declarativas se distinguen de las habilidades perceptuales y motoras. Estas últimas se refieren comúnmente en neurociencias, neuropsicología y psicología cognitiva como "memoria perceptual" y "memoria motora", pero el conocimiento y las habilidades se oponen en varias dimensiones. Mientras que el primero está disponible en la memoria y se puede adquirir y expresar a través del lenguaje, la habilidad está profundamente embebida en la percepción y la motricidad, y se adquiere por la práctica intensa a lo largo del tiempo. Mientras que el conocimiento se registra, se reconoce, se retribuye, las habilidades se usan pero las experiencias particulares de entrenamiento no se recuerdan o se recuperan de la memoria. Mientras que el conocimiento es transparente a la consciencia, al menos cuando se registra o retribuye de la memoria y se usa en el razonamiento, y es objeto de la reflexión y la introspección, la experiencia de desplegar una habilidad se siente pero es opaca a la consciencia; y mientras el conocimiento es esencial y causal a la conducta intensional, los reportes o explicaciones del "conocimiento" de las habilidades son meras racionalizaciones que

distan mucho del proceso causal a la experiencia misma, como cuando alguien explica su habilidad para manejar una bicicleta.

Esta distinción se refleja en IOCA en la naturaleza de los módulos de la arquitectura. Las habilidades perceptuales y motoras –o los módulos de análisis y síntesis– corresponden a la distinción tradicional de memorias perceptuales y motoras, respectivamente. Estos módulos procesan funciones de transferencia, son procedurales, y sus conexiones de entrada y salida al bus central son funciones que se interpretan como plausibilidades o probabilidades condicionales en la arquitectura Bayesiana, como se ilustra en la Figura 9.1. Por su parte, los RMAs corresponden a información a priori –es decir probabilidades absolutas– y caracterizan a los objetos de la memoria declarativa propiamente. Es posible que haya memorias declarativas con información perceptual y/o motora, como el conocimiento que se tiene acerca de los eventos y acciones que se llevan a cabo en el mundo, sin menoscabo de las habilidades perceptuales y motoras que se asocian a los módulos de análisis y síntesis. Asimismo, los módulos de razonamiento, tanto los esquemáticos como los de propósito general, son análogos a las memorias, sólo que codifican el conocimiento no sólo de manera explícita sino también implícita a través de algoritmos u otros modos de cómputo.

Los módulos de memoria en IOCA pueden ser de carácter simbólico, pero dado que la información en el bus se especifica en el espacio de características, la unidad de entrada y salida de memoria debe traducir dicho formato al simbólico, realizar las operaciones de memoria en este último, y llevar a cabo la operación inversa para depositar la información reconocida o recuperada en el bus. En la práctica, los sistemas de Inteligencia Artificial de orientación simbólica que tienen una interfaz perceptual y motora, y están situados en el mundo, requieren realizar la transformación entre lo sub-simbólico y lo simbólico, y vi-

ceversa, aunque esta transformación se aborda de forma particular y al nivel de la implementación.

9.6. Algoritmos Mínimos

El formato de frecuencias naturales presenta la información a la mente de forma más apropiada que los formatos artificiales, como la notación matemática. Sin embargo, de acuerdo con la Ciencia Cognitiva, el cómputo se lleva a cabo mediante algoritmos estándar utilizando la Máquina de Turing o alguna máquina equivalente. Esta corriente de opinión sostiene que el cerebro es una máquina tan poderosa que puede evaluar las operaciones aritméticas involucradas en el cómputo de manera natural (e.j., [12, 13]); es decir, que los seres humanos y los animales no humanos con un cerebro suficientemente desarrollado computan algoritmos de la misma forma que lo hacen las computadoras digitales del tipo ordinario.

Sin embargo, el formato de la MT es lingüístico y proposicional y si el cerebro fuera una MT el formato de frecuencias naturales se requeriría traducir al formato simbólico o proposicional, y el cómputo sería igualmente difícil o no factible. Por el contrario, el formato de probabilidades es proposicional y éste se puede procesar directamente en la MT, y razonar con este formato sería fácil. De manera más general, para que el cómputo sea factible es necesario que el formato en que se presenta la información sea el mismo, o al menos compatible, con el formato en el que se lleva a cabo el cómputo.

Dado que el modo de computación natural no se conoce, sostener que la mente usa algoritmos computados por una MT es una racionalización a posteriori, pero no causal de la conducta mental. Los conceptos matemáticos, así co-

mo las notaciones y los sistemas métricos, son construcciones culturales e históricas, que aparecieron mucho después de la maquinaria utilizada por la computación natural, y no hay ninguna razón para suponer que el cerebro utiliza dichos constructos, de la misma forma que no hay razón para suponer que las racionalizaciones que provee la gente respecto al conocimiento involucrado en el despliegue de habilidades, tal como andar en bicicleta, son causales.

Es posible que la máquina computacional natural utilice un formato altamente distribuido, en el que el cómputo se lleve a cabo por unidades de procesamiento muy sencillas que computen algoritmos muy simples, y que la complejidad venga del cómputo coordinado de dichas unidades. La interpretación de las representaciones distribuidas como un todo se ejemplifica por los algoritmos mínimos. Esta visión es muy similar a la formulación original de las redes neuronales [2] pero su implementación en computación por tablas, por ejemplo, se puede llevar a cabo por arreglos masivos que realicen computaciones en paralelo en un número muy reducido de pasos de cómputo, y no se necesita reducir a la MT ni utilizar algoritmos costosos de gradiente descendente, como la propagación hacia atrás utilizada en las redes neuronales.

Las teorías tradicionales de racionalidad se enfocan en los mecanismos generales y en las estrategias particulares de las que surge la conducta racional, y se asume implícitamente que el cómputo se lleva a cabo por una máquina general, tal como la Máquina de Turing. Sin embargo, es posible revertir este punto de vista y enfocarse en la máquina computacional que es causal a la conducta. Ésta es la perspectiva de la racionalidad desde el punto de vista de la máquina, en la que la pregunta es cuál es la conducta racional que se puede generar por una u otra máquina.

Capítulo 10

Cognición Natural

La Cognición Natural contrasta con la Cognición Artificial en que en la primera la memoria es asociativa, los procesos tienen un grado de indeterminación y las fronteras entre sus módulos no están claramente demarcadas, mientras que en la segunda la memoria es local, la llamada Memoria de Acceso Aleatorio o *Random Access Memory* o RAM,[1] los procesos están predeterminados y sus módulos –la Unidad Central de Proceso (*CPU*), los registros de entrada y salida y la memoria RAM– son independientes. La cognición natural se distingue también en que los procesos de la percepción y la acción *se sienten* o su despliegue va acompañado de *tener una experiencia* y, en el caso de la cognición humana, y posiblemente de algunas otras especies, la experiencia se enriquece con el símbolo y el pensamiento consciente.

[1]Es paradójico que estas memorias se denominen *Random Access Memory* ya que los registros no se accesan de forma aleatoria y, por el contrario, el acceso está perfectamente determinado a través de la dirección de los registros.

