Fundamentos filosóficos de la Inteligencia Artificial

Luis Alonso Romero
Catedrático de Ciencia de la Computación e Inteligencia Artificial
Universidad de Salamanca

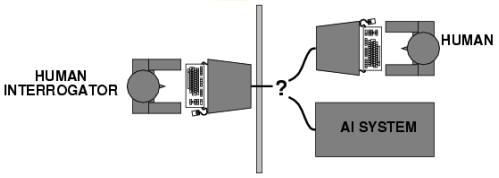
Objetivos

- Desde que surgió el concepto de Inteligencia Artificial (Darmouth 1956), se han ido planteando algunas preguntas:
  - ¿Pueden pensar las máquinas?
  - ¿Cómo funciona la mente?
  - ¿Qué implicaciones éticas tendrá la Inteligencia Artificial?
  - .....
- La Inteligencia Artificial se ha planteado con dos enfoques:
  - **Débil** : es posible construir máquinas que se comporten como si fuesen inteligentes ("conductista")
  - **Fuerte**: es posible construir máquinas inteligentes.

IA débil: simulación de inteligencia

- "La IA, perseguida como una especie de culto a la computación ...no tiene la más mínima posibilidad de producir resultados duraderos y, por tanto, es hora de dedicar los esfuerzos, y el dinero, a otras tareas" (Sayre 1993).
- Obviamente, todo depende de la definición de IA.
- Desde un punto de vista ingenieril la respuesta es SI : las máquinas pueden simular la inteligencia.
- Pero los filósofos plantean la pregunta en otros términos:
  - ¿Pueden pensar las máquinas?
- La pregunta está mal planteada porque no tenemos una idea clara de lo que significa "pensar".
  - ¿Pueden volar las máquinas?
  - ¿Y nadar?

Test de Turing (1950)



- Turing responde a la pregunta planteando una prueba de inteligencia
- Turing afirmó que en el 2000 un ordenador tendría un 30% de probabilidades de engañar a una persona durante 5 minutos.
- El test de Turing no ha recibido mucha atención del mundo de la IA.
- Pero Turing también planteó las objeciones a la posibilidad de una máquina inteligente.

Argumento de incapacidad

- Una máquina nunca podrá :
  - Ser amable, ser hermosa, tener sentido del humor, equivocarse, enamorarse, disfrutar de las fresas con nata, aprender de la experiencia, ... (Turing 1950).
- Mirando hacia atrás (2006), es innegable que los ordenadores hacen cosas que antes eran privativas de los humanos:
  - Jugar al ajedrez, a las damas, vigilar líneas de producción, pilotar aviones, diagnosticar, ...
  - Han hecho descubrimientos significativos en astronomía, matemáticas, química, biología, ciencia de la computación, etc. Estos descubrimientos habrían sido imposibles sin el computador.
- Está claro que los ordenadores pueden hacer muchas cosas igual, o mejor, que los humanos.
  - Lo que no implica que usen comprensión o intuición.
  - Puramente conductista.

Argumento matemático

- El teorema de incompletitud de Godel (1931) estableció que hay cuestiones matemáticas que no pueden resolverse con un sistema axiomático formal.
- Lucas (1961) , Penrose (1994),.. establecieron que esto demuestra que las máquinas son mentalmente inferiores a los humanos:
  - Los ordenadores son sistemas axiomáticos formales y los humanos no.
- Contraejemplos:
  - 1.- “Manuel no puede asegurar que esta sentencia es cierta”
    - ¿Significa esto que Manuel no es inteligente?
  - 2.- Esta máquina puede sumar mucho más rápidamente que cualquier persona.
    - ¿Significa que la máquina es más inteligente?
  - 3.- Las máquinas tienen limitaciones... Y nosotros también. ¿O es que somos absolutamente consistentes en nuestros razonamientos?

Argumento de la informalidad

- Problema de cualificación:
  1. El comportamiento humano es demasiado complicado como para ser capturado por un sistema simple de reglas.
  2. Los ordenadores solamente pueden seguir un sistema simple de reglas.
  3. Conclusión: nunca podrán generar un comportamiento tan inteligente como el humano.
  - Hubert y Stuart Dreyfus (1972, 1986, 1992)
- Esta objeción puede ser cierta en sistemas de IA basados en lógica de primer orden, pero no en sistemas más complejos (probabilistas, por ejemplo).
- Dreyfus no explica cómo razonan los humanos:
  - Dennet (1994) : "Es como si los filósofos se proclamaran expertos en los métodos de los prestidigitadores. Si se les pregunta algo dicen "está claro que el mago no ha serrado a la chica, pero cómo lo ha hecho no es problema mío".

