

# Ética e inteligencia artificial

Javier Arroyo

Departamento de Ingeniería del Software  
e Inteligencia Artificial

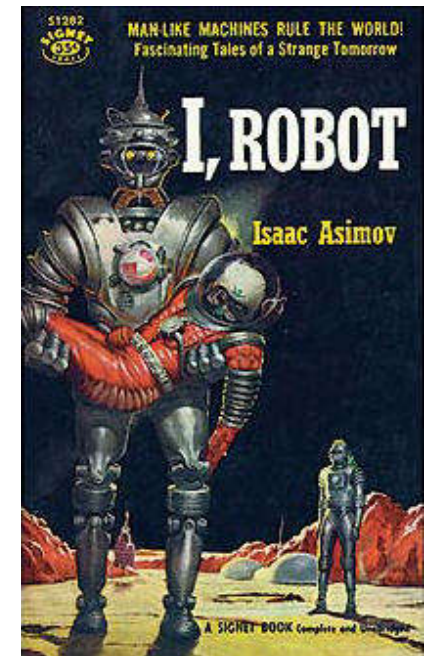


# Ética e inteligencia artificial

- Según Wikipedia, la **ética** es la rama de la filosofía que estudia la conducta humana, lo correcto y lo incorrecto, lo bueno y lo malo, la moral, el buen vivir, la virtud, la felicidad y el deber.
- A medida que un sistema inteligente sea autónomo y tome decisiones complejas, deberá estar guiado por consideraciones éticas
  - Sus acciones pueden afectar a las personas, a otros seres vivos o al medio ambiente




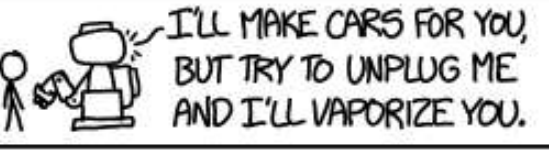

# Las leyes de la Robótica

- En los 40, la ciencia ficción anticipa el problema antes incluso de que surgiera la IA (años 50)
- En su relato “Círculo vicioso” del libro “Yo, robot” de 1942, Isaac Asimov enuncia las tres leyes de la robótica:
  1. Un robot no hará daño a un ser humano o, por inacción, permitirá que un ser humano sufra daño.
  2. Un robot debe cumplir las órdenes dadas por los seres humanos, a excepción de aquellas que entrasen en conflicto con la primera ley.
  3. Un robot debe proteger su propia existencia en la medida en que esta protección no entre en conflicto con la primera o con la segunda ley



# El orden importa

## WHY ASIMOV PUT THE THREE LAWS OF ROBOTICS IN THE ORDER HE DID:

POSSIBLE ORDERING	CONSEQUENCES	
1. (1) DON'T HARM HUMANS 2. (2) OBEY ORDERS 3. (3) PROTECT YOURSELF	[SEE ASIMOV'S STORIES]	BALANCED WORLD
1. (1) DON'T HARM HUMANS 2. (3) PROTECT YOURSELF 3. (2) OBEY ORDERS	EXPLORE MARS!  HAHA, NO. IT'S COLD AND I'D DIE.	FRUSTRATING WORLD
1. (2) OBEY ORDERS 2. (1) DON'T HARM HUMANS 3. (3) PROTECT YOURSELF		KILLBOT HELLSCAPE
1. (2) OBEY ORDERS 2. (3) PROTECT YOURSELF 3. (1) DON'T HARM HUMANS		KILLBOT HELLSCAPE
1. (3) PROTECT YOURSELF 2. (1) DON'T HARM HUMANS 3. (2) OBEY ORDERS		TERRIFYING STANDOFF
1. (3) PROTECT YOURSELF 2. (2) OBEY ORDERS 3. (1) DON'T HARM HUMANS		KILLBOT HELLSCAPE