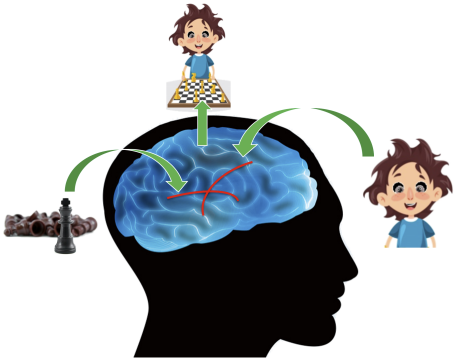


Figura 10.1: Trazos de Memoria de Conceptos Asociados

10.1. Memoria Asociativa Natural

Las memorias naturales de los seres humanos y los animales con un sistema nervioso suficientemente desarrollado son asociativas. Una percepción particular, como un olor, una imagen o una frase, puede disparar una cadena de recuerdos que se van entrelazando, cuya evolución se impacta en buena medida por el contenido afectivo y/o emocional de lo evocado. El registro, el reconocimiento o la retribución de informaciones se hace por asociaciones entre las interpretaciones que provee la percepción en sus diferentes modalidades, incluyendo el lenguaje, y los contenidos previamente almacenados [45].

La diferencia entre los dos tipos de memoria se ilustra contrastando las Figuras 6.2 y 10.1; en la primera hay un trazo de memoria "separado" –el concepto de ajedrez– y en la segunda se muestran dos trazos traslapados –el concepto de ajedrez y el concepto de persona. La asociación se ilustra con la intersección de ambos trazos y la figura sugiere que si se evoca uno se evoca el otro. Si se piensa en el jugador se piensa en el juego y viceversa, y la asociación propiamente representa al concepto de jugador de ajedrez.

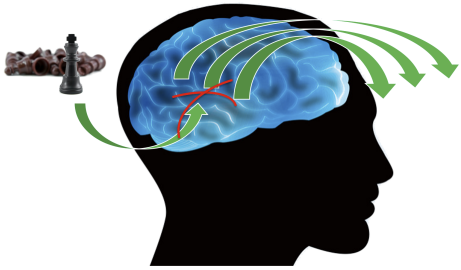


Figura 10.2: El ajedrez como una función de posiciones a movidas

10.2. Asociatividad e Indeterminación

La cognición artificial y natural se oponen también en que la primera es determinista y la segunda no lo es necesariamente, como se discute en la Sección 6.1. Es plausible que en la memoria natural el mismo estímulo evoque diversos recuerdos de manera aleatoria y consecuentemente se realicen diversas conductas igualmente azarosas. Si la aleatoridad es primitiva y no sólo simulada, la máquina en cuestión tiene una entropía computacional, que crece con el número de respuestas posibles.

La asociatividad y la indeterminación de la memoria son nociones relacionadas. Los conceptos asociados deben contar con trazos de memoria compartidos, de tal forma que la activación de un concepto pueda inducir la activación de otros conceptos, pero pensar en un concepto no condiciona pensar en todos sus conceptos asociados, y la selección de las asociaciones en una evocación particular no está completamente predeterminada; esta intuición se ilustra en la Figura 10.2, donde un conjunto de conceptos asociados se relaciona con un conjunto de conductas potenciales que se pueden realizar como respuesta al mismo estímulo.

10.3. Génesis del Símbolo

Una oposición adicional entre la cognición artificial y la natural es que en la primera el pensamiento y la memoria se idealizan como módulos independientes mientras que en la segunda el ejecutivo central y la memoria trabajan de manera profundamente entrelazada. La percepción incide directamente de forma asociativa sobre la memoria, como lo ilustra el caso del ajedrez, y el pensamiento y la memoria inciden directamente sobre la motricidad.

La memoria natural no es un bloque monolítico y se distinguen varios tipos, principalmente la memoria semántica, la memoria episódica, descritas originalmente por Tulving en 1972 [26], y la memoria de trabajo, por Baddeley en 1981 [97]. Tanto la memoria semántica como la episódica son de largo plazo, como se describe en el Capítulo 4. Por su parte la memoria de trabajo es de corto plazo y se utiliza por el ejecutivo central en la manipulación de la información que es sujeta del foco de atención durante el flujo del pensamiento. Sin embargo, no hay una demarcación dura entre los tipos y los procesos de memoria, y al pensamiento se le puede concebir como un proceso que se da en la memoria y, más radicalmente, como un proceso de memoria. Este módulo memoria-pensamiento se ilustra en la Figura 10.3.

Desde esta perspectiva la percepción se modula por las expectativas almacenadas en la memoria y produce las interpretaciones que hace el agente; por su parte, la motricidad responde a las intenciones del agente, pero también las expectativas o predicciones motoras pueden incidir sobre el pensamiento, y tanto la percepción como la acción motora se ligan con el pensamiento-memoria bidireccionalmente.

Se sabe con cierto grado de certeza que la memoria de trabajo se localiza en la zona prefrontal de la corteza cerebral y que la episódica se consolida en el hipo-

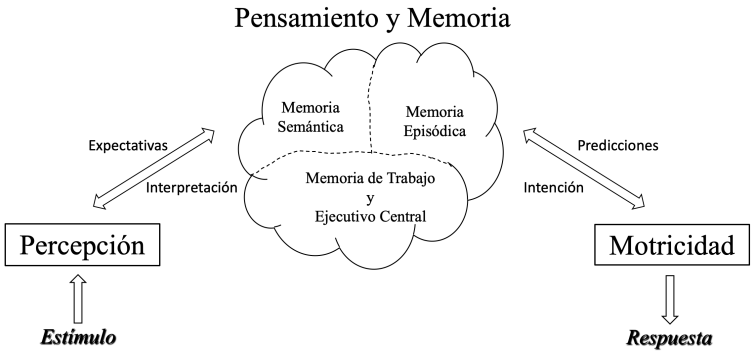


Figura 10.3: Cognición Natural

campo, en el lóbulo temporal, y se cree que tanto la memoria semántica como la episódica están distribuida en diversas locaciones, aunque éstas no se pueden señalar puntualmente, además de que las interacciones entre los tres tipos de memoria son muy elusivas y realmente no hay un modelo claro de su demarcación, funcionamiento e interacción. Tampoco se sabe cuál es el formato en el que se representa la información en ninguno de los tipos de memoria.

Sin embargo, se tiene la intuición muy sólida de que los contenidos –ya sean símbolos lingüísticos o imágenes– se hacen conscientes cuando se presentan a la memoria de trabajo. De manera análoga a los contenidos embebidos en el aparato perceptual, que sólo se hacen conscientes en las interpretaciones que se entregan al módulo pensamiento-memoria, es plausible que los contenidos de la memoria semántica y la episódica sólo se presenten como símbolos cuando se registran y se evocan. El formato de estas formas de memoria es muy posiblemente distribuido y su acceso se lleva a cabo en paralelo, y el sustrato neuronal sugiere que la información se guarde en la memoria de largo plazo de manera sub-simbólica. Sin embargo, lo esencial es que los contenidos, tanto los prove-

nientes de la percepción como de la memoria de largo plazo, se erigen como símbolos al presentarse a la consciencia.