Argumento de la informalidad II

- Los hermanos Dreyfus proponen un proceso en cinco etapas para adquirir conocimiento (1986). Todas sus propuestas se han incorporado en la IA actual (redes neuronales, aprendizaje, incertidumbre, ...)
- Por lo tanto, el argumento de la informalidad mostró las limitaciones de la IA de la época, no la imposibilidad.

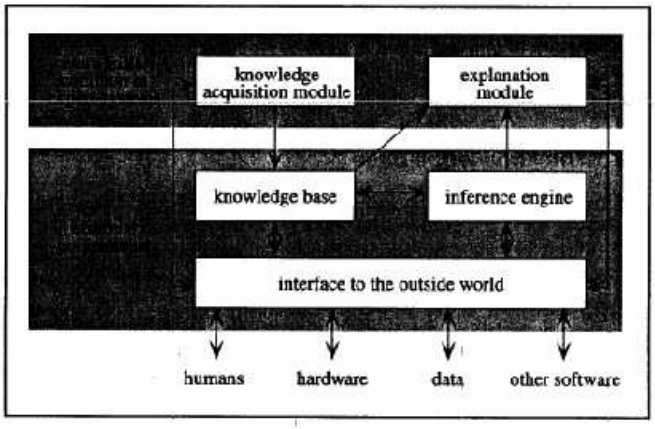
Algunos objetivos de la IA

- A pesar de todo lo anterior, la IA ha conseguido una serie de objetivos que podemos clasificar en:
- Razonamiento:
  - Dados {conocimiento, hechos}, deducir consecuencias. Por ejemplo, dado un cierto conocimiento sobre enfermedades, y una relación de síntomas (hechos) efectuar un diagnóstico.
- Planificación:
  - Dados {conocimiento, situación actual, objetivo deseado}, deducir la secuencia de acciones para alcanzar el objetivo. (*razonamiento dirigido por objetivos*).
- Aprendizaje:
  - Dados {conocimiento, hechos}, deducir de nuevos hechos posibles modificaciones sobre el conocimiento.
- En aplicaciones concretas (reconocimiento de habla, visión artificial, robótica) pueden coexistir varios de los objetivos anteriores.

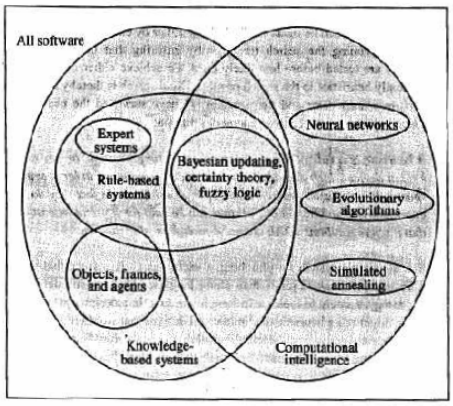
Desarrollos en IA

- Los desarrollos producidos en la Inteligencia Artificial en todos estos años pueden dividirse, grosso modo, en tres categorías:
  - Sistemas basados en conocimiento:
    - Sistemas expertos, sistemas basados en reglas y sistemas basados en marcos y objetos.
  - Sistemas de Inteligencia computacional:
    - Redes neuronales, algoritmos genéticos.
  - Sistemas híbridos:
    - Sistemas basados en lógica borrosa.
- Las dos últimas categorías constituyen lo que se ha dado en llamar computación “soft”.

Sistemas expertos



Software de Sistemas Inteligentes.



IA fuerte: ¿pueden pensar las máquinas?

- Jefferson (1949):
  - "Hasta que una máquina no pueda escribir un soneto... Basado en sus conocimientos y emociones... No podremos estar de acuerdo en que las máquinas sean iguales al cerebro"
- Turing llamó a esto el argumento de la consciencia (conocimiento de lo que pasa en la propia mente).
  - Jefferson realmente planteaba el problema *fenomenológico*, o estudio de la experiencia directa (sentir las emociones).
  - Otros autores se centran en la *intencionalidad* de la máquina en la expresión de sus deseos o sentimientos.
- Turing, en cierto modo, despreció bastante este argumento:
  - "¿Por qué tenemos que exigir más a las máquinas que a nosotros? En la vida ordinaria, no tenemos ninguna evidencia de los estados mentales internos de las demás personas...En lugar de discutir sobre esto solemos tener la convención educada de que todo el mundo piensa."

