Problemas éticos y morales de la Inteligencia Artificial. Evolución a lo largo de la historia

Historia de las matemáticas - Universidad de Granada

Antonio Martín Ruiz
Laura Gómez Garrido
Fernando de la Hoz Moreno

12 de diciembre de 2019

Contenido

Introducción. ¿Qué es la IA?

Cuatro visiones:

- Sistemas que actúan como humanos
- Sistemas que piensan como humanos
- Sistemas que piensan racionalmente
- Sistemas que actúan racionalmente

Historia de la IA

- Génesis de la Inteligencia Artificial (1943-1955)
- Nacimiento de la Inteligencia Artificial (1956)
- Entusiasmo inicial, grandes esperanzas (1952-1969)
- Una dosis de realidad (1966-1973)

Historia de la IA

- Sistemas basados en conocimiento (1969-1979)
- La IA se convierte en una industria (desde 1980 hasta el presente)
- Regreso de las redes neuronales (desde 1986 hasta el presente)
- La IA se convierte en una ciencia (desde 1987 hasta el presente)

Objeciones de Turing

Objeción Teológica:

> *El pensamiento es una función del alma inmortal del hombre. Dios ha proporcionado un alma inmortal a todos los hombres y mujeres, pero no así a ningún otro animal, ni tampoco a las máquinas. Por consiguiente, ningún animal o máquina puede pensar.*

El argumento de la percepción extrasensorial.

Si estas objeciones fueran ciertas, ya no podríamos considerar que nuestros cuerpos se mueven de acuerdo a las leyes físicas conocidas ni por las que aún están por descubrir.

Objeciones de Turing

La objeción de la cabeza en la arena:

> *Las consecuencias de que las máquinas pensaran serían demasiado terribles. Esperemos y creamos que no pueden hacerlo.*

Argumento de la conciencia: Se basa en que no podemos considerar que una máquina iguala al cerebro humano hasta que no sea realmente consciente de lo que ella misma esta creando y pueda tener sentimientos y emociones.

Objeciones de Turing

La objeción matemática:

Existen muchos resultados de lógica matemática que pueden utilizarse para demostrar que hay limitaciones al potencial de las máquinas de estado discreto. [...] Este es el resultado matemático: se afirma que prueba que las máquinas adolecen de una incapacidad a la que no se encuentra sujeto el intelecto humano.

Objeciones de Turing

Argumentos sobre diversas incapacidades:
Son de la forma: "*Acepto que puedas hacer que las máquinas hagan todo lo que hasta ahora has mencionado, pero nunca podrás hacer que una de ellas haga X*"

La objeción de Lady Lovelace

> *La máquina no pretende crear nada. Puede hacer lo que sea que sepamos ordenarle. - Ada Lovelace, 1842.*

Objeciones de Turing

El argumento de la continuidad del sistema nervioso

El sistema nervioso no es una máquina de estado discreto y, por ello, no deberíamos de poder ser capaces de imitar su comportamiento de forma discreta.

Objeciones de Turing

El argumento de la informalidad del comportamiento

Este dice que es inviable establecer un conjunto de reglas de conducta que los seres humanos puedan cumplir para cualquier eventualiad. Además,a través de estas reglas de comportamiento que poseen las máquinas, deberíamos de poder ser capaces de predecir en un tiempo razonable todas sus acciones.

Singularidad tecnológica

- La singularidad tecnológica. Crecimiento tecnológico incontrolable e irreversible.
- Cerebro VS Inteligencia Artificial.
- Explosión de la inteligencia. Capacidad de automejora que permitetirá superar la inteligencia humana.

Peligros no intencionados. Seguridad en IA

En la IA, pueden surgir situaciones en la que su comportamiento sea inesperado.

La función objetivo indica a la IA si esta realizando la tarea que tiene encomendada.

Problema cuando la función objetivo no esta bien especificada. Puede no tener en cuenta el resto de elementos del medio y provocar efectos dañinos.

Peligros intencionados. Uso malicioso de la IA

Características que hacen peligrosa a la IA:

- Posible utilización destructiva
- Eficientes y escalables
- Aumentan anonimato y distancia psicológica
- Rápida difusión
- Vulnerabilidades

Seguridad digital

Los sistemas distrubuidos favorecen la automatización de ataques.

En la actualidad solo se conocen ataques llevados a cabo por sombreros blancos.

Actual utilización de sistemas de IA en ciberseguridad.

Algunos casos concretos:

- Automatización de ataques de ingeniería social.
- Automatización del descubrimiento de vulnerabilidades.
- Ataques DDoS imitando el comportamiento humano.

Seguridad física

Aplicaciones armamentísticas.

Aumento de la diferencia entre potencial ofensivo y defensivo.

- Reutilización terrorista de sistemas comerciales.
- Aumento de la capacidad ofensiva.
- Aumento de la escala de los ataques.

Seguridad política

Avance tecnológico e inestabilidad política están ligados.

Producción y detección de información manipulativa.

Control de la población.

Ataques contra la seguridad política:

- Plataformas de vigilancia automatizada.
- Informes de noticias falsa.
- Campañas de desinformación.

Consecuencias de la IA sobre el empleo

Los robots, la automatización y la inteligencia artificial están ya transformando el mundo del empleo.

Se destruirá empleo de baja cualificación, mientras que se crearán nuevos puestos de trabajo más especializados.

Problema con la reinserción laboral de mucha gente. Una posible solución puede ser la renta universal.

Para ampliar

Computing Machinery and Intelligence Alan Turing (1950)

Artifficial Intelligence: A Modern Aproach Stuart J. Russell y Peter Norvig

Concrete Problems in AI Safety Dario Amodei, Crhis Olah, Jacob Steinahrdt, Paul Christiano, John Schulman, Dan Mané

The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation Miles Brundage, Shahar Avin et al.