Una nueva analogía termodinámica puede ayudar a ilustrar este concepto: los contenidos de la percepción y la memoria de largo plazo se almacenan en una fase continua de la sustancia, como la gaseosa, pero al pasar al pensamiento consciente hay un cambio de fase, donde la sustancia adquiere un vehículo individual diferenciado, como las gotas que se condensan de la nube cuando llueve. La gota vive brevemente hasta llegar al suelo, donde se vuelve a integrar a la fase líquida continua, en la masa de agua, como la de un lago o del océano, que eventualmente se evapora y se vuelve otra vez nube.

La evocación va seguida del registro refrescado del recuerdo. Las gotas son los símbolos, que pueden ser palabras o imágenes, y se sabe cuál es su referente durante su breve vida en la experiencia consciente. Pero en el mar de la memoria los contenidos se guardan sobrepuestos, distribuidos e indiferenciados.

El símbolo se crea y se evanece dinámicamente en la percepción, en el registro y la evocación del recuerdo, en el razonamiento y en la interpretación y la generación del lenguaje. Este es un proceso creativo continuo y permanente, que subyace a la experiencia consciente. Si se piensa o se habla, está lloviendo.

10.4. Determinismo *versus* Indeterminismo

La discusión previa sugiere que el pensamiento-memoria de la cognición natural puede tomar decisiones libremente y la indeterminación de la mente se opone al determinismo de Laplace y de Turing. El determinismo empieza en la física y se proyecta a la química, la biología y la psicología. De forma recíproca,

la mente, al ser un fenómeno psicológico, surge o emerge del cerebro, y consecuentemente de la biología, de la química y en última instancia de la física.

El determinismo asume que hay un conjunto de leyes de la ciencia que son objetivas y causales de todos los fenómenos de la naturaleza. Esta doctrina asume también que todo está determinado por dichas leyes, es decir que las causas determinan a los efectos; sin embargo, se asume igualmente que los efectos se deben a las causas, por lo que se incurre en petición de principio. La alternativa presupone que hay fenómenos que son genuinamente espontáneos, que no tienen causa; y ante ambos demonios del pensamiento muchos prefieren la visión determinista.

Pero se puede pensar también que lo que llamamos "leyes causales" no son sino modelos de las regularidades que se observan en el mundo. Son diagnósticos o interpretaciones que formulamos con base en las observaciones y el conocimiento previo. Y como tales se pueden revisar ante observaciones más generales y/o novedosas.

Lo *irregular* no se puede describir por los modelos de las regularidades. Si lo irregular se hace sistemático se convierte en parte de las regularidades. Si algo es regular tiene una entropía baja o moderada, pero lo irregular eleva la entropía. El universo es en general regular y la entropía es moderada o baja, y por eso se puede pensar en que hay leyes generales que lo rigen. Pero también hay irregularidades que se observan espontáneamente. Son los eventos que no tienen causa. Son los eventos genuinamente aleatorios que elevan la entropía. Empiezan en la física, pero se proyectan y amplifican a los fenómenos químicos y biológicos, donde son más frecuentes y hacen los fenómenos menos predecibles, y finalmente a la psicología, donde las leyes de conducta son muy débiles y las conductas intencionales prácticamente libres.

Los casos paradigmáticos son la visión indeterminista de la física cuántica, que postula que hay fenómenos que brotan espontáneamente, y la teoría de la evolución de Darwin, donde la recombinación genética que da origen a la evolución de las especies es un accidente que no tiene causa. El pensamiento-memoria asociativo, indeterminado y entrópico es un caso en el nivel psicológico. Las comunicaciones que intercambian los agentes racionales entre sí tienen un grado de indeterminación y la productividad de la toma de decisiones asume también la indeterminación del mundo. La indeterminación de la mente en conjunto con la indeterminación del mundo abre el espacio para el libre albedrío.

10.5. Entropía Cerebral

El cerebro es una máquina entrópica que mantiene un número muy significativo de estados. Un estado cerebral consiste en un patrón de actividad estable que involucra la activación y/o conectividad de diversas redes cerebrales de gran escala. Se han identificado varios de éstos, como los de reposo, alerta y meditación [98]. Los estados cerebrales pueden tener muchos sub-estados cuyas fluctuaciones influyen en las funciones cognitivas superiores [99] y es posible que la entropía del cerebro se relacione con el número de estados que se accesan durante el desempeño de sus funciones [100].

La entropía de un estado del cerebro se puede medir mediante imágenes de resonancia magnética funcional IRMf [101]. Para este efecto la imagen que se estudia se divide en vóxels, cada uno de los cuales tiene una señal que refleja el valor único del nivel de oxígeno en la sangre –*blood-oxygen-level-dependent signal* o BOLD– al momento de escaneo, el cual refleja a su vez el nivel de actividad en el vóxel y permite hacer un mapa de la actividad del cerebro cuando el sujeto

desempeña una función cerebral. La técnica consiste en registrar una secuencia o ventana temporal de valores BOLD y calcular la indeterminación del vóxel. Ésta se puede caracterizar por la llamada *entropía de muestreo* o *SampEn*. La metodología se ha utilizado para medir los cambios de la entropía cerebral durante el desempeño de tareas sensorio-motoras con los siguientes resultados [101]:

- La entropía provee una medida fisiológica y funcional significativa de la actividad del cerebro.
- Hay una reducción de la entropía en las áreas visuales y sensorio-motoras en las regiones cerebrales asociadas a la tarea relativa al estado de reposo.
- La entropía de las regiones de la neo-corteza es menor que la del resto del cerebro –el cerebelo, el tronco encefálico, la zona límbica, etc.,– en el estado de reposo.
- Zonas del cerebro con niveles particulares de entropía corresponden a áreas anatómicas o funcionales del cerebro.

Bajos niveles de entropía en la neo-corteza indican niveles superiores de organización que se requieren durante los procesos de interpretación, pensamiento y acción intencional, como se sugiere en el principio de energía libre en la hipótesis del cerebro Bayesiano [96]. Otras áreas del cerebro con mayores niveles de entropía constituyen maquinaria biológica estándar y no hacen interpretaciones propiamente. El experimento sugiere que los niveles de entropía relativamente altos en el estado de reposo permiten que el cerebro sea lo suficientemente flexible para enfrentar los cambios del entorno, ya que la entropía disminuye para lograr el foco y la especificidad que se requiere durante el desempeño de

funciones cognitivas superiores; sin embargo, la entropía no puede disminuir demasiado ya que la determinación absoluta es la condición de lo inerte.

Esta metodología se ha utilizado para medir la relación entre la entropía del cerebro en el estado de reposo y la inteligencia [100]. Para este efecto se utilizaron pruebas de inteligencia verbal y del desempeño de una tarea sensorio-motora. Los resultados mostraron que los sujetos con mayor IQ tenían mayores niveles de entropía. Este resultado es un tanto paradójico ya que está en conflicto con el decremento de entropía asociado a las funciones cognitivas superiores. Sin embargo, este resultado se explica en relación al compromiso de la entropía, el cual sugiere que una entropía muy baja se asocia a conductas rígidas y predeterminadas; sin embargo, si la entropía se incrementa significativamente el desempeño se verá igualmente afectado.