IA fuerte

- Más de Turing:
  - "No quiero dar la impresión de que pienso que no hay ningún misterio con la consciencia... Pero no creo que esto tenga que resolverse antes de que podamos contestar a la pregunta sobre la actuación inteligente de las máquinas".
- Un ejemplo de otro campo: En 1848, F. Wöhler sintetizó urea por primera vez, empleando técnicas de Química Inorgánica y Orgánica, posibilidad que había sido negada hasta entonces. A partir de ahí, la urea natural y la artificial no son distinguibles...
- Un ordenador puede "jugar" al ajedrez, o puede "sumar" dos números. ¿Por qué no puede "pensar"?

Funcionalismo vs. naturalismo

- Funcionalismo : cualquier condición causal intermedia entre una entrada y una salida es un estado mental.
  - Bajo esta teoría, dos sistemas cualesquiera con estados causales isomórficos tienen los mismos estados mentales.
  - Es decir, un programa que "juegue" al ajedrez pasa por los mismos estados mentales que un jugador humano.
- Naturalismo biológico: los estados mentales son características emergentes de alto nivel causadas por procesos neurológicos de bajo nivel en las neuronas, cuyas propiedades son las que importan.
  - Es decir, un programa no solo tiene que tener la misma entrada-salida que el humano sino que, además, debería ejecutarse en una arquitectura con la misma potencia de causalidad que las neuronas.

Problema mente-cuerpo

- ¿Cómo se relacionan los estados mentales con los estados y procesos corporales (o cerebrales)?
  - Este es uno de los problemas más viejos de la filosofía de la mente.
- Teoría dualista (Descartes, "Discurso del método" 1637) : el alma y el cuerpo son dos cosas distintas.
- Teoría monista (materialismo) : Los estados mentales son estados cerebrales. "Los cerebros producen las mentes" (Searle, 1984)
  - Obstáculos:
    - Libre albedrío : si las mentes están gobernadas por leyes de la física, ¿dónde queda el libre albedrío).
    - Consciencia: ¿por qué algunos sistemas físicos son conscientes de que lo son y otros no? (las piedras, por ejemplo)
- ¿Qué es el estado de un cerebro?

“Cerebro en una cubeta”

- Este es un experimento (virtual) muy clásico:
  - Imaginemos que al nacer un niño se le extrae el cerebro que se mantiene vivo, y creciendo, en una cubeta.
  - Este cerebro se conecta a un simulador del mundo a través de las interfaces entrada/salida apropiadas. (Película Matrix 1999)
- Los estados mentales de este cerebro, ¿serían los mismos si estuviese alojado en su cráneo y creciendo de forma natural?
  - Para un observador externo la respuesta es claramente NO. (*contenido extenso*)
  - Desde un punto de vista subjetivo interno, la respuesta es SI (*contenido estrecho*)
- “Qualia” o experiencias intrínsecas:
  - ¿El color rojo para mí es el mismo que para ti? ¿Cómo saber si lo que yo veo rojo tú lo ves verde, y viceversa?
- Este experimento plantea una serie de interrogantes muy presentes en la literatura de ficción.

Prótesis cerebral

- Creado por Glymour (1970), Searle (1980) y Moravec (1988)
- Imaginemos que:
  - Conocemos perfectamente el funcionamiento del cerebro humano.
  - Podemos construir circuitos electrónicos que emulen perfectamente el funcionamiento de partes del cerebro.
  - Podemos sustituir parte de un cerebro por circuitos, sin interrumpir el funcionamiento del cerebro.
- Experimento : sustituir gradualmente todas las neuronas del cerebro por circuitos y, un tiempo después, invertir el proceso.
- ¿Qué pasaría con el sujeto objeto del experimento?
  - Moravec : su consciencia no se vería afectada.
  - Searle: la consciencia desaparece pero el comportamiento externo no se altera.
- ????

La habitación china(Searle 1980)

- Una habitación cerrada con una ventanilla “entrada” y otra “salida”
- Dentro de la habitación hay una persona que solo entiende inglés, con un libro de reglas, un lápiz y un montón de hojas de papel.
  - Persona : CPU,
  - Libro de reglas : programa
  - Papeles : memoria de almacenamiento.
- Por la ventanilla de entrada se introducen hojas con textos escritos en chino.
- La persona consulta su libro de reglas, hace sus anotaciones en los papeles y al cabo de un rato, por la ventanilla de salida saca hojas con el texto correspondiente en inglés.
  - La persona no entiende chino: “la ejecución del programa no genera comprensión de lo que se está haciendo (Searle)”

Más sobre la habitación china

- Mc Carthy : aunque la persona no entiende el chino, el sistema (habitación, libro, papeles) se comporta como si lo entendiese.
- Aplicando la convención educada de Turing, esto sería suficiente.
- Searle: la comprensión no está en el hombre, y no puede estar en el libro o los papeles ….por tanto no hay comprensión.
- ¿Podemos decir que la habitación china es una mente?

