

Inteligencia Artificial

Ética e Inteligencia Artificial

Prof. Wladimir Rodríguez
wladimir.rodriguez@outlook.com
Departamento de Computación
Escuela de Ingeniería de Sistemas

Traducción de las diapositivas:

CS 343

AI: Ethics and Society

Prof. Scott Niekum — The University of Texas at Austin

Verdadero o Falso

- * La IA reemplazará a los seres humanos en la mayoría de los trabajos
- * La IA sobrepasará la inteligencia humana en X años
- * La IA funciona de manera similar al cerebro humano
- * Los sistemas de IA "piensan"
- * Los sistemas de IA tienen sus propios deseos y metas
- * Los sistemas de IA pueden hacer cosas para las cuales no fueron diseñados
- * Los sistemas de IA se pueden convertir en conscientes
- * Se puede confiar en los sistemas de IA
- * Los algoritmos de IA pueden discriminar o exhibir prejuicios

Ética en las Ciencias de las Computadoras

- * Ética de las máquinas
 - * El comportamiento moral de los seres artificiales inteligentes
- * Ética robótica
 - * El comportamiento moral de diseñar seres artificiales inteligentes
- * Ética computacional
 - * El comportamiento moral de usar computadores y sistemas computarizados

¿Qué son Marcos Éticos?

- * Sistemas que guían las decisiones éticas y proveer una razón por esa decisión.
- * Esto es un problema no resuelto.
 - * Numerosos enfoques que resultan en muy diferentes resultados y comportamientos
- * Tres marcos generales:
 - * Marco basado en responsabilidad
 - * Marco consecuencialista
 - * Marco de virtudes

Teorías Éticas

- * No consecuencialistas
 - * Se ocupan de la intención del agente en vez de las consecuencia
- * Consecuencialistas
 - * Se ocupan de las consecuencias de las acciones del agente
- * Centradas en el agente
 - * Se ocupan de la composición ética del agente

No Consecuencialista (Basada en Deberes)

- * Comúnmente asociada con el “*imperativo categórico*” de Immanuel Kant
 - * “*Obra sólo según una máxima tal que puedas querer al mismo tiempo que se torne ley universal*”
- * Conducta ética significa elegir acciones que son correctas y buenas
- * Considerar deberes y obligaciones al elegir

¿Problemas?

- * Buenas intensiones se valoran sobre los buenos resultados
- * No responde cómo actuar cuando dos deberes entran en conflicto
- * No proporciona una definición de comportamientos éticos

Consecuencialista

- * Basada en la filosofía Utilitaria
- * Pesar lo positivo y negativo producido por la acción para determinar la mejor acción en general
- * Conducta ética significa tratar de hacer el mayor bien y el menor mal
- * Considerar el impacto sobre todos los individuos involucrados al elegir

¿Problemas?

- * Las necesidades de muchos anulan las necesidades de unos pocos.
- * Cualquier acción puede justificarse si sale suficiente bien
- * No aborda cómo predecir resultados basados en acciones

Centrado en Agente (Virtud)

- * Basado en ideas de Aristóteles y Confucio.
 - * Los agentes deben actuar de acuerdo con su yo ideal
- * Conducta ética significa determinar los rasgos y comportamientos de un agente y construir sobre aquellos que fomentan el bien
- * Considera la totalidad de la vida de un agente en lugar de acciones individuales

¿Problemas?

- * Se centra en el carácter personal en lugar de un sistema para determinar la acción.
- * El enfoque de alto nivel requiere una profunda comprensión e interpretación para implementarlo de manera efectiva
- * No define rasgos virtuosos

