100 COSAS QUE HAY QUE SABER SOBRE INTELIGENCIA ARTIFICIAL

Ramon López de Mántaras Badia

CIENX100

Lectio Ediciones

100 cosas que hay que saber sobre inteligencia artificial

• Colección Cien × 100 – 37 •

100 cosas que hay que saber sobre inteligencia artificial

Ramon López de Mántaras Badia

Instituto de Investigación en Inteligencia Artificial
Consejo Superior de Investigaciones Científicas

Lectio Ediciones

Primera edición: septiembre de 2024

© de esta edición:
9 Grupo Editorial
Lectio Ediciones
C/ Mallorca, 314, 1º 2ª B – 08037 Barcelona
Tel. 977 60 25 91 – 93 363 08 23
lectio@lectio.es
www.lectio.es

Diseño y composición: 3 × Tres

Impresión: Romanyà Valls, SA

ISBN: 978-84-18735-68-4

DL T 804-2024

ÍNDICE

Machinas ad sanitatis

Cambio climático, biodiversidad y geociencia

Explorando y observando el universo

Machina ludens

Creatividad en artes y ciencias

Grandes retos

AGRADECIMIENTOS

He tenido la fortuna de contar con la ayuda de Jordina Biosca, quien ha revisado exhaustiva y pacientemente todo el contenido del libro con gran sentido crítico. Sin duda, sus comentarios han mejorado significativamente el resultado final. Muchas gracias, Jordina.

La maestra Carme Ortoll ha revisado muchos capítulos del libro y sus pertinentes observaciones también han contribuido mucho a mejorarlo. Muchas gracias, Carme.

El Dr. Cesc Múrria, ecólogo, ha revisado los capítulos sobre biodiversidad. Sus observaciones me han permitido aprender mucho sobre este importante tema. Muchas gracias, Cesc.

Finalmente, quiero agradecer a Jordi Ferré y Josep Maria Olivé, director y director adjunto de 9 Grup Editorial, el haber confiado en mí al proponerme, en 2021, que escribiera la versión original en catalán de este libro, que fue publicada, en septiembre de 2023, por la editorial Cossetània. También quisiera agradecer a Marta Ferré, responsable de comunicación y prensa de Cossetània, su excelente labor de comunicación y promoción.

PREFACIO

A pesar de los impresionantes éxitos recientes de la Inteligencia Artificial (IA), en particular la IA generativa, actualmente todavía nos encontramos con importantes dificultades para que una máquina sea capaz de llevar a cabo tareas que para nosotros son sencillas. Posiblemente la lección más importante que hemos aprendido a lo largo de los casi setenta años de existencia de la IA, es que lo que parecía más difícil, como diagnosticar enfermedades, jugar al ajedrez, o al Go, ya se ha logrado; en cambio, lo que parecía más fácil, como entender el significado profundo del lenguaje o la interpretación general de escenas, ha resultado ser tan difícil que todavía no lo hemos conseguido. Las capacidades más complicadas de alcanzar son aquellas que requieren interaccionar con entornos no restringidos que requieren entender el mundo mediante la percepción y comprensión del lenguaje, así como tomar decisiones con información incierta e incompleta.

Como se verá a lo largo del libro, la comprensión profunda del lenguaje y de lo que percibimos con nuestros sentidos solo es posible si, entre otras cosas, tenemos conocimientos de sentido común. La adquisición de conocimientos de sentido común es el principal problema con el que se enfrenta la IA. Poseer sentido común es el requerimiento fundamental para que las máquinas den el salto cualitativo desde la IA especializada a la IA de tipo general. Hay millones de conocimientos de sentido común que las personas utilizamos fácilmente y que nos permiten comprender cómo es y cómo funciona el mundo que nos rodea pero que las máquinas no poseen.

A pesar de estas dificultades, las tecnologías basadas en la IA han empezado ya a cambiar nuestras vidas en aspectos como la

salud, la productividad o el ocio. A corto y medio plazo tendrán un gran impacto en la energía, el transporte, la educación y en nuestras actividades domésticas, así como en actividades artísticas. Entre las actividades futuras, creo que los temas de investigación más importantes seguirán siendo el aprendizaje automático, los sistemas multiagente, el razonamiento, la planificación de acciones, la visión artificial, la comunicación multimodal persona-máquina, la robótica humanoide y los robots sociales.