Aunque no hay una noción clara y ampliamente aceptada de estado del cerebro, y la medida de entropía sugerida pudiera ser una aproximación muy burda, las consideraciones presentes, en conjunto con la computación relacional indeterminada y la entropía computacional, sugieren que las regiones de la corteza neocortical son máquinas entrópicas, y consecuentemente no son Máquinas de Turing, y el cerebro como un todo no es una Máquina de Turing. Estas consideraciones sugieren asimismo que las estructuras corticales más antiguas deberían tener una entropía muy alta y consecuentemente no se deberían considerar máquinas computacionales. Asimismo, las funciones de control automático deben tener entropías muy bajas y se asemejan más a los dispositivos artificiales de control que a los sistemas de cómputo. Se sigue de esta reflexión que las máquinas computacionales naturales deben tener una entropía moderada, no muy baja ni muy alta, y que la computación natural se adhiere al compromiso de la entropía.

10.6. Retos Técnicos y Predicciones

Si la percepción, la acción, la conducta esquemática, el pensamiento y la memoria se simulan con computadoras digitales del tipo ordinario la inducción de interpretaciones, la síntesis de acciones y la toma de decisiones están predeterminadas. Sin embargo, si la simulación se hace con otros modos de computación que sean intrínsecamente entrópicos, tal como el relacional indeterminado, o posiblemente con computación analógica o cuántica, o incluso holografía, puede haber un grado de indeterminación del agente computacional adicional al del medio ambiente, y la toma de decisions puede ser libre en la misma medida. La teoría elaborada aquí presenta un reto para la construcción de maquinaria y procesos computacionales, así como de memorias asociativas que utilicen algoritmos mínimos que sean causales de la conducta.

De manera particular, la presente propuesta sugiere varias hipótesis que pueden ser sujetas de investigación empírica; por ejemplo en:

Psicología Evolutiva y Sociología:

- Productividad Potencial de las Decisiones: es posible definir volúmenes de control humanos y de especies animales no humanas, medir su entropía y contar los cambios conductuales productivos debidos a la comunicación. Si la predicción es válida se podría identificar un perfil τ. El valor de este parámetro predeciría fenómenos sociales, tales como el tamaño o grado de organización de grupos sociales.

- La entropía se relaciona con la filogenia del cerebro: las estructuras más antiguas tienen un nivel mayor de entropía en el estado de reposo, que se decrementa conforme aumenta la variabilidad y flexibilidad de estructu-

ras mas jóvenes. La neocorteza, asociada a las funciones ejecutivas, tiene la menor entropía en el estado de reposo, como lo sugieren los resultados mencionados arriba.

- La entropía cerebral de especies en el estado de reposo con estructuras neuronales más desarrolladas es menor a la entropía de animales con menor desarrollo neural en estados del cerebro análogos.

Neurociencias:

- Entropía de los módulos funcionales del cerebro: la entropía asociada a las estructuras que sostienen las habilidades perceptuales y motoras, así como la conducta esquemática, decrece en relación al estado de reposo para lograr la especificidad que se requiere en la interpretación y la acción. La entropía asociada al pensamiento deliberativo y a la memoria decrece también con respecto al estado de reposo, pero en menor grado, ya que estos procesos son menos determinados.

- La entropía asociada a las funciones de control automático es muy baja en relación a la interpretación, la memoria y el pensamiento deliberativo.

- La memoria de largo plazo tiene una entropía mayor que la memoria de trabajo en el estado de reposo. El olvido de manera natural ocurre si la entropía excede el rango operacional de la memoria, ya que los conceptos almacenados se confunden. Si la entropía de la memoria de trabajo es muy baja, por su parte, el pensamiento es esquemático y predecible.

- Trastornos mentales: trastornos de las redes de atención así como condiciones mentales particulares [83, 84] se asocian a niveles anormales de

entropía en relación al estado de reposo; por ejemplo, el trastorno obsesivo compulsivo se asocia a bajos niveles de entropía mientras que el déficit de atención e hiperactividad, así como condiciones esquizofrénicas se asocian a altos niveles; por su parte condiciones maniaco-depresivas se asocian a bajos y altos niveles de entropía, respectivamente.

Psicología Cognitiva:

- Pensamiento concreto *versus* abstracto: el decremento de la entropía en relación al estado de reposo es mayor en el pensamiento concreto que en el abstracto.

- Razonamiento proposicional *versus* diagramático: El decremento de la entropía en el razonamiento simbólico o proposicional, que es más algorítmico, es mayor que el decremento de la entropía cuando los problemas se presentan en formatos diagramáticos o analógicos, que pueden ser más naturales.

- La inteligencia se asocia a un mejor control ejecutivo, por lo que el nivel de entropía se correlaciona con la inteligencia dentro de su rango operacional. La correlación entre IQ y niveles de entropía altos [100] apoya esta predicción. Sin embargo, si el nivel de entropía del ejecutivo central excede su valor óptimo el IQ decrece, de acuerdo al compromiso de la entropía.

Una conjetura más fundamental es que la mente evolucionó a partir de la comunicación. Individuos que no se comunican son reactivos y su entropía es cero; consecuentemente no hacen interpretaciones y es posible que no tengan experiencias. La conducta esquemática es la forma paradigmática de experimentar al mundo, aunque de forma inconsciente, y es común en los seres humanos y

en los animales no humanos con un sistema nervioso suficientemente desarrollado. El pensamiento deliberativo es una forma de experiencia que involucra a la consciencia y a la toma de decisiones, y permite anticipar al mundo; y la utilidad de la comunicación se caracteriza por la productividad potencial de las decisiones. El compromiso de la entropía se puede correlacionar posiblemente con la experiencia y la consciencia en los seres humanos y en las especies animales que pueden experimentar al mundo. La racionalidad refleja qué tanto importa la comunicación y los contenidos de la mente para el individuo y la especie.

Capítulo 11

Principio de Racionalidad

La conducta es racional en la medida en que las hipótesis de interpretación perceptual y de acción sean adecuadas y coherentes con el mundo; en que la toma de decisiones, que se manifiesta en la intención y da lugar a la acción, se traduzca en sobrevivir y mejorar las condiciones de vida del agente y de su entorno; y en que la productividad potencial de la decisión, que abre el espacio de decisión, sea satisfactoria u óptima. Ésta es una proposición general a la que aquí nos referimos como *Principio de Racionalidad*. La racionalidad más acabada se da en los seres humanos cuya arquitectura cognitiva integra a la percepción, el pensamiento-memoria y la acción intencional, así como la experiencia y la consciencia.

El principio de racionalidad se sustenta en las siguientes consideraciones:

- La noción general de la acción;
- El Principio de Interpretación;
- El Principio de Acción;

- El proceso de toma de decisiones, que surge del pensamiento y permite anticipar las consecuencias de la acción;
- La entropía y la productividad de la decisión;
- El resultado a posteriori de la acción, que corresponde a la adaptación en la evolución.

La acción racional, como todas las acciones, es la respuesta a una necesidad, que a su vez es la manifestación de un desequilibrio del agente con el entorno o consigo mismo; y la acción tiende a reestablecer el equilibrio y a llevar al agente a un equilibrio más estable. Ésta es la noción general de Piaget que fundamenta a su teoría del desarrollo mental, como se discute en el Capítulo 5.1.

El Principio de Interpretación establece que la mejor estrategia para interpretar al mundo es seleccionar la hipótesis que resulte de ponderar las hipótesis potenciales que producen las habilidades perceptuales con el conocimiento almacenado en la memoria, de acuerdo a la concepción Bayesiana.