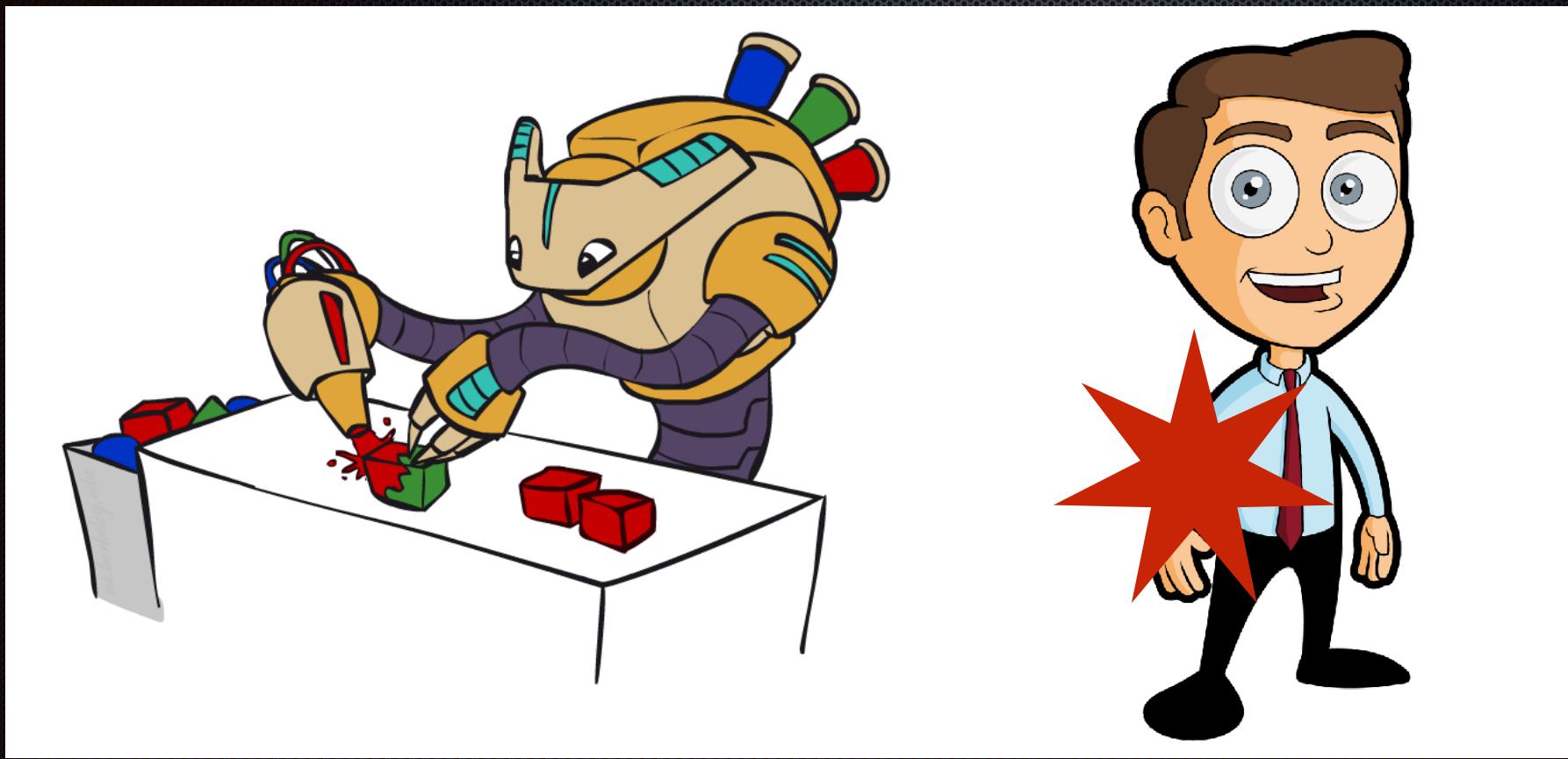
Dilemas Éticos

- * *Rushworth Kidder define los dilemas éticos como elecciones correctas versus correctas:*
- * *Verdad versus lealtad*
- * *Justicia versus misericordia*
- * *Uno versus muchos*
- * *Corto plazo versus largo plazo*

Actuar Racionalmente

- * *Actuar racionalmente simplemente significa maximizar la utilidad*
- * *¿pero esto puede salir mal?*

¿Consecuencias imprevistas de maximizar la utilidad?



¿Qué salió mal?

- * ¿Es esto realista?
 - * Los robots no son lo suficientemente inteligentes como para aprender de un ejemplo. Pero supongamos que podrá en el futuro.
 - * No tendría el concepto de "ser humano". Solo sabe de colores y cubos.

¿Qué salió mal?

- * Mal diseño!
- * Los objetivos deben diseñarse cuidadosamente: el robot solo debe ser recompensado por hacer cubos rojos, no por ningún otro objeto.
- * Las acciones deben ser limitadas: solo las acciones disponibles deben ser recoger un bloque o pintar un bloque.
- * Se deben verificar los planes de seguridad antes / durante la ejecución: cancele cualquier trayectoria que entre en contacto con un humano.
- * No continúe aprendiendo después de la implementación.

¿Qué salió mal?

- * ¿Es esto más peligroso que cualquier fábrica con maquinaria no inteligente que no se detiene automáticamente si hay alguien en el camino?
- * Es un mal diseño, ¡pero sabemos cómo usar la ingeniería para evitar estas situaciones!