Este libro, cuya versión en catalán se publicó en septiembre de 2023, aborda todos estos aspectos, desde los inicios de la IA hasta los desarrollos recientes, haciendo énfasis en sus limitaciones tanto técnicas como éticas y su impacto social. No he tenido la pretensión de ser exhaustivo, ya que es imposible teniendo en cuenta la enorme cantidad y variedad de resultados existentes. Por otra parte, la IA está avanzando rápidamente y por lo tanto muchos de los resultados que se presentan serán superados a corto plazo, pero creo que este libro muestra una foto fiel de la situación de la IA actual. He priorizado mostrar resultados de investigación recientes, a veces curiosos, que han sido mayoritariamente publicados en revistas del más alto prestigio. Solo me queda esperar que este libro aporte respuestas a algunas preguntas y sobre todo que plantee otras. También espero que los lectores disfruten leyendo tanto como yo he disfrutado escribiendo.

Instituto de Investigación en Inteligencia Artificial, Bellaterra,
y Western Sydney University, Sydney, 1 de febrero de 2023

Nota: A pesar de no ser estrictamente lo mismo, los términos: *software*, *sistema*, *agente* y *algoritmo* se usan indistintamente en este libro.

EL INICIO DE UN LARGO CAMINO

01/100

EL LARGO CAMINO HACIA LA INTELIGENCIA ARTIFICIAL

¿Es posible construir máquinas inteligentes? ¿Es el cerebro una máquina? Estas son dos preguntas que han obsesionado a grandes pensadores durante siglos. El desarrollo de la IA ha acercado estas dos preguntas e incluso, para muchos investigadores, las ha unificado en el sentido de que se están utilizando conceptos, técnicas y experimentos similares en los intentos de diseñar máquinas inteligentes e investigar la naturaleza de la mente. Actualmente sabemos todavía relativamente poco sobre el cerebro, pero estamos siguiendo un camino que implica considerarlo un sistema computacional y hemos empezado a explorar el espacio de posibles modelos computacionales que permitan emular su funcionamiento.

El objetivo último de la IA, conseguir que una máquina tenga una inteligencia de tipo general, similar a la humana, es uno de los objetivos más ambiciosos que se ha planteado la ciencia. Por su dificultad, es comparable a otros grandes objetivos científicos como explicar el origen de la vida, el origen del universo o conocer la estructura de la materia. A lo largo de los últimos siglos, este empeño por construir máquinas inteligentes nos ha conducido a inventar modelos o metáforas del cerebro humano. Por ejemplo, en el siglo XVII, Descartes se preguntó si un complejo sistema mecánico compuesto de engranajes, poleas y tubos podría, en principio, emular el pensamiento. Dos siglos después, la metáfora fue los sistemas telefónicos, puesto que parecía que sus conexiones se podían asimilar a una red neuronal. Actualmente el modelo dominante es el modelo computacional basado en el ordenador digital y, por lo tanto, es el modelo que hay detrás de todo lo que contiene este libro.

Hasta muy recientemente, el modelo dominante en IA ha sido el simbólico. Es un modelo *top-down* basado en el razonamiento lógico y la búsqueda heurística como pilares para la resolución de problemas. Es decir, la IA simbólica opera con representaciones abstractas del mundo real que se modelan mediante lenguajes de representación basados principalmente en la lógica matemática y sus extensiones. Simultáneamente con la IA simbólica también se ha desarrollado una IA bioinspirada llamada conexionista. Los sistemas conexionistas, contrariamente a la IA simbólica, siguen una modelización *bottom-up*, ya que se basan en la hipótesis de que la inteligencia emerge a partir de la actividad distribuida de un gran número de unidades interconectadas que procesan la información en paralelo. En IA conexionista, estas unidades son modelos muy aproximados de la actividad eléctrica de las redes neuronales biológicas.

