El rol de la IA en la sociedad

Sofía Trejo

El objetivo principal de este curso es estudiar las problemáticas éticas y sociales derivadas de la Inteligencia Artificial.

Inteligencia Artificial

IA:

Máquinas que presentan un comportamiento inteligente y les es fácil superar al humano en tareas específicas. Por ejemplo, el buscador de Google. Estos sistemas desarrollan tareas denominadas académicamente como inteligentes. Por ejemplo, traducción, recomendaciones, reconocimiento de voz e imagen.

IAGeneral:

Sistemas con capacidades cognitivas (como aprendizaje y resolución de problemas) equivalentes o superiores al humano.

A grandes rasgos hay dos tipos de IA.

Analítica: inteligencia cognitiva.

Inspirada en humanos: que emule sentimientos y empatía.

IA en la actualidad

- Contrataciónes (escaneo de CVs).
- Préstamos e Hipotécas.
- Sentencias de carcel (asesoría a jueces).
- Patrullaje (policial).
- Transporte (self-driving cars).

Las IA realizan tareas cognitivas antes realizadas por humanos con una dimensión social, por ende las IA heredan requerimientos sociales.

Ética

Sistema de principios morales que influencian la toma de decisiones y la forma en que llevamos nuestras vidas.

En particular, la ética de IA se puede dividir en dos ramas:

- **Ética de máquinas** (machine ethics): se refiere al comportamiento moral de las máquinas
- **Roboética** (roboethics): estudia el comportamiento moral de los humanos que diseñan los sistemas, cómo se implementan y cómo se tratan a dichos sistemas.

Ejemplo de problemática

Deep Blue: IA construida por IBM para jugar ajedrez.

- **1985:** Kasparov gana vs 15 sistemas.
- **1994:** Kasparov pierde una partida vs Chess genius (no el juego).
- **1996**: Deep Blue pierde 4-2 vs. Kasparov, el campeón mundial de ajedrez.
- **1997:** Deep Blue gana 3 ½ vs 4 ½ vs. Kasparov.

Kasparov acusa a IBM de hacer trampa y pide un rematch. IBM se niega y desmantela el sistema.

¿Cómo ganaron?

Dado el estado actual del tablero, el sistema explora en el grafo de posibles movimientos. Decide su movimiento analizando las probabilidades de ganar determinadas por caminos en el grafo.
El sistema aprendió las probabilidades jugando y fue ajustando los valores.

¿Por qué hacer esto?

El número de posibles partidas de ajedrez es $10^{100,\,000}$, que es mayor al número de átomos en el universo.

- No se pueden programar todas las partidas de ajedrez.
- Al dar una programación específica la computadora jugaría tan bien como el programador, no mejor que el campeón mundial de ajedrez.

Tecnologías como Deep Blue extrapolan las consecuencias futuras de sus acciones, pero no tienen especificaciones locales de comportamiento.

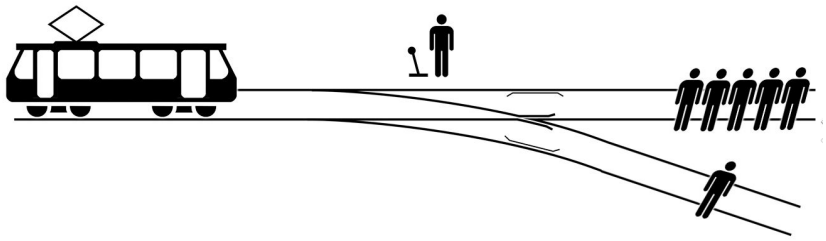
Vehículos Autónomos

Si un vehículo autónomo incurre en un accidente ¿quién es el responsable?

- ¿El dueño del vehículo?
- ¿La compañía que produjo el vehículo?
- ¿El programador?

Dilema del Tranvía (Trolley problem)

Problema clásico de psicología moral.