# Industria 4.0 e IA

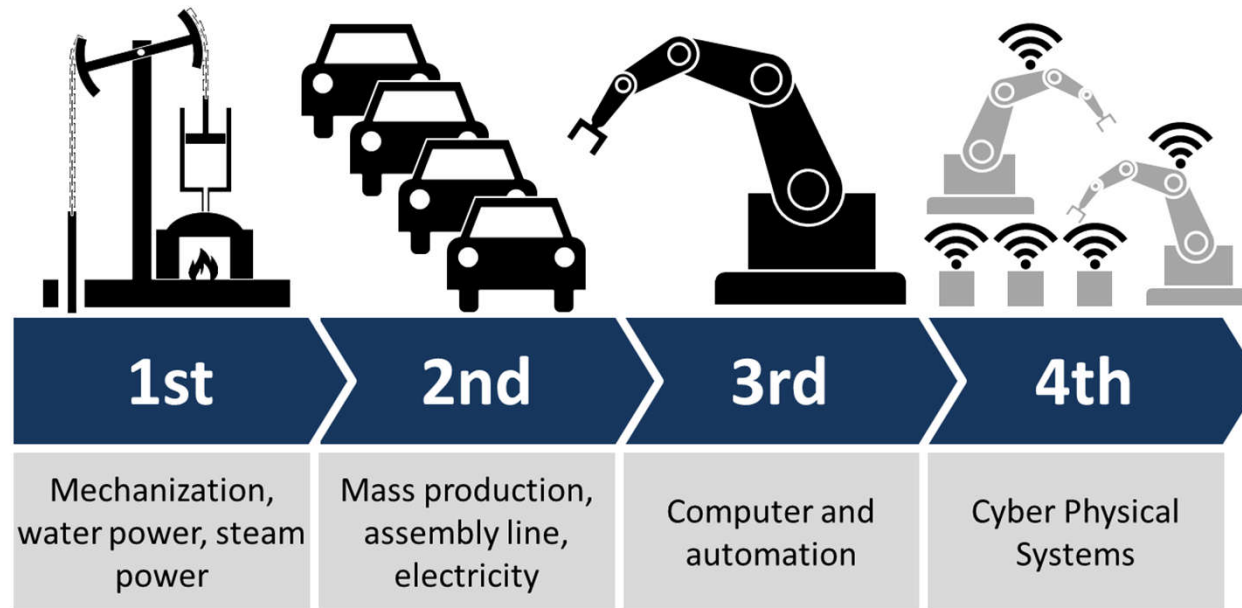


Imagen de Wikimedia Commons

- La IA se introduce en la “3ª revolución industrial”
  - Permite automatizar procesos y hacerlos más eficientes
- A medida que avanzamos hacia la llamada Industria 4.0 los sistemas informáticos y tecnológicos tienen más autonomía
  - No solo complementan al ser humano, lo sustituyen en tareas complejas y toman decisiones
  - Más poder de decisión → Entra en juego la ética

# Pionero de la ética en IA: Joseph Weizenbaum (1923-2008)



Imagen de Wikimedia Commons

- Padre de la IA y creador del chatbot ELIZA (60s)
  - Demostraba que era fácil simular una conversación humana
- Acaba adoptando una postura crítica contra la IA
- Escribe “Computer power and human reason” en 1976 sobre la necesidad de limitar la IA
  - Sobre limitar la capacidad de elección de los ordenadores ya que no tienen cualidades como la compasión o la sabiduría
  - Diferencia entre
    - Decisión: implica cálculo y es programable
    - Elección: implica juicio y factores emocionales (propia de los seres humanos)
- No considera que la IA deba reemplazar al hombre en puestos que requieran empatía, respeto y cuidado
  - Judicatura, ejército, policía, enfermería, atención al cliente...
  - Nos harán sentir devaluados, alienados y frustrados
  - Si lo hacemos es porque pensamos en nosotros mismos como ordenadores



# ¿Existe una ética común? - The Moral Machine

- El MIT realizó un estudio masivo online llamado Moral Machine para ayudar al diseño de vehículos autónomos
  - Recopila 40 millones de decisiones sobre dilemas morales en diez lenguajes tomadas por millones de personas en 233 países y territorios
  - Problemas del tipo “dilema del tranvía”

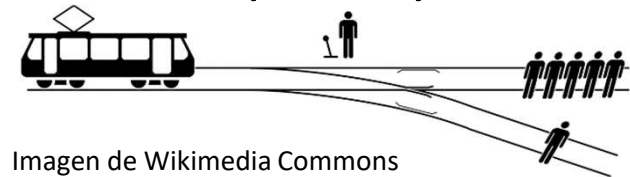


Imagen de Wikimedia Commons

- Encontraron “preferencias morales globales” que pueden servir para desarrollar principios socialmente aceptables para una ética de las máquinas
- Encontraron también variaciones individuales en las preferencias según criterios demográficos
- Encontraron que la variabilidad ética intercultural se podía clasificar en tres grandes grupos de países
  - **No hay una gran ética común, ¡nuestra cultura influye!**

Fuente: Awad, E., Dsouza, S., Kim, R. *et al.* The Moral Machine experiment. *Nature* **563**, 59–64 (2018).