A pesar de lo que últimamente afirman algunos sobre la generalidad de la IA generativa, basada en modelos conexionistas, estamos todavía muy lejos de conseguir una IA de tipo general. De hecho, prácticamente todos los esfuerzos en IA se han centrado en construir inteligencias artificiales especializadas y los logros alcanzados, en sus casi setenta años de existencia, son muy impresionantes; en particular durante el último decenio, principalmente gracias a la conjunción de dos elementos: la disponibilidad de enormes cantidades de datos y el acceso a la computación de altas prestaciones para poder analizarlos.

El camino hacia la IA de tipo general seguirá siendo, pues, largo y difícil. Al fin y al cabo, la IA tiene solo siete décadas de existencia y, como diría Carl Sagan, setenta años son un brevísimo momento en la escala cósmica del tiempo; o, como muy poéticamente dijo Gabriel García Márquez:

> Desde la aparición de vida visible en la Tierra tuvieron que transcurrir 380 millones de años para que una mariposa aprendiera a volar, 180 millones de años más para fabricar una rosa sin otro compromiso que el de ser hermosa y cuatro eras geológicas para que los seres humanos fueran capaces de cantar mejor que los pájaros y morir de amor.

02/100

EL VIEJO SUEÑO DE CREAR INGENIOS A NUESTRA IMAGEN Y SEMEJANZA

Construir máquinas con forma humana y dotarlos de vida propia, conciencia y sentimientos es uno de los sueños más antiguos de la humanidad. En la antigua Grecia, en el siglo octavo antes de cristo, en la *Ilíada*, el poeta Homero relata la construcción de nuevas armas para Aquiles en el taller de Hefesto. El taller venía a ser un santuario de la robótica: trípodes que van y vienen solos sobre pequeñas ruedas de oro; androides metálicos fabricados de oro, que hacen de criados, así como otros robots. Unos diez siglos más tarde, otro poeta, Ovidio, en las *Metamorfosis*, explica que el Rey Pigmalión esculpió la estatua en marfil de una doncella, Galatea, tan bella que se enamoró de ella. Pigmalión pidió a la diosa Afrodita que diera vida a su amada y esta lo complació. En esta prehistoria del sueño de crear ingenios animados, encontramos muchos ingredientes de la relación humana con la IA, en particular su papel de esclavo o dominante y su capacidad para despertar amor u odio. Lo que también explican estos relatos mitológicos es que el poder de crear máquinas, a nuestra imagen y semejanza, está relacionado con el poder divino, como si los humanos siempre hubiéramos querido jugar a ser Dios a base de progreso científico.

El filósofo griego Aristóteles, en el siglo IV antes de Cristo, soñaba con automatizar el razonamiento. Aristóteles identificó un tipo de razonamiento llamado silogismo, que nos permite sacar conclusiones a partir de premisas. El silogismo más conocido es el siguiente: a partir de 1) Todos los humanos son mortales (primera premisa) y 2) Todos los catalanes son humanos (segunda premisa), podemos concluir 3) Todos los catalanes son mortales. La importancia para la IA de la contribución de Aristóteles tiene que ver con

la forma del silogismo: no nos limita a hablar específicamente de seres humanos, catalanes o mortalidad. Podríamos estar hablando de cualquier otra cosa. Esto es obvio si reescribimos el silogismo mediante símbolos arbitrarios. Es decir, podemos escribir: 1) Todos los B son C (en lugar de todos los humanos son mortales); 2) Todos los A son B (en lugar de todos los catalanes son humanos); por tanto, 3) Todos los A son C (todos los catalanes son mortales). En otras palabras, se puede sustituir cualquier cosa que se desee por A, B y C para llegar a una conclusión válida, respetando el significado de las palabras que forman la frase. Por supuesto, la conclusión no será cierta si no lo son las premisas.