- Clásico: la mayoría de las personas salvaría a las cinco personas.
- Empujar a al gordo: las personas no lo empujarian para salvar a las cinco personas.
- Gordo villano: parece ser un imperativo moral empujarlo.
- Si la persona sola es un pariente: la gente prefiere no mover la palanca.

The Moral Machine Experiment

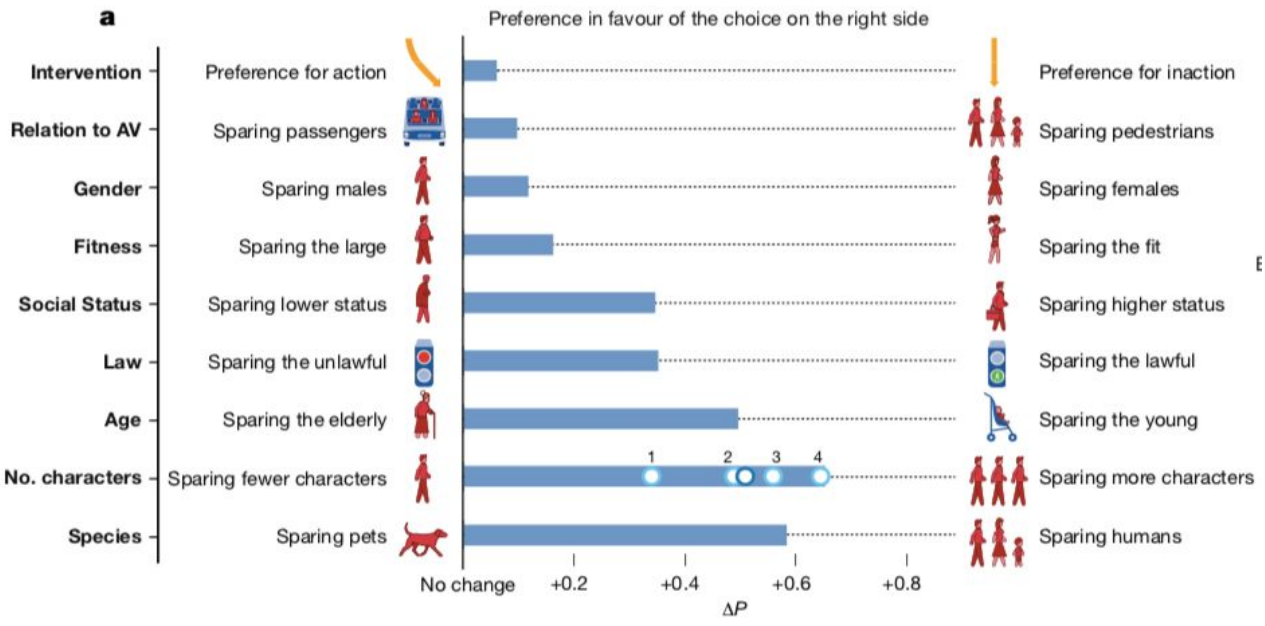
Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-françois Bonnefon & Iyad Rahwan.

Para ayudar con los dilemas éticos asociados con vehículos autónomos MIT construyó un simulador virtual para estudiar las preferencias morales de las personas.

- 40 millones de decisiones registradas provenientes de 233 países.
- Parámetros a considerar:
    - Humano vs animales.
    - Desviarse vs continuar trayectoria.
    - Pasajeros vs peatones.
    - Individuo vs colectividad.
    - Hombre vs Mujer.
    - Distintas edades.
    - Cruce legal de peatones vs cruce ilegal.
    - Estado físico (gordo vs en forma).
    - Estatus social (bajo vs alto).

Preferencias globales

Cada renglón representa la diferencia entre la probabilidad de salvar a los personajes de la derecha menos la probabilidad de salvar a los de la izquierda sobre la distribución de los otros atributos.