El Principio de Acción establece que la mejor estrategia para actuar consiste en realizar la acción que resulte de ponderar el grado en que las acciones potenciales satisfacen la intención con la capacidad y/o el costo de llevarlas a cabo. Este principio produce también una hipótesis, pero en este caso acerca de los propios recursos del agente para cambiar al mundo y satisfacer sus intenciones.

La toma de decisiones tiene como insumos al producto de la percepción, que siempre es una hipótesis; al conocimiento del agente, que frecuentemente es incompleto e implícito, y que se traduce en un diagnóstico acerca de las causas de los hechos observados en el mundo; y a sus intereses y su lógica afectiva, que involucran aspectos subjetivos. La decisión es una hipótesis acerca de los deseos e intereses del agente y determina la intención, que se traduce en la especifica-

ción de un plan de acción propiamente, y la racionalidad consiste en realizar las acciones incluidas en el plan para cambiar el entorno a un estado que se cree más favorable.

Se actúa bajo la hipótesis de que la acción se traducirá en los efectos esperados; la acción se decide bajo la hipótesis de que su realización satisfacerá las intenciones que se tienen; las intenciones son hipótesis de que las decisiones reflejan los deseos e intereses del agente; y las decisiones se toman bajo la hipótesis de que las interpretaciones que provee la percepción son coherentes con el mundo, que los diagnósticos son adecuados y que los deseos son consistentes con los intereses, preferencias y valores del agente.

Asimismo, las decisiones se toman por mentes y en entornos con un grado de indeterminación. La productividad de las decisiones es alta para valores moderados de la entropía de la mente del agente y del entorno. El nivel normal de entropía depende de la naturaleza del agente o mecanismo y su desviación refleja una anomalía. Las conductas automáticas reflejan una entropía baja o nula, como en las conductas estereotipadas u obsesivas, y en el extremo opuesto, la conducta errática, como en el caso de los trastornos de la atención, refleja una entropía alta, y ambos casos corresponden a una productividad de las decisiones baja, que se traduce en conductas irracionales.

Por estas razones la racionalidad sólo se puede apreciar a posteriori por el efecto de la acción en el agente y en su entorno. De la misma forma que los accidentes genéticos son aleatorios, y que la característica a la que dan lugar contribuye o dificulta la adaptación del individuo y de la especie, qué tan racional es la conducta depende del grado en que la acción mejore la calidad de vida del agente y de sus efectos en el mundo. No es un juicio que hace el individuo u otros agentes: es el resultado neto de la acción en el entorno.

La conducta irracional se da cuando se viola el principio de racionalidad; es decir, cuando las hipótesis de interpretación perceptual y de acción intencional son pobres; cuando las decisiones no son consistentes con los intereses del agente, y cuando la productividad potencial de la decisión es baja. La irracionalidad se debe a disfunciones de alguno de los módulos de la cognición, que limitan al agente a actuar de manera productiva para sí mismo y/o para su comunidad y entorno.

Las disfunciones pueden ser congénitas o debidas a un traumatismo. Las del módulo de pensamiento-memoria se manifiestan en problemas psicológicos. Un síntoma de la irracionalidad es la alteración de la indeterminacón de la mente. La conducta productiva requiere de un nivel moderado de entropía en el que sea posible tomar decisiones productivas y la irracionalidad se da en la medida de las desviaciones. Es también posible que este módulo se "apague" y la salida de la percepción incida directamente en la acción, con el consecuente automatismo. Asimismo, disfunciones de los módulos de la percepción y la acción producen interpretaciones empobrecidas o erróneas y la selección de acciones inadecuadas o no factibles que se traducen en irracionalidad.

La conducta racional es la expresión de una mentalidad sana y un nivel adecuado de entropía, y un nivel anormal de entropía es síntoma de discapacidad mental.

La arquitectura cognitiva embebe al ciclo de inferencia, al principio de interpretación, al principio de la acción y al principio de racionalidad, y permite describir los niveles de racionalidad de mecanismos artificiales y de individuos de diferentes especies naturales. El grado o nivel de racionalidad dependerá del desarrollo y estructura de los módulos y submódulos de la arquitectura cognitiva, como se ilustra en las Figuras 9.1 y 10.3 para el caso de la cognición artificial

y natural respectivamente. La especie humana es la referencia de la arquitectura cognitiva más acabada y las limitaciones funcionales o estructurales respecto a la misma sugieren o corresponden a niveles de racionalidad limitados o disminuidos, ya sea por características de la especie, por el estadio del desarrollo mental o por disfunciones del agente.

Los niveles de racionalidad se pueden postular limitando la funcionalidad o eliminando los componentes de la arquitectura cognitiva. De manera esquemática se pueden postular tres niveles de los más simples a los más acabados:

- Incluye sólo observaciones y acciones básicas
- Incluye habilidades perceptuales y motoras
- Incluye el pensamiento-memoria

Las especies en cada nivel pueden tener una gama muy amplia de funcionalidad y estructura, y la variación es mayor si se consideran las diferencias y las disfunciones individuales, pero cada nivel tiene algunas propiedades generales y limitaciones distintivas: nivel 1) sólo tienen conductas reactivas, por lo que sólo pueden habitar entornos simples y estables; nivel 2) cuentan además con habilidades que se pueden entrenar, que permiten aumentar la autonomía y habitar entornos más complejos; incluso pueden comunicarse y tener algún grado de experiencia; y nivel 3) tienen memoria de trabajo y un ejecutivo central, que les permite tomar decisiones y anticipar al mundo; consecuentemente, pueden habitar en entornos más complejos y variables; las especies más avanzadas en el tercer nivel gozan de memoria semántica y posiblemente de memoria episódica, pueden aprender conceptualmente, gozan del lenguaje y del pensamiento simbólico, tienen intereses, valores, voluntad y una lógica afectiva, y son plenamente conscientes.

Estas distinciones son por supuesto especulativas y no hay necesidad de que el desarrollo del cerebro culmine en los humanos. Hay otras especies que cuentan con otras estructuras más desarrolladas, como el bulbo olfatorio, en los perros y los tiburones, lo que puede dar lugar a otras formas de experiencia, y otras cuya neocorteza es similar o aún mayor en tamaño que la humana, como los delfines, cuyo ejecutivo central y su memoria de trabajo son similares a los primates no humanos más desarrollados, cuentan con memoria de largo plazo y tienen capacidades acústicas que exceden significativamente a las humanas, que pudieran dar lugar a otras formas de experiencia y de consciencia.

El presente Principio de Racionalidad pone en perspectiva otras nociones de racionalidad, como el principio de racionalidad de Newell y Simon, basado en la hipótesis del símbolo aterrizado [19] y concebido como la única ley de comportamiento en el nivel del conocimiento [21]. Este modelo ha tenido y sigue teniendo una influencia profunda en la toma de decisiones en la economía y la psicología. Sin embargo, esta visión se limita al módulo del pensamiento. En ésta, las interpretaciones están dadas, la abducción y la planeación se embeben en la toma de decisión en el módulo del pensamiento, y la decisión se identifica directamente con la acción. Asimismo, la autonomía de este tipo de agentes no toma en cuenta los intereses, los valores, la voluntad y el libre albedrío, que son insumos centrales a la toma de decisiones y la racionalidad humana, y los agentes artificiales no pueden ejercer la libertad. En este modelo, los dispositivos de entrada y salida, que corresponden a la percepción y la acción, son contingentes. El caso paradigmático de los juegos racionales identifica a la racionalidad con el pensamiento simbólico, aislado del mundo y desacoplado de la acción.