Persiguiendo sus propios sueños visionarios, el filósofo, teólogo y fraile mallorquín Ramon Llull en el siglo XIII ideó unos conjuntos de discos concéntricos rotativos con los que pretendía convertir a la fe cristiana a musulmanes y judíos a través de la lógica y la razón. Los discos llevaban inscritas letras del alfabeto que representaban algunos de los atributos de Dios, como bondad, grandeza, eternidad, poder, sabiduría, voluntad, virtud y gloria. La adecuada rotación de los discos se supone que producía respuestas a preguntas teológicas. Este sistema ideado por Llull podía aplicarse a otras cuestiones no teológicas e influyó, cuatro siglos más tarde, en el proyecto del filósofo y matemático Gottfried Wilhelm Leibniz de diseñar un lenguaje universal, basado en números en lugar de letras, con el que se pudiera formular todo el conocimiento humano. Su idea era que si los conceptos eran representados por números, podría obtenerse un razonamiento complejo a partir de los conceptos así representados, multiplicando los números correspondientes a estos conceptos. Leibniz estaba convencido de que de esta manera todas las preguntas podrían reducirse a operaciones matemáticas y que, para resolver cualquier cuestión, solo había que calcular. Este es el significado de la famosa exclamación "Calculemus!", de Leibniz.

Estas propuestas de Aristóteles, Llull y Leibniz abonaron el terreno para una investigación exhaustiva sobre la naturaleza del razonamiento automatizado. Sin embargo, no fue hasta la segunda mitad del siglo XX que, gracias a los ordenadores digitales, se pudieron empezar a poner en práctica estas ideas gracias a la IA.

03/100

ALAN TURING: UN CIENTÍFICO VISIONARIO

Alan Turing fue uno de los científicos más importantes del siglo XX. En 2012, con motivo del centenario de su nacimiento, hubo actos de homenaje en casi todo el mundo y, en particular, en el Reino Unido, su país de origen. Homenajes que nunca tuvo en vida, sino todo lo contrario. Turing, en 1952, fue condenado por homosexual en base a una ley homófoba. Se le dio a elegir entre la cárcel o la castración química. Optó esta última opción, lo que le causó importantes secuelas físicas y psíquicas que, junto al rechazo social por la condena, provocaron su suicidio por envenenamiento al morder una manzana que contenía cianuro potásico.

Turing fue víctima de una sociedad que debía haberle reconocido como un héroe por haber jugado un papel fundamental en el equipo de matemáticos que, durante la Segunda Guerra Mundial, logró descifrar los mensajes cifrados que los mandos del ejército nazi se intercambiaban mediante las máquinas Enigma. Se estima que el descifrado de estos mensajes acortó la guerra en al menos un par de años, evitando cientos de miles de víctimas. Pero la genialidad de Turing no se limitó a sus extraordinarias capacidades para descifrar mensajes. Durante su corta vida hizo contribuciones fundamentales en informática y hoy en día es considerado uno de los padres de esa ciencia. En 1936, mucho antes de que se construyeran los primeros ordenadores, Turing desarrolló los fundamentos teóricos de la computación mediante la introducción de un concepto matemático, ahora conocido como máquina de Turing, sobre el que se basan todos los ordenadores actuales. La máquina de Turing es una rigurosa formalización de conceptos tan básicos en informática como algoritmo y computabilidad y, gracias a esta formalización, podemos determinar dónde están los límites de lo

que es calculable por un ordenador. Demostrar imposibilidades es de extraordinaria importancia en ciencia. Por ejemplo, la imposibilidad de construir máquinas con movimiento perpetuo condujo al descubrimiento de las leyes de la termodinámica en física. Del mismo modo, conocer los límites de las matemáticas y de la computación nos permite saber qué es imposible y, por lo tanto, no es necesario intentar.

Además, Turing es considerado el padre de la IA. En el artículo "Computing machinery and intelligence", publicado en la revista *Mind*, en 1950, argumentaba que en un plazo de unos 50 años habría ordenadores capaces de realizar deducciones lógicas, de aprender adquiriendo nuevos conocimientos, tanto inductivamente como por experiencia, y de comunicar mediante interfaces humanizadas. Era una idea muy radical en ese momento. La argumentación de Turing se basaba en otro importantísimo concepto matemático, el de máquina universal, propuesto también por él. La máquina universal de Turing es capaz de emular cualquier otra máquina, aunque sea más compleja que ella misma. Dado que los seres humanos somos complejas máquinas biomoleculares, pero máquinas al fin y al cabo, podemos pensar, como hizo Turing, que su máquina universal debería poder emular la inteligencia humana.