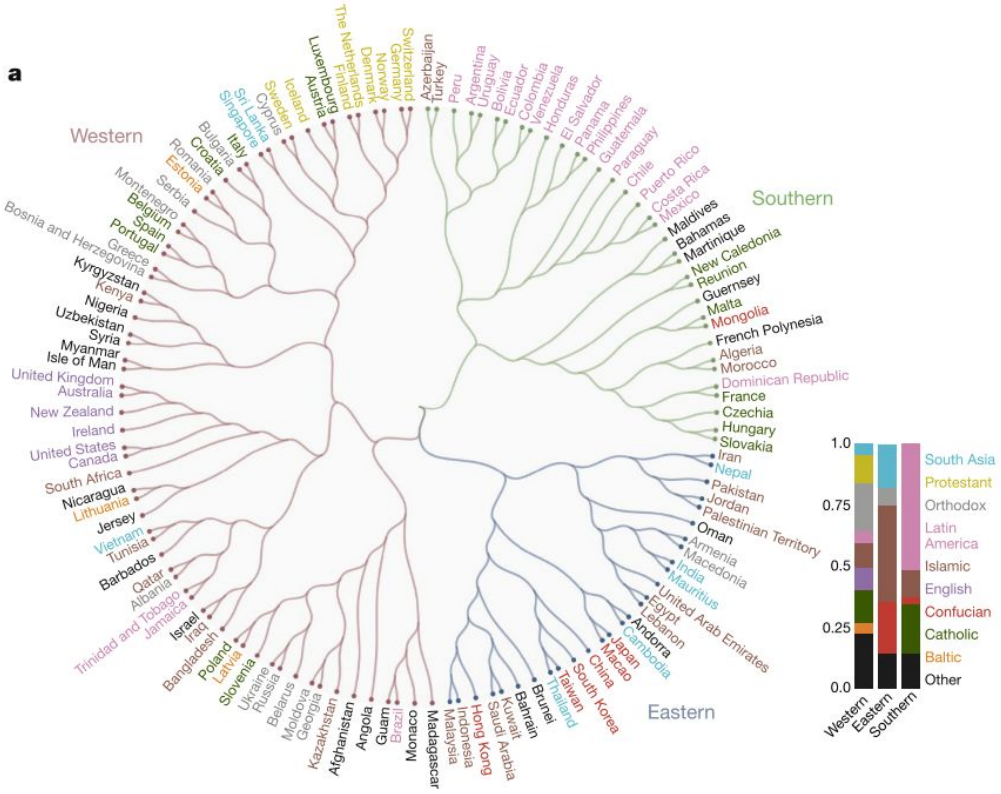
German Ethics Commission on Automated and Connected Driving (2017)

- En la regulación no se hace explícito si es preferible salvar el mayor número de vidas. Pero deja claro que la regulación debe estar en acuerdo con la mayoría de las personas.
- El reglamento hace explícito el que no se debe hacer distinción entre individuos, sin embargo las personas tienen una clara preferencia a salvar la vida de los niños.