NEST (6 programa Marco): New and Emerging Science and Technology

Human Mind Project: Objetivos

§ The genetics of human cognition – how our species, despite its remarkable genetic resemblance to other apes, has evolved and sustains such an extraordinary complex mind.
§ The developing mind – how various life experiences influence the development, maturation and aging of a normal human brain.
§ The process of thinking – new insights into reasoning, learning and memory that will impact on education, communication and the development of intelligent technologies.
§ Motivation and decision making – insights into what motivates people to cooperate or, conversely, to behave with disregard for others, and what factors influence us to make the choices we do.
§ Cultural context – how the human mind manifests itself as culture, how cultures change and how they endure. Which behaviours and ways of thinking are part of our culture and which are part of our nature.

La ética y la I.A.

- Todo avance científico y tecnológico plantea dilemas morales : ¿se debería hacer?
- La IA no es ajena a este planteamiento. Las principales objeciones que se le plantean son:
  1. Las personas podrían perder su trabajo por culpa de los ordenadores.
  2. Las personas podríamos tener demasiado ocio (o ninguno).
  3. Las personas podrían perder la sensación de ser únicos.
  4. Las personas podrían ver alterados sus derechos.
  5. El uso de sistemas de IA podría cambiar la noción de responsabilidad.
  6. El éxito de la IA podría suponer la extinción de la especie humana.

Objeciones

- 1.- Laboral: Hasta ahora, la informatización ha creado bastantes más puestos de trabajo que los que ha destruido. Como todo avance científico tecnológico, lo que se produce es un cambio del tipo de trabajo y una creación de nuevos trabajos generalmente más interesantes y mejor pagados.
- 2.- Ocio : la experiencia demuestra que aunque la jornada laboral media tiende a disminuir, los sistemas informatizados exigen una atención continuada: o las jornadas se alargan, o se crean más puestos.
  - El principio neoliberal “el ganador se queda con todo” hace que la competitividad exagerada produzca una fuerte presión en el trabajo. Una empresa un 10% mejor que la competencia se puede quedar con el 100% del mercado.

Más objeciones éticas

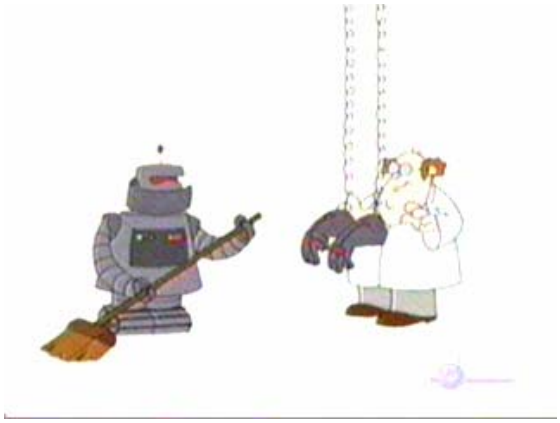
- 3.- Unicidad del ser humano: Esta idea es muy antigua (L'Homme Machine, La Metrie 1748). Es un pensamiento que puede tener algo de razón, pero no está claro si en sentido positivo o negativo.
- 4.- Derechos: el argumento está fundado en la experiencia. Pero no depende de la IA, sino de los políticos que reducen, o eliminan, los derechos privados valiéndose de cualquier herramienta: escuchas, espías informáticos, Guantánamo (nada que ver con la IA)
- 5.- Responsabilidad : si un sistema experto hace (o sugiere) un mal diagnóstico, ¿quién es responsable? (Gawande 2002)
  - Si tenemos un agente inteligente que hace inversiones en Bolsa, ¿quién responde de posibles deudas?
  - Los legisladores van a remolque de la tecnología, en general. O legislan de forma muy genérica, sin entender realmente qué es lo que tratan de regular.