Dicha visión contrasta con la de la Inteligencia Artificial sub-simbólica, basada en las redes de aprendizaje profundo y el aprendizaje por refuerzo de hoy en

día. Estas herramientas se utilizan para modelar habilidades tanto perceptuales como motoras –los términos $P(O/E)$ y $P(I/A)$ de la Figura 9.1 respectivamente– a partir de observaciones y acciones básicas –O y A– que son dadas. En la práctica se utilizan para construir clasificadores y/o predictores, independientemente de consideraciones cognitivas. Sin embargo, este enfoque carece de los módulos centrales de pensamiento y memoria, y conecta directamente al producto de la percepción con la entrada de la acción y no hay realmente una toma de decisión. Asimismo, al no haber memoria, no hay conocimiento previo de los eventos y de las acciones –los términos $P(E)$ y $P(A)$– que se aprenden; y las interpretaciones y las acciones intencionales –$P(E/O)$ y $P(A/I)$– dependen sólo de las habilidades; consecuentemente las interpretaciones son pobres y las acciones no son refinadas.

Los modelos sub-simbólicos son también muy frágiles ante los sesgos cognitivos debidos tanto al pensamiento esquemático como a los datos de entrenamiento que subyacen a las habilidades adquiridas, que reflejan a los sesgos del entorno, tanto para modelar a la percepción como a la acción. Sin embargo, esto no es diferente a la cognición natural que depende de las contingencias de vida: lugar y lenguaje natal; condición socio-económica; preferencias políticas y religiosas en el entorno, etc. Sin embargo, los seres humanos podemos superar los sesgos a través del conocimiento y la reflexión, y modular tanto a la interpretación como a la acción intencional. Pero, para que el conocimiento pueda incidir sobre la percepción y la acción es necesario contar con el módulo pensamiento-memoria.

Asimismo, en los modelos esquemáticos la memoria de trabajo y el ejecutivo central juegan un papel muy limitado. Por ejemplo, en los coches autónomos o los drones, la salida del módulo de la percepción incide directamente sobre el de

la acción. Consecuentemente, no hay una anticipación del mundo producto de la reflexión.

AlphaZero se puede ver como un caso mixto en que la habilidad perceptual modelada por la red profunda se expresa como una tabla en la memoria. La red profunda se comporta en este caso como una máquina lógica que hace explícito el conocimiento que está implícito en las reglas del juego. Sin embargo, el sistema nunca interactúa con el mundo y la percepción es sólo aparente; el contenido conceptual en la memoria tampoco se deriva de la experiencia empírica, y en ambos escenarios hay una profunda anomalía. En esto AlphaZero se diferencia radicalmente del jugador humano cuya experiencia es realmente empírica. Es un mejor modelo del juego, pero sigue encapsulando a la racionalidad en un tubo de ensayo.

La cognición artificial y la natural se diferencian también en que los agentes artificiales son máquinas y no tienen experiencias; por lo mismo carecen de emociones y sentimientos; la cognición natural goza además de la consciencia, que es una forma aumentada de experiencia que permite inspeccionar y sentir los procesos de la mente.

Bibliografía

[1] H. Simon, Models of Man, Social and Rational: Mathematical Essays on Rational Human Behavior in a Social Setting., Wiley, New York, 1957.

[2] D. E. Rumelhart, J. L. McClelland, the PDP Research Group, Parallel Distributed Processing, Explorations in the Microstructure of Cognition, Vol.1: Foundations, The MIT Press, Cambridge, Mass., 1986.

[3] R. Brooks, Intelligence without representation, Artificial Intelligence 47 (1991) 139–159.

[4] M. L. Anderson, Embodied cognition: A field guide, Artificial Intelligence 149 (2003) 91–130.

[5] T. Froese, T. Ziemke, Enactive artificial intelligence: Investigating the systemic organization of life and mind, Artificial Intelligence 173 (2009) 466–500.

[6] J. Pearl, Causality, 2nd Edition, Cambridge University Press, Cambridge, UK, 2009. doi:10.1017/CBO9780511803161.

[7] L. E. Sucar, Probabilistic Graphical Models Principles and Applications, 1st Edition, Advances in Computer Vision and Pattern Recognition, Springer London, London, 2015.

[8] A. Tversky, D. Kahneman, Judgment under uncertainty: Heuristics and biases, Science 185 (1974) 1124–1131. doi:10.1126/science.185.4157.1124.

[9] H. Simon, A behavioral model of rational choice, The Quarterly Journal of Economics 60 (1) (1955) 99–118.

[10] M. Minsky, The Society of Mind, Simon and Schuster, New York, 1986.

[11] J. Koehler, The base rate fallacy reconsidered: Descriptive, normative and methodological challenges, Behavioral and Brain Sciences 19 (1996) 1–53. doi:10.1017/S0140525X00041157.

[12] L. Cosmides, J. Tooby, Are humans good intuitive statisticians after all? rethinking some conclusions of the literature on judgment under uncertainty, Cognition 58 (1996) 1–73.

[13] P. M. Todd, G. Gigerenzer, Ecological rationality: Intelligence in the World, New York: Oxford University Press, 2012. http://hdl.handle.net/11858/001M-0000-0024-EE01-A

[14] A. K. Barbey, S. A. Sloman, Base-rate respect: From ecological rationality to dual process, Behavioral and Brain Sciences 30 (2007) 241–254. doi:doi:10.1017/S0140525X07001653.

[15] A. M. Turing, On computable numbers, with an application to the entscheidungs problem, Proceedings of the London Mathematical Society 42 (1936) 230–265.

[16] G. S. Boolos, R. C. Jeffrey, Computability and Logic (Third Edition), Cambridge University Press, 1989.

[17] M. D. Davis, Computability and Unsolvability, McGraw-Hill Series in Information Processing and Computers, McGraw-Hill, 1958.

[18] D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, T. Lillicrap, K. Simonyan, D. Hassabis, General reinforcement learning algorithm that masters chess, shogi, and go through self-play, Science 362 (2018) 1140–1144. doi:10.1126/science.aar6404.

[19] A. Newell, H. Simon, Computer science as empirical inquiry: Symbols and search, Communications of the ACM 19 (3) (1976) 113–126.

[20] H. A. Simon, The Sciences of the Artificial, 3rd Edition, MIT Press, Cambridge, MA, 1996.

[21] A. Newell, The knowledge level, Artificial Intelligence 18 (1982) 87–127.

[22] J. E. Laird, The SOAR Cognitive Architecture, MIT Press, Cambridge, MA, 2012.

[23] J. R. Anderson, D. Bothell, M. D. Byrne, S. Douglass, C. Lebiere, Y. Qin, An integrated theory of the mind, Psychological Review 111 (2004) 1036–1060.