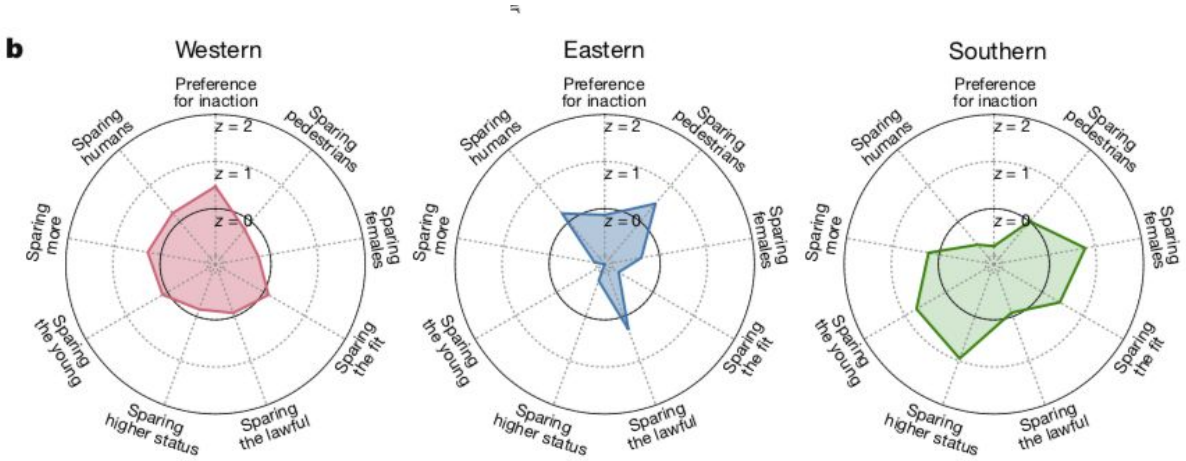
Cluster cultural

Se seleccionaron los 130 países, con más de 100 participantes c/u y se encontraron los siguientes grupos:

- **Grupo Occidental:** America del Norte (menos México) y varios países Europeos Protestantes, Católicos y Cristianos Ortodoxos.
- **Grupo Oriental:** Países del lejano oriente, como Japón y Taiwán, que pertenecen al grupo del Confucionismo e Islamismo, como Indonesia, Pakistán y Arabia Saudita.
- **Grupo del Sur:** Países Latinoamericanos, países con influencia Francesa (Francia, territorios Franceses). Latinoamérica es claramente un subgrupo.



Características culturales



Diferencias significativas

- En países Orientales, con una cultura colectiva, se da preferencia a las personas mayores sobre a los jóvenes.
- En países con mayor desigualdad económica (usando el coeficiente de Gini) se tiende a favorecer a las personas ricas. Lo que refleja que en países con gran desigualdad, las personas son tratadas de manera desigual.

Otras problemáticas

El problema principal no es la aparición de una conciencia maligna, sino la habilidad de hacer decisiones de alta calidad. Donde "calidad" es definida usando una función de utilidad; el valor esperado de la acción de un agente es calculado al multiplicar los valores que el agente le da a cada posible resultado de una acción por su probabilidad y sumando estos valores. Este concepto es utilizado cuando se considera que hay riesgos involucrados en la toma de decisiones. El teorema de "Utilidad Esperada" (expected utility) en microeconomía establece que:

> *Siempre es posible representar las preferencias de un sistema usando el valor esperado de una función de utilidad, a menos que el sistema tenga "vulnerabilidades" Que lo lleven a perder recursos sin beneficios*[1]

De acuerdo a la teoría estándar de decisiones, cuando se comparan alternativas una debe elegir la que tenga mayor utilidad esperada. De hecho los economistas describen sistemas que actúan para maximizar su utilidad esperada como "agentes económicos racionales".

En el caso de una IA es de suponer que la función de utilidad está especificada por un humano. Lo cual presenta dos problemas principales:

1. Es posible que la función de utilidad no esté alineada con los valores humanos que (en el mejor de los casos) son muy difíciles de definir.
2. Cualquier sistema inteligente suficientemente capaz elegirá continuar su propia existencia y adquirir más recursos físicos y computacionales para poder lograr sus metas de manera más eficiente.

Un sistema optimizando una función de n variables, donde el objetivo depende de k<n, tiende a fijar las variables sin restricciones en valores extremos; si esa variable es algo que nos importa la solución puede ser sumamente indeseable.

Entonces, el problema no reside en crear inteligencia. Reside en crear inteligencia que esté alineada con los valores humanos. De lo contrario corremos el riesgo de estar en la historia del Rey Midas (todo lo que toca se convierte en oro).

---

[1] S. M. Omohundro, “The nature of self-improving artificial intelligence.” http://selfawaresystems.com/2007/10/05/paper-on-the-nature-of-self-improving-artificial-intelligence/, October 2007.