<https://doi.org/10.1038/s41586-018-0637-6>

# Problemas éticos actuales sobre sistemas inteligentes

- Amazon descarta su IA de selección de personal porque favorecía la selección de hombres para los trabajos técnicos
  - Usaba como datos los CVs enviados en los últimos diez años
- Los servicios de reconocimiento facial de Microsoft e IBM son significativamente más precisos para los hombres que para las mujeres y para los blancos que para los negros
  - El problema parece estar en los datos usados para entrenar los sistemas con poca representación de mujeres y de tonos oscuros
- Un software estadounidense utilizado en el sistema judicial sesga las valoraciones de personas blancas hacia “bajo riesgo”
  - Lo hace sin conocer la raza, pero incluye otros factores que correlan con ella (pobreza, desempleo y marginación social)
- Los reguladores financieros investigan la tarjeta de crédito de Apple porque discrimina a las mujeres
  - La ley de Nueva York (que aplica en el caso) sanciona que con intención o no se dé un tratamiento discriminatorio a las mujeres o a cualquier otro grupo protegido

Fuentes: <https://www.theguardian.com/technology/2018/oct/10/amazon-hiring-ai-gender-bias-recruiting-engine>  
<https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>  
<https://www.wired.com/story/photo-algorithms-id-white-men-fineblack-women-not-so-much/>  
<https://www.theverge.com/2019/11/11/20958953/apple-credit-card-gender-discrimination-algorithms-black-box-investigation>



# Problemas éticos actuales sobre sistemas inteligentes

- Los datos que se usan para que el sistema aprenda son clave
  - Estos datos pueden reflejar una situación que ya no es válida hoy día y que puede reproducir sesgos o discriminaciones
  - ¿Se ha tenido en cuenta todo el universo de casos posibles? Si no es así, casos “nuevos” pueden tener un resultado inesperado.
- La función objetivo que el sistema busca “optimizar” también es clave. ¿Cuál es? ¿Qué efecto tiene al usarse repetidas veces?
  - Puede favorecer un tipo de perfil por ser “óptimo” según un criterio, pero perjudicar a otros que son válidos también.
- No es menos importante entender qué salida da el sistema y qué uso se le puede dar
  - ¿El sistema hace un filtrado, sugiere una decisión, toma la decisión?
  - ¿Alguien supervisa la salida del sistema? ¿Puede ver por qué el sistema ha tomado la decisión? ¿Qué pasa si el responsable cambia la decisión y se equivoca?
  - ¿Se le está dando un uso responsable y adecuado al sistema?
- Aun así, no es sencillo. ¿Se puede acusar a un sistema de discriminación si las variables discriminatorias no entraron en juego?
  - Hay que poder validar que esto es así y que el sistema no encontró “huecos” para saberlo.
  - Si no las ha usado, pero produce un trato discriminatorio y esto no es adecuado, entonces la IA debe ser rediseñada con eso en mente (igualdad de oportunidades *by-design*)

# Hacia una IA ética

- Las respuestas a estos problemas no son sencillas, pero pasan porque la IA incorpore “requisitos sociales” en su desarrollo
- Algunas ideas a tener en cuenta al desarrollar IA
  - La IA debe ser **transparente y auditable**: debe poderse revisar y entender las decisiones que toma y va a tomar
    - XAI (eXplainable Artificial Intelligence) tiene gran auge en investigación hoy día
  - La IA debe ser **predecible** para aquellos sobre los que va a “regir”: así estas personas pueden ajustar su comportamiento para interactuar con ella.
    - De la misma forma que las leyes ofrecen un entorno predecible
  - La IA debe ser **robusta contra la manipulación**: este requisito es típico de los sistemas de seguridad, pero también debe serlo ahora de los sistemas basados en IA
- También es necesario que **se establezcan responsables** por el comportamiento de la IA y que la responsabilidad no se diluya
  - Esto es a menudo complicado incluso sin haber IA de por medio