La última y sorprendente noticia sobre la genialidad de Turing se dio a conocer en 2012 cuando investigadores del King's College de Londres confirmaron experimentalmente una teoría que Turing había formulado más de sesenta años atrás, que explicaba cómo se generan los patrones biológicos que dan lugar, por ejemplo, a las rayas en los tigres o las manchas en los leopardos. El estudio, publicado en la prestigiosa revista *Nature Genetics*, demuestra que estos patrones se deben a la interacción de un par de morfógenos, uno inhibidor y el otro activador, tal y como predecían las ecuaciones que había formulado Turing. Este resultado es de tal magnitud que incluso puede tener aplicaciones importantes en medicina regenerativa. Nos resulta evidente pensar cuántas veces más nos habría sorprendido Alan Turing con contribuciones científicas de primer orden si la intolerancia no se hubiera cruzado en su camino.

04/100
EL JUEGO DE LA IMITACIÓN

En 1950, Alan Turing planteó la cuestión de cómo averiguar si una máquina es o no inteligente. Para responder a esta pregunta propuso una prueba que actualmente lleva su nombre: test de Turing. Esta prueba es una variante del llamado juego de la imitación. En este juego, que era bastante popular en la Inglaterra de principios del siglo pasado, participaban tres personas: un interrogador, un hombre y una mujer. El interrogador, que puede ser hombre o mujer, se sitúa en una sala diferente y se comunica con las otras dos personas mediante mensajes de texto escritos a máquina y dispone de cinco minutos para determinar quién es el hombre y quién es la mujer en base a las respuestas que recibe a sus preguntas. Esto sería fácil si no fuera porque las reglas de este juego permiten que el hombre mienta, con el objetivo de confundir al interrogador, pretendiendo ser la mujer. La mujer, por su parte, intenta, con sus respuestas, ayudar al interrogador a discernir correctamente quién es quién. Si pasados los cinco minutos el interrogador no es capaz de saber con una certeza superior al 70% quién es quién, entonces el hombre gana el juego, ya que ha logrado confundir al interrogador haciéndole creer que era la mujer. Pues bien, el test de Turing para determinar si una máquina es inteligente consiste simplemente en sustituir en este juego el papel del hombre por un ordenador, de tal forma que, si logra confundir al interrogador, haciéndole creer que es un ser humano, diremos que el ordenador es inteligente.

A pesar de algunas noticias aparecidas recientemente en los medios de comunicación, hasta ahora no hay ningún programa de ordenador que haya superado este test. De todas formas, hay que decir que tampoco es realmente un objetivo de los investigadores en IA conseguir superarlo y, por lo tanto, no se han dedicado demasiados esfuerzos a ello. El principal motivo es que este test,

basándonos en el estado actual de la IA, no es un buen indicador para determinar si una máquina es inteligente, ya que, por un lado, es más bien un test sobre la capacidad de engañar que un test de presencia de inteligencia y, por otro lado, como mucho, solo evalúa procesos cognitivos que son susceptibles de ser expresados verbalmente. Hay otros procesos cognitivos fundamentales que no se pueden verbalizar y su modelización y evaluación son imprescindibles en IA. El ejemplo más paradigmático es la actual investigación en robots autónomos con el objetivo de dotarles de sofisticadas habilidades sensoriales y motoras, que permitirán que puedan aprender a reconocer y comprender lo que vean, toquen, escuchen e incluso huelan. También tendrán que tener capacidades de razonamiento espacial para aprender a interpretar su entorno, que generalmente incluirá a otros robots y también a seres humanos, lo que requerirá que desarrollen capacidades de socialización. Para poder medir los progresos hacia estos objetivos, un test como el propuesto por Turing no sirve. Necesitamos un conjunto de pruebas que evalúen todo el abanico de capacidades que conforman la inteligencia y, en particular, la capacidad de adquirir conocimientos de sentido común, que es el problema más importante que debemos resolver para conseguir inteligencias artificiales de propósito general. De hecho, ya hay muchas propuestas de pruebas alternativas. En el artículo "Benchmarks for automated commonsense reasoning: a survey", publicado en octubre de 2023, en la revista *ACM Computing Surveys*, Ernest Davis hace un análisis exhaustivo de doce pruebas cuyo fin es comprobar si los sistemas de IA han adquirido conocimientos de sentido común y capacidad para razonar en base a dichos conocimientos. La conclusión es que dichas pruebas muestran importantes carencias, por lo que muchos aspectos del sentido común siguen sin poder comprobarse. En consecuencia, actualmente no existe un modo fiable de medir hasta qué punto los sistemas de IA existentes han logrado estas capacidades.