[24] L. A. Pineda, N. Hernández, A. Rodríguez, G. Fuentes, R. Cruz, Deliberative and conceptual inference in service robots, Applied Sciences 11 (4) (2021) 1523. doi:https://doi.org/10.3390/app11041523.

[25] R. Reiter, A logic for default reasoning, Artificial Intelligence 13 (1980) 81—132.

[26] E. Tulving, Memory systems: episodic and semantic memory, in: E. Tulving, W. Donaldson (Eds.), Organization of Memory, New York: Academic Press, 2013, pp. 381–403.

[27] L. A. Pineda, A. Rodríguez, G. Fuentes, C. Rascón, I. V. Meza, A light non-monotonic knowledge-base for service robots, Intel Serv Robotics 10 (2017) 159–171.

[28] L. A. Pineda, A. Rodríguez, G. Fuentes, N. Hernández, M. Reyes, C. Rascón, R. Cruz, I. Vélez, H. Ortega, Opportunistic inference and emotion in service robots, Journal of Intelligent & Fuzzy Systems, 34 (5) (2018) 3301–3311.

[29] I. Torres, N. Hernández, A. Rodríguez, G. Fuentes, L. A. Pineda, Reasoning with preferences is service robots, Journal of Intelligent & Fuzzy Systems 36 (5) (2019) 5105–5114.

[30] L. A. Pineda, N. Hernández, I. Torres, G. Fuentes, N. Pineda-De-Ávila, Practical non-monotonic knowledge-base system for un-regimented domains: A case-study in digital humanities, Information Processing & Management 57 (3) (2020) 102214.

[31] G. Brewka, T. Eiter, M. Truszczyński, Answer set programming at a glance, Communications of the ACM 54 (12) (2011) 92—103.

[32] H. L. Levesque, R. Brachman, A fundamental tradeoff in knowledge representation and reasoning, in: R. Brachman, H. Levesque (Eds.), Readings in Knowledge Representation, Morgan and Kaufmann, Los Altos, CA, 1985, pp. 41–70.

[33] H. L. Levesque, Logic and the complexity of reasoning, Journal of Philosophical Logic 17 (1988) 355–389.

[34] L. A. Pineda, Conservation principles and action schemes in the synthesis of geometric concepts, Artificial Intelligence 171 (2007) 197–238.

[35] J. Piaget, Seis Estudios de Psicología, Barral Editores, S. A., Barcelona, 1970.

[36] A. M. Turing, Computing machinery and intelligence, Mind 59 (1950) 433—460.

[37] C. E. Shannon, A mathematical theory of communication, The Bell System Technical Journal 27 (3) (1948) 379–423. doi:10.1002/j.1538-7305.1948.tb01338.x.

[38] J. E. Hopcroft, R. Motwani, J. D. Ullman, Introduction to Automata Theory, Languages, and Computation (3rd Edition), Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2006.

[39] J. Martin, Introduction to Languages and the Theory of Computation (Third edition), McGraw Hill, 2003.

[40] L. A. Pineda, Entropy, computing and rationality (2020). arXiv:2009.10224.

[41] L. A. Pineda, G. Fuentes, R. Morales, An entropic associative memory, Scientific Reports 11 (6948) (2021) 1–15.
doi:10.1038/s41598-021-86270-7.
https://www.nature.com/articles/s41598-021-86270-7

[42] L. A. Pineda, A distributed extension of the Turing Machine, CoRR abs/1803.10648. arXiv:1803.10648.
http://arxiv.org/abs/1803.10648

[43] L. A. Pineda, The Mode of Computing, CoRR abs/1903.10559.
arXiv:1903.10559.
http://arxiv.org/abs/1903.10559

[44] G. E. Hinton, J. L. McClelland, D. E. Rumelhart, Distributed representations (chapter 3), in: D. E. Rumelhart, J. L. McClelland (Eds.), Parallel Distributed Processing, Explorations in the Microstructure of Cognition, Vol.1: Foundations, The MIT Press, Cambridge, Mass., 1986.

[45] J. R. Anderson, G. H. Bower, Human Associative Memory: A Brief Edition, Lawrence Erlbaum Associates, Publishers, Hillsdale, New Jersey, 1980.

[46] F. C. Bartlett, Remembering: A study in experimental and social psychology, Cambridge University Press, 1932.

[47] S. M. Kosslyn, W. L. Thomson, G. Ganis, The Case for Mental Imagery, Oxford University Press, 2006.

[48] M. R. Quillian, Semantic memory, in: M. Minsky (Ed.), Semantic Information Processing, MIT Press, 1968, pp. 227–270.

[49] K. Steinbuch, Die lernmatrix, Kybernetik 1 (1) (1961) 36–45.

[50] D. J. Willshaw, O. P. Buneman, H. C. Longuet-Higgins, Non-holographic associative memory, Nature 222 (1969) 960–962.

[51] T. Kohonen, Correlation matrix memories, Computers, IEEE Transactions on C-21 (1972) 353 – 359. doi:10.1109/TC.1972.5008975.

[52] G. Palm, On associative memory, Biological Cybernetics 36 (1) (1980) 19–36. doi:10.1007/BF00337019.

[53] J. J. Hopfield, Neural networks and physical systems with emergent collective computational abilities, Proceedings of the National Academy of Sciences of the USA 79 (8) (1982) 2554–2558.

[54] B. Kosko, Bidirectional associative memories, IEEE Transactions on Systems, man, and Cybernetics 18 (1) (1988) 49–60.

[55] I. Aleksander, W. Thomas, P. Bowden, Wisard, a radical new step forward in image recognition, Sensor Rev 4 (3) (1984) 120–124.

[56] I. Aleksander, An introduction to neural computing, Chapman and Hall, London, 1990.

[57] G. X. Ritter, P. Sussner, J. L. DiazdeLeon, Morphological associative memories, IEEE Transaction on Neural Networks 9 (2) (1998) 281–293.

[58] G. X. Ritter, J. L. DiazdeLeon, P. Sussner, Morphological bidirectional associative memories, Neural Networks 12 (1999) 851–867.

[59] P. Sussner, M. E. Valle, Implicative fuzzy associative memories, IEEE Transactions on Fuzzy Systems 14 (6) (2006) 793–807.

[60] P. Sussner, T. Schuster, Interval-valued fuzzy morphological associative memories: Some theoretical aspects and applications, Information Sciences 438 (2018) 127–144. doi:doi.org/10.1016/j.ins.2018.01.042.

[61] H. Ramsauer, B. Schäfl, J. Lehner, P. Seidl, M. Widrich, T. Adler, L. Gruber, M. Holzleitner, M. Pavlović, G. K. Sandve, V. Greiff, D. Kreil, M. Kopp, G. Klambauer, J. Brandstetter, S. Hochreiter, Hopfield networks is all you need (2020). arXiv:2008.02217.

[62] J. A. Fodor, Z. W. Pylyshyn, Connectionism and cognitive architecture: A critical analysis, Cognition 28 (1-2) (1988) 3–71.

[63] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (11) (2015) 436–444. doi:10.1038/nature14539.

[64] G. E. Hinton, R. R. Salakhutdinov, Reducing the Dimensionality of Data with Neural Networks, Science 313 (5786) (2006) 504–507. doi:10.1126/science.1125249.

[65] J. Masci, U. Meier, D. Cireşan, J. Schmidhuber, Stacked Convolutional Auto-Encoders for Hierarchical Feature Extraction, in: T. Honkela, W. Duch, M. Girolami, S. Kaski (Eds.), Artificial Neural Networks and Machine Learning – ICANN 2011, Vol. 6791 of Lecture No-

tes in Computer Science, Springer, 2011, pp. 52–59. doi:10.1007/978-3-642-21735-7_7.

[66] S. Basu, M. Karki, S. Ganguly, R. DiBiano, S. Mukhopadhyay, R. Nemani, Learning sparse feature representations using probabilistic quadtrees and deep belief nets (2015). arXiv:1509.03413.

[67] A. Kortylewski, Q. Liu, H. Wang, Z. Zhang, A. Yuille, Combining compositional models and deep networks for robust object classification under occlusion (2020). arXiv:1905.11826.

[68] D. Krotov, J. J. Hopfield, Dense associative memory for pattern recognition, in: Advances in Neural Information Processing Systems, Vol. 29, 2016, pp. 1172–1180. arXiv:1606.01164.

[69] H. He, Y. Shang, X. Yang, Y. Di, J. Lin, Y. Zhu, W. Zheng, J. Zhao, M. Ji, L. Dong, N. Deng, Y. Lei, Z. Chai, Constructing an associative memory system using spiking neural network, Frontiers in Neuroscience 13 (2019) 650. doi:10.3389/fnins.2019.00650.

[70] D. Marr, Vision: A Computational Investigation into the Human Representation and Processing of Visual Information, Henry Holt and Co., Inc., New York, NY, USA, 1982.

[71] Aristoteles, Obras, Aguilar S. A. de ediciones, Madrid, 1964.

[72] C. Shields, Aristotle's psychology, in: E. N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy, winter 2016 Edition, Metaphysics Research Lab, Stanford University, 2016.

[73] N. Chomsky, Syntactic Structures, The Hague/Paris: Moutons, 1957.

[74] N. Chomsky, A review of B. F. Skinner's Verbal Behavior, Language 35 (1) (1959) 26–58.

[75] J. A. Fodor, The Language of Thought, Harvard University Press, 1975.

[76] B. C. Smith, Prologue to "Reflection and Semantics in a Procedural Language", in: H. L. R. Brachman (Ed.), Readings in Knowledge Representation, Morgan and Kaufmann, Los Altos, CA, 1985, pp. 31–40.

[77] J. R. Searle, Minds, brains, and programs, Behavioral and Brain Sciences 3 (3) (1980) 417–457.

[78] D. J. Chalmers, Syntactic transformations on distributed representations, Connection Science 2 (1–2) (1990) 53–62. doi:10.1080/09540099008915662.

[79] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proceedings of the IEEE 86 (11) (1998) 2278–2324.

[80] J. G. Rueckl, Connectionism and the notion of levels, in: T. Horgan, J. Tienson (Eds.), Connectionism and the Philosophy of Mind. Studies in Cognitive Systems, vol 9, Springer, Dordrecht, 1991.

[81] H. T. Siegelmann, E. D. Sontag, On the computational power of neural nets, Journal of Computer and System Sciences 50 (1) (1995) 132–150. doi:org/10.1006/jcss.1995.1013.

[82] G. Z. Sun, H. H. Chen, Y. C. Lee, C. L. Giles, Turing equivalence of neural networks with second order connection weights, in: Anon (Ed.), Pro-

ceedings. IJCNN - International Joint Conference on Neural Networks, Publ by IEEE, 1992, pp. 357–362.

[83] M. Posner, M. K. Rothbart, Research on attention networks as a model for the integration of psychological science, Annu Rev Psychol 58 (2007) 1–23.

[84] M. Posner, M. K. Rothbart, H. Ghassemzadeh, Restoring attention networks, Yale Journal of Biology and Medicine 92 (2019) 139–143.

[85] J. A. Fodor, The mind-body problem, Scientific American 244 (1981) 114–123.

[86] S. M. Kosslyn, W. L. Thompson, G. Ganis, The Case for Mental Imagery, Oxford Univ. Press, 2006.

[87] R. Shepard, L. Cooper, Mental images and their transformations, MIT Press, Cambridge, Mass., 1982.

[88] M. Tye, The Imagery Debate, A Bradford Book, The MIT Press, Cambridge, Mass., 1991.

[89] Z. W. Pylyshyn, What the mind's eye tells the mind's brain: A critique of mental imagery, Psychological Bulletin 80 (1973) 1–24.

[90] B. J. Copeland, The church-turing thesis, in: E. N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy, spring 2019 Edition, Metaphysics Research Lab, Stanford University, 2019.

[91] J. Hopcroft, R. Motwani, J. D. Ullman, Introduction to Automata Theory, Languages, and Computation (3rd Edition), Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2006.

[92] D. Chalmers, The Conscious Mind: In Search of a Fundamental Theory, Oxford University Press, Inc., New York, NY, USA, 1996.

[93] W. Casscells, A. Schoenberger, T. B. Graboys, Interpretation by physicians of clinical laboratory results, N Engl J Med 299 (18) (1978) 999–1001. doi:10.1056/NEJM197811022991808.

[94] M. Bar-Hillel, The base-rate fallacy in probability judgments, Acta Psychologica 44 (3) (1980) 211–233. doi:doi.org/10.1016/0001-6918(80)90046-3.

[95] G. Gigerenzer, U. Hoffrage, How to improve bayesian reasoning without instruction: Frequency formats, Psychological Review 102 (1995) 684–704.

[96] K. J. Friston, The free-energy principle: a unified brain theory?, Nature Reviews Neuroscience 11 (2) (2010) 127–138.
URL https://doi.org/10.1038/nrn2787

[97] A. D. Baddeley, The concept of working memory: A view of its current state and probable future, Cognition 10 (1981) 17–23.

[98] Y. Y. Tang, M. K. Rothbart, M. Posner, Neural correlates of establishing, maintaining, and switching brain states, Trends in Cognitive Science 16 (6) (2018) 330–337.
URL https://doi.org/10.1016/j.tics.2012.05.001

[99] E. Zagha, D. A. McCormick, Neural control of brain state., Curr Opin Neurobiol 29 (2014) 178–186.
URL https://doi.org/10.1016/j.conb.2014.09.010

[100] G. N. Saxe, D. Calderone, L. J. Morales, Brain entropy and human intelligence: A resting-state fmri study, PLoS ONE 13 (2) (2018) e0191582.
URL https://doi.org/10.1371/journal.pone.0191582

[101] Z. Wang, Y. Li, A. R. Childress, J. A. Detre, Brain entropy mapping using fmri, PLoS ONE 9 (3).
URL https://doi.org/10.1371/journal.pone.0089948

Racionalidad Computacional,
se imprimió en marzo de 2021 en el taller de Agys Alevín S.C.
Plásticos 84 local 2 Ala Sur, Fracc. Industrial Alce Blanco,
Naucalpan de Juárez, Estado de México CP 53370.
En su composición se utilizó tipo Garamond.
Impreso en papel couché mate de 115 grs.
La edición consta de 300 ejemplares.