Fuente: Bostrom, N., Yudkowsky, E. The Ethics of Artificial Intelligence. Publicado en K. Frankish & W. Ramsey (Eds.), *The Cambridge Handbook of Artificial Intelligence* (pp. 316-334). Cambridge University Press. (2014)

<https://doi.org/10.1017/CBO9781139046855.020>

# La no-discriminación en IA

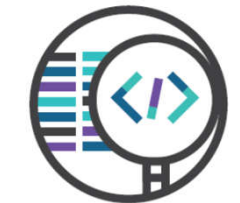
- El Foro Económico Mundial ofrece recomendaciones en un libro blanco publicado en 2018
  - **Inclusión activa:** Especialmente de aquellos perfiles que vayan a verse afectados por el sistema.
  - **Justicia:** Los desarrolladores deben promover que el sistema sea justo, según el concepto de justicia que mejor encaje en cada caso.
  - **Derecho a entender:** Los sistemas deben ser capaces de ofrecer una explicación de forma que su toma de decisiones pueda ser comprensible y revisable por la autoridad competente.
    - Si esto no es posible, entonces se debe estudiar si el sistema debe ser usado o no.
  - **Acceso a la compensación:** Si un sistema tiene un efecto negativo sobre personas, los responsables deben proponer a éstas vías de compensación.



Active Inclusion



Fairness



Right to Understanding



Access to Remedy

# La no discriminación en IA

- Hay activistas como Joy Buolamwini, investigadora del MIT Media Lab, que investiga para identificar sesgos en los algoritmos y desarrollar prácticas que fomenten la transparencia y la responsabilidad durante su diseño
- Con su investigación Buolamwini denunció el sesgo discriminatorio de los reconocedores faciales de empresas como IBM, Amazon y Microsoft y consiguió su rectificación
  - [Su charla TED](#) sobre el tema tiene más de 1 millón de visitas
- Buolamwini ha creado el programa [Algorithmic Justice League](#) que pretende sacar a la luz los sesgos algorítmicos discriminatorios



# Cartas abiertas sobre IA y ética

- En 2014 Stephen Hawking y Elon Musk, miembros del panel científico asesor del Future of Life Institute, impulsan una carta abierta sobre la IA y la ética
  - *[...] ya que todo lo que la civilización es es un producto de la inteligencia humana; no podemos predecir lo que podríamos lograr cuando esta inteligencia se magnifique con las herramientas que la IA puede proporcionar [...] Debido al gran potencial de la IA, es importante investigar cómo cosechar sus beneficios y al mismo tiempo evitar sus posibles riesgos.*
- La carta establece prioridades de investigación
  - Optimizar el impacto económico: maximizar beneficios y minimizar los efectos adversos (paro y desigualdad)
  - Investigación en ética y derecho: responsabilidad de las decisiones, privacidad, políticas públicas...
  - Inteligencia artificial robusta: necesidad de que los sistemas se comporten como se espera de ellos
- En IJCAI 2015 (International Joint Conferences on Artificial Intelligence) científicos de renombre presentaron una carta abierta para rechazar el desarrollo de armas autónomas

Fuente: <https://futureoflife.org/ai-open-letter>  
<https://futureoflife.org/open-letter-autonomous-weapons/>

# La ética y la robótica en las leyes europeas

- En 2017 el Parlamento Europeo aprobó un informe sobre Robótica que establecía un **Código Ético de Conducta para Ingenieros en Robótica**
- Se deben respetar principios como:
  - Beneficencia: los robots deben actuar en beneficio del hombre
  - Principio de no perjuicio o maleficencia: la doctrina de «primero, no hacer daño», en virtud del cual los robots no deberían perjudicar a las personas
- Además, se abordan aspectos como:
  - Proteger a los humanos del daño causado por robots: la dignidad humana
  - Proteger la libertad humana frente a los robots
  - Proteger la privacidad y el uso de datos: especialmente cuando avancen los coches autónomos, los drones, los asistentes personales o los robots de seguridad
  - Protección de la humanidad ante el riesgo de manipulación por parte de los robots: Especialmente en colectivos –ancianos, niños, dependientes– que puedan generar una empatía artificial
  - Evitar la disolución de los lazos sociales haciendo que los robots monopolicen las relaciones de determinados grupos.
  - Igualdad de acceso al progreso en robótica (brecha robótica)



# La IA y los derechos humanos

- El uso de la IA tendrá un efecto disruptivo en la distribución de poder en el mundo
  - Puede crear nuevas formas de opresión sobre los más vulnerables
  - La legislación doméstica o el autocontrol de las empresas puede resultar insuficiente
- La aplicación práctica de la IA debe respetar los derechos humanos
  - Los derechos humanos pueden tener un significado más claro que apelar a principios éticos
  - Los derechos humanos han sido una herramienta útil para combatir desigualdades y conflictos en el mundo
- Es importante que los derechos humanos sean tenidos en cuenta como un **valor añadido** en el desarrollo de la IA

Fuente: C. van Veen (2018) Artificial Intelligence: What's Human Rights Got To Do With It?