Características humanas versus IA

Humano	IA
Evolucionado para sobrevivir	Diseñado por ingenieros
Establece objetivos propios	Objetivos programados explícitamente
Sistema complejo de uso general	Sistema específico y restringido
Continuamente aprende	Puede desactivar el aprendizaje o no usar el aprendizaje
Aprende de todos los datos observados.	El acceso a los datos puede ser controlado
Aprende solo de las propias experiencias.	Puede compartir datos con otros robots
Puede hacer cualquier elección en cualquier momento	Las acciones disponibles pueden restringirse

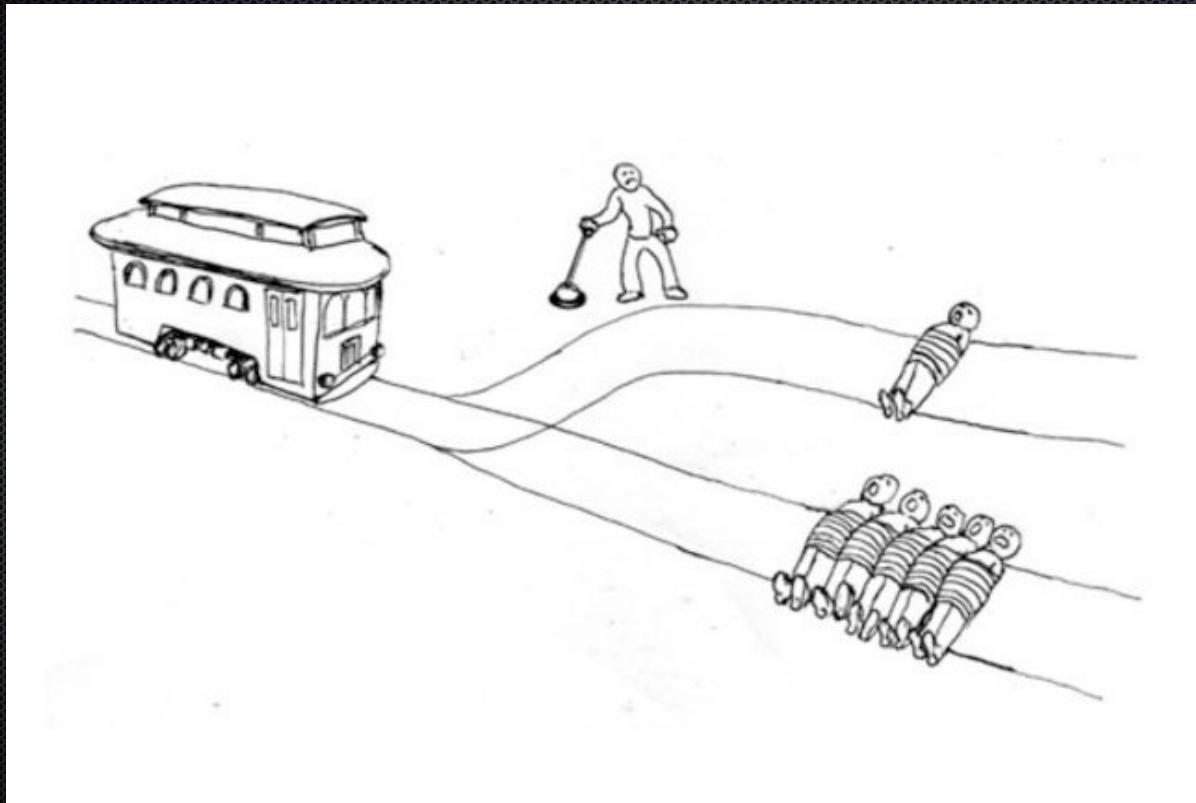
Riesgos reales de la IA: Paro masivo debido a la automatización



Riesgos reales de la IA: Pruebas de calidad inferior / poca comprensión del usuario



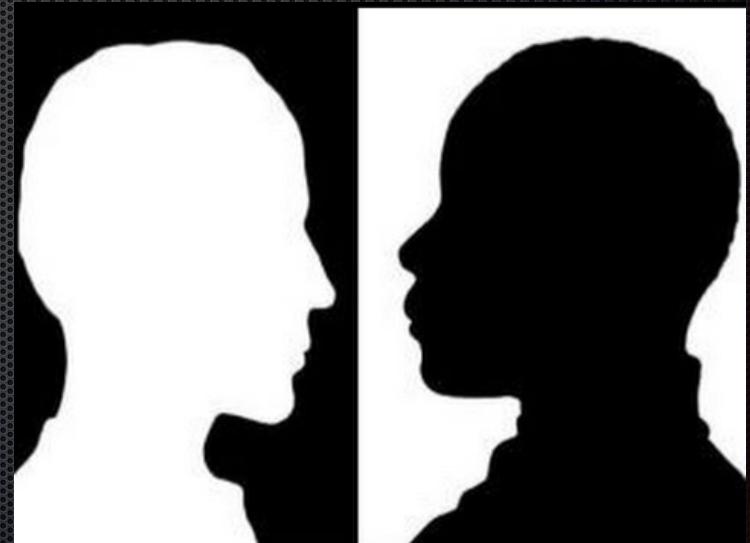