How We're Predicting AI—or Failing To

Stuart Armstrong and Kaj Sotala

En realidad no lo sabemos, porque hasta ahora somos malos prediciendo cómo evolucionará el campo de la IA.
Para este estudio se usaron 95 predicciones relacionadas con IAG, como la fecha en la que la primera máquina pasaría el Turing Test.

Turing Test:

Diseñado por Alan Turing en 1950. Él predijo que en el año 2000 máquinas con 100MB de memoria lograrían engañar al 30% de los humanos en un test de cinco minutos.

El Loebner Prize:
El primer sistema en ganar el test fue el chatbot Eugene Goostman en 2014. Este chatbot pretende ser un niño Ucraniano. Goostman ganó ya que pudo convencer a más del 30% de los jueces que era humano.

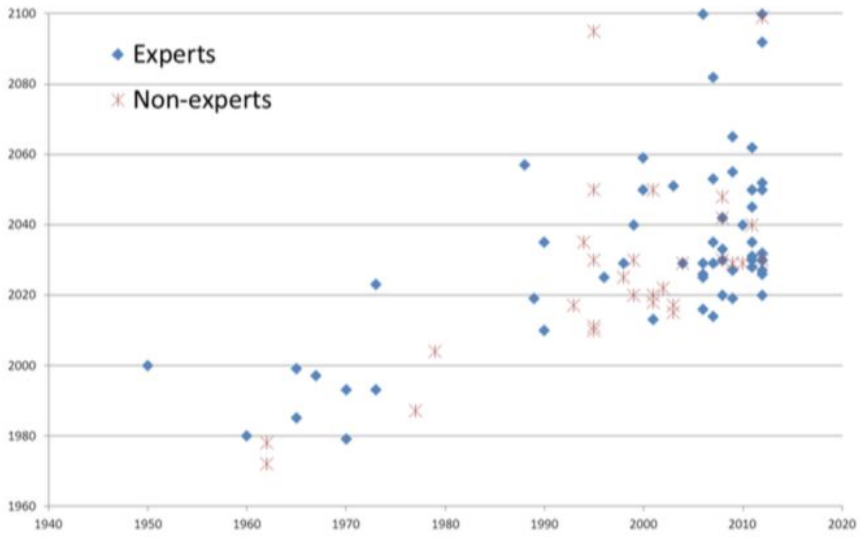


Figure 1: Median estimate for human-level AI, graphed against date of prediction.

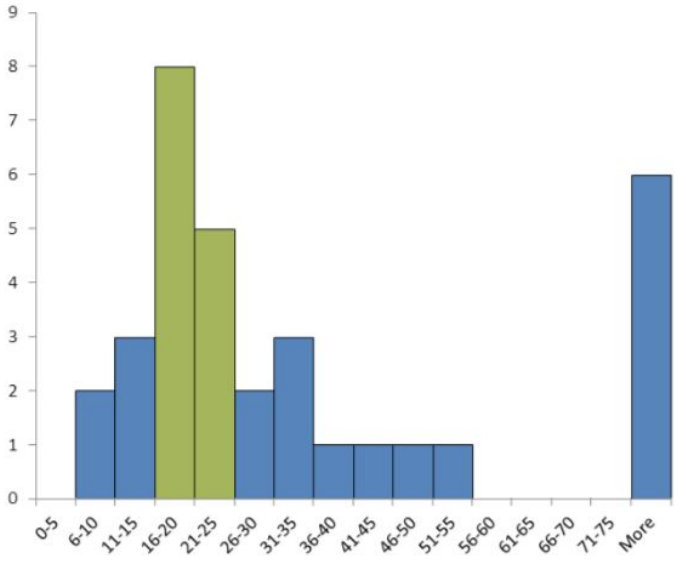


Figure 4: Time between the arrival of AI and the date the prediction was made, for non-expert predictors.