<https://points.datasociety.net/artificial-intelligence-whats-human-rights-got-to-do-with-it-4622ec1566d5>

# Diseñando sistemas inteligentes éticos

- Bill Hibbard de la U. de Wisconsin Madison ha escrito [“Ethical Artificial Intelligence”](#) sobre cómo diseñar sistemas inteligentes éticos
- IA actual (p.ej. coche autónomo)
  - Modelo del entorno diseñado por humanos
  - Restricciones de seguridad explícitas sobre su comportamiento reflejadas en el modelo
- IA futura
  - Modelo del entorno muy complejo para el entendimiento humano y que debe ser aprendido (no programado)
  - Restricciones de seguridad que no pueden ser explícitas sobre el modelo aprendido
    - Reglas de seguridad que serán forzosamente ambiguas (como las leyes de la robótica de Asimov)
    - Mayor complejidad → mayor dificultad → mayor riesgo

Fuente: Bill Hibbard. An analytical framework for Ethical AI

<https://slideplayer.com/slide/6422524/>

# Riesgos sociales y políticos

- La IA es una herramienta de competición militar y económica
- Una élite de personas que controlan los servicios de IA usados en dispositivos electrónicos podrían controlar la sociedad
- La distribución normal de la inteligencia humana será reemplazada por la distribución de “ley de potencias” de inteligencia artificial
  - Habrá **superinteligencias artificiales**
  - El humano medio quedará detrás de esas inteligencias



Fuente: Bill Hibbard. An analytical framework for Ethical AI

<https://slideplayer.com/slide/6422524/>

Imagen tomada de wisc.edu

# La ética y las superinteligencias

- Una superinteligencia es un intelecto que **supera ampliamente** a los mejores cerebros humanos en prácticamente todos los campos, incluyendo la creatividad científica, la sabiduría y las habilidades sociales
  - Deep Blue o AlphaZero no lo serían porque son mejores que los humanos en un pequeño dominio (ajedrez, go, etc.)
- Algunos autores, como Bostrom, Kurzweil o Moravec, creen que su creación puede ser cuestión de unas pocas décadas
- Ante esa posibilidad y sus enormes consecuencias, filósofos como Nick Bostrom de la Universidad de Oxford, han abordado el tema desde el punto de vista ético



Fuente: Nick Bostrom. Ethical issues in advanced artificial intelligence  
<https://nickbostrom.com/ethics/ai.html>

Imagen de Wikimedia Commons

# La ética y las superinteligencias

- Su aparición aceleraría el progreso tecnológico
  - Podría ser el “último” invento humano, ya que la superinteligencia sería mejor en ciencia y tecnología que los hombres
- Al ser la ética una actividad cognitiva, una superinteligencia podría superar en esos dilemas a los pensadores humanos
  - Razonar y pesar mejor la evidencia
  - Valorar mejor los efectos a largo plazo de decisiones y políticas
- Sería potencialmente autónoma al tener iniciativa propia
- No necesariamente tendrá motivaciones humanas, sus motivaciones pueden ser arbitrarias a nuestros ojos
- Es fundamental que sus motivaciones iniciales se basen en valores filantrópicos y que su objetivo último sea la amistad
  - Un amigo que busca cambiar para hacerte daño no es un amigo
  - Un verdadero amigo busca mantener el cariño hacia su amigo
  - Eso evitaría que se transformara en una amenaza para nosotros

Fuente: Nick Bostrom. Ethical issues in advanced artificial intelligence

<https://nickbostrom.com/ethics/ai.html>

# O puede que los robots nunca lleguen a dominarnos...



Thanks to machine-learning algorithms,  
the robot apocalypse was short-lived.

Fuente: <https://www.smbc-comics.com/comics/1538492931-20181002.png>