- 6.- Fin de la especie humana: este es un argumento muy usado en las novelas de ciencia ficción (y películas) : Frankestein, Terminator (1964), Matrix (1999)
  - Los robots inteligentes, ¿son más amenazadores que los fantasmas o las brujas?
  - Las máquinas que se construyan no tiene que ser agresivas...a menos que se desee hacerlas así.
- Un escenario más amenazante es aquel en el que los ordenadores se vuelvan absolutamente imprescindibles para la vida humana...y sean conscientes de ello.
- Kurweil (2000) : "Hacia el 2099 ya no existirá una distinción clara entre los hombres y los computadores" (transhumanismo)
- Punto de vista de las máquinas: si adquieren consciencia, ¿será ético seguir tratándola como máquinas?

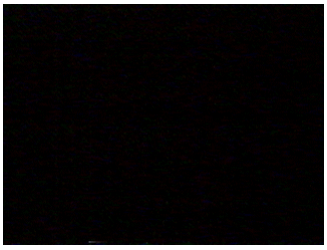
Singularidad Tecnológica: ¿Ciencia ficción?

- Vernor Vinge, matemático (1993): " Dentro de 30 años tendremos los medios tecnológicos para crear una inteligencia superhumana...Algún tiempo después, la era humana habrá terminado"
  - Vinge se basa en la ley de Moore que parece que está alcanzando la fase de saturación...¿O no?
- Moravec (2000) en *Robot, Mere Machine to Trascendent Mind*: "De manera bastante rápida, podríamos quedar desplazados y fuera de la existencia...Al igual que los hijos biológicos de generaciones anteriores, las máquinas representan la mejor esperanza de la humanidad para un futuro a largo plazo. Nos corresponde a nosotros ofrecerles todas las ventajas y cómo retirarnos cuando ya no podamos contribuir"

Robótica "Inteligente"



Inteligencia Animal

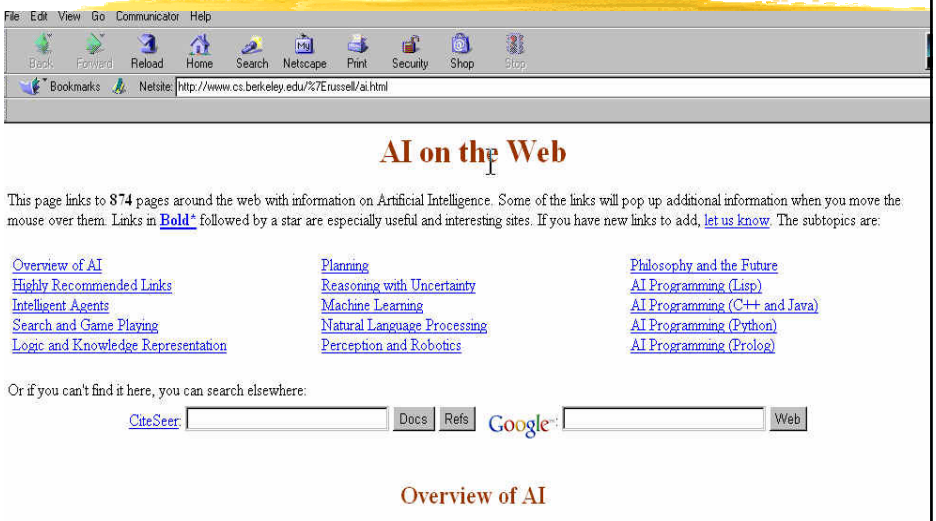


Inteligencia de insectos: R. Brooks

Algunas cifras para pensar...

	Computador	Cerebro humano
Unidades	1 CPU, 10^9 puertas	10^{11} neuronas
Almacenamiento	10^{11} Bits RAM 10^{13} Bits disco	10^{11} neuronas 10^{14} sinapsis
Tiempo de ciclo	10^{-9} sg	10^{-3} sg
Ancho de banda	10^{10} bits/sg	10^{14} bits/sg
Actualizaciones de memoria	10^9 /sg	10^{14} /sg

Algunas páginas Web



Algo de bibliografía

- Russell, Norvig.- *Inteligencia Artificial, Un Enfoque Moderno.*- Prentice Hall, 97
- Bender.- *Mathematical Methods in Artificial Intelligence*.- IEEE Press, 96
- Haykin: *Neural Networks*, IEEE Press, 1994
- Nillson.- *Principios de Inteligencia Artificial*.- Diaz de Santos, 94
- Aranda, et al.- *Fundamentos de Lógica Matemática*. Sanz y Torres 03
- Bratko.- *PROLOG programming for Artificial Intelligence.* Addison-Wesley 2001
- Hopgood .- *Intelligent Systems for Engineers and Scientists.*- CRC Press, 2001
- Minsky.- *The Emotion Machine*.- Simon&Schuster 2006