05/100

EL SUEÑO DE SESENTA NOCHES DE VERANO

A mediados de los años 50 del siglo pasado, John McCarthy, entonces profesor ayudante en el Dartmouth College en New Hampshire (EE.UU.), organizó un encuentro con un grupo selecto de científicos para hacer un *brainstorming*, sobre la idea de que "cualquier aspecto del aprendizaje o cualquier otro rasgo de la inteligencia humana puede, en principio, ser descrito con un nivel de detalle tal que permite ser simulado en una máquina". McCarthy convenció a Claude Shannon, el padre de la teoría matemática de la información, de los laboratorios Bell, y a Marvin Minsky, entonces en Harvard, para, entre los tres, redactar una propuesta sobre la base de esta idea. La titularon *Summer research project in artificial intelligence* y solicitaron financiación a la Fundación Rockefeller. Fue financiada y el encuentro, que duró unas ocho semanas, tuvo lugar en el verano de 1956 en Dartmouth. En la propuesta se afirmaba que, en tan solo unos dos meses, un grupo de científicos cuidadosamente seleccionados podría conseguir avances significativos en aspectos como la comprensión del lenguaje, la abstracción de conceptos mediante aprendizaje y la resolución de problemas que hasta entonces solo habían sido resueltos por seres humanos.

Además de los tres proponentes, también estuvieron en Dartmouth los siguientes investigadores: Nathaniel Rochester, Arthur Samuel y Alex Bernstein, de IBM; Oliver Selfridge y Ray Solomonoff, del Massachusetts Institute of Technology; Allen Newell, de Rand Corporation, y Herbert Simon, del Carnegie Institute of Technology (actualmente Universidad Carnegie Mellon). Rochester estaba interesado en la aproximación conexionista a la IA para modelizar matemáticamente el funcionamiento de las redes neuronales; Samuel había diseñado un programa para jugar a *checkers* —un

juego muy similar a las damas— que jugando contra una copia de sí mismo era capaz de aprender a mejorar su juego mediante la estimación de una función matemática que evaluaba la calidad de las jugadas. Bernstein también estaba interesado en los juegos y había trabajado en un programa para jugar al ajedrez. Selfridge estaba interesado en resolver el problema del reconocimiento de patrones y Ray Solomonoff trabajaba en una teoría general de la inferencia y sus posibles implicaciones para modelizar inteligencia artificial general. Newell y Simon llegaron a Dartmouth con algo más tangible, un programa de ordenador, llamado Logic Theorist, capaz de demostrar 38 de los 52 teoremas incluidos en el libro *Principia Mathematica*, de Alfred North Whitehead y Bertrand Russell. John McCarthy, además de proponer, con éxito, que el nuevo campo de estudio se llamara inteligencia artificial, estaba interesado en diseñar un lenguaje, interpretable por un ordenador, con el que programar aspectos como la recursividad para hacer cálculos con expresiones simbólicas. Esta idea dio lugar pocos años después al lenguaje de programación LISP. Shannon estaba interesado en la aplicabilidad de su teoría matemática de la información para modelizar el funcionamiento del cerebro, pero después de la reunión de Dartmouth dejó de interesarse por la IA. Por último, Minsky planteó la posibilidad de una máquina, nunca construida, capaz de generar un modelo abstracto del mundo, de forma que, a la hora de resolver cualquier problema, primero intentara encontrar la solución usando dicho modelo abstracto interno y si esto no daba resultados intentara solucionarlo planificando experimentos interactuando con el mundo.