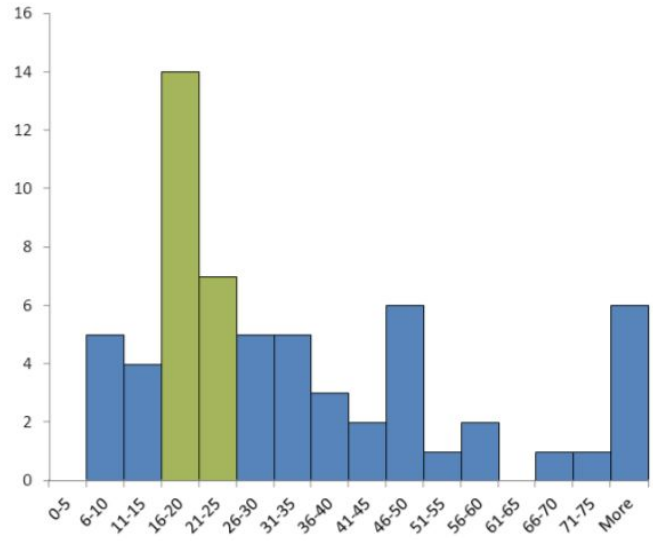


Figure 3: Time between the arrival of AI and the date the prediction was made, for expert predictors.

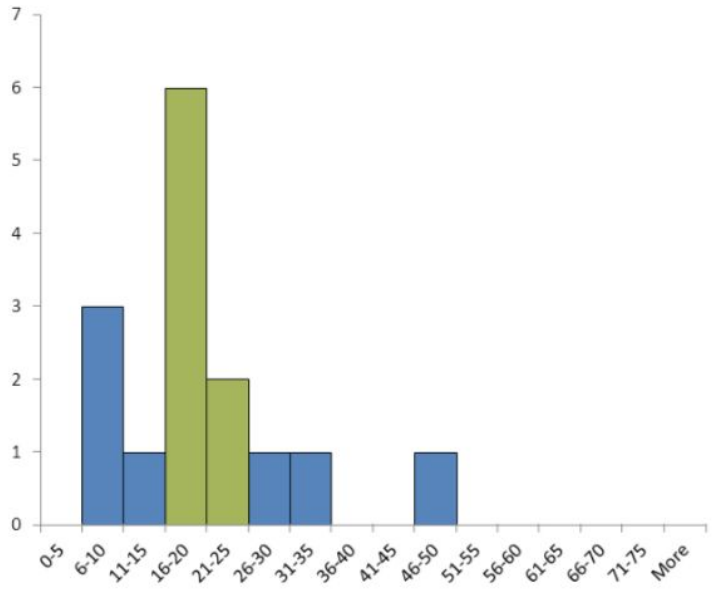


Figure 5: Time between the arrival of AI and the date the prediction was made, for failed predictions.

El único resultado consistente es que los expertos prefieren sis-
temáticamente hacer predicciones de "15 a 25 años en el futuro". Sin embargo, los expertos son indistinguibles de los no expertos haciendo predicciones y tampoco se distinguen predicciones pasadas (fallidas) de las actuales.

Bibliografía:

- Omohundro, Stephen M. *The Basic AI Drives,* Artificial General Intelligence 2008: Proceedings of the First AGI Conference, Frontiers in Artificial Intelligence and Applications 171, pp. 483-492 (2008).
- Stuart Russell, *An Open Letter Research Priorities For Robust And Beneficial Artificial Intelligence,* Future of Life blog (https://futureoflife.org/ai-open-letter/?cn-reloaded=1)
- E. Awad, Sohan Dsouza, R. Kim, J. Schulz, J. Henrich, A. Shariff, J.F. Bonnefon & I. Rahwan, T*he Moral Machine experiment*, Nature, 563, pp. 59-64 (2018).
- S. Armstrong, S., & K. Sotala, *How we're predicting AI or failing to*, Beyond Artificial Intelligence, pp. 11-29 (2015).