El tiempo ha demostrado que estos pioneros fueron exageradamente optimistas, ya que de hecho fueron necesarias varias décadas para poder hablar efectivamente de progresos significativos en los temas que se discutieron en Dartmouth. Uno de los errores de estos pioneros de la IA fue su excesivo optimismo, consecuencia de subestimar la enorme complejidad del problema de modelizar procesos cognitivos. De hecho, en 2006, durante la celebración, también en Dartmouth, del 50 aniversario de la famosa reunión, los cuatro supervivientes del encuentro de 1956, McCarthy, Minsky, Selfridge y Solomonoff, reconocieron que la IA es un objetivo mucho más difícil de lo que nunca podrían haber llegado a imaginar.

06/100

IA GENERAL Y LA HIPÓTESIS DEL SISTEMA FÍSICO DE SÍMBOLOS

En una ponencia, en 1975, en ocasión de la recepción del prestigioso Premio Turing, Allen Newell y Herbert Simon formularon una hipótesis para la IA: la hipótesis del sistema físico de símbolos (SFS). Según esta hipótesis, todo sistema capaz de procesar símbolos posee los medios necesarios y suficientes para ser inteligente en el sentido general del término. Aunque estrictamente la hipótesis SFS se formuló en 1975, de hecho, ya estaba implícita en las ideas de los pioneros de la IA en los años 50 e incluso en las ideas de Alan Turing en sus escritos sobre máquinas inteligentes de finales de los años 40.

Por otro lado, dado que los seres humanos somos inteligentes en el sentido general, entonces, de acuerdo con la hipótesis, nosotros somos también sistemas físicos de símbolos. Conviene aclarar a qué se refieren Newell y Simon al hablar de sistema físico de símbolos. Un SFS consiste en un conjunto de entidades llamadas símbolos que, mediante relaciones, pueden ser combinados formando estructuras más grandes —como los átomos que se combinan formando moléculas— y pueden ser transformados aplicando un conjunto de procesos. Estos procesos pueden introducir nuevos símbolos, crear y modificar relaciones entre símbolos, almacenar símbolos, comparar si dos símbolos son iguales, etcétera. Estos símbolos son físicos en tanto que tienen un sustrato electrónico (en el caso de los ordenadores) o biológico (en el caso de los seres humanos). Efectivamente, en el caso de los ordenadores, los símbolos se realizan mediante circuitos electrónicos digitales y, en el caso de los seres humanos, mediante redes de células nerviosas

(neuronas). En definitiva, de acuerdo con la hipótesis SFS, la naturaleza del sustrato (circuitos electrónicos o redes de neuronas) no tiene importancia siempre que este sustrato permita procesar símbolos. No olvidemos que se trata de una hipótesis y, por tanto, no debería ser ni aceptada ni rechazada a priori. Su validez o refutación deberá verificarse, de acuerdo con el método científico, con ensayos experimentales. La IA es precisamente el campo científico dedicado a intentar verificar esta hipótesis en el contexto de los ordenadores digitales, es decir, verificar si un ordenador convenientemente programado es capaz o no de tener inteligencia general.

Es importante matizar que debería tratarse de inteligencia general y no especializada, ya que la inteligencia de los seres humanos es general. Exhibir inteligencia específica es algo bien diferente. Por ejemplo, los programas que juegan al ajedrez a nivel de Gran Maestro son incapaces, sin un rediseño y reentrenamiento extensivo, de jugar a las damas a pesar de ser un juego mucho más sencillo. En el caso de los seres humanos no es así puesto que cualquier jugador de ajedrez puede aprovechar sus conocimientos sobre este juego para, en cuestión de pocos minutos, jugar a las damas perfectamente. El diseño y realización de inteligencias artificiales que únicamente muestran comportamiento inteligente en un ámbito muy especializado está relacionado con lo que se conoce por IA débil en contraposición con la IA fuerte, a la que, de hecho, se referían Newell y Simon y otros padres fundadores de la IA. La IA fuerte guarda relación con la IA general, aunque, contrariamente a lo que muchos creen, no son lo mismo. Una IA fuerte, según la definió el filósofo John Searle, no simula una mente, sino que es una mente y, por tanto, será necesariamente general. Pero puede haber IA general, es decir, capaz de llevar a cabo una amplia gama de tareas muy distintas, sin ser una mente. En definitiva, IA fuerte implica IA general pero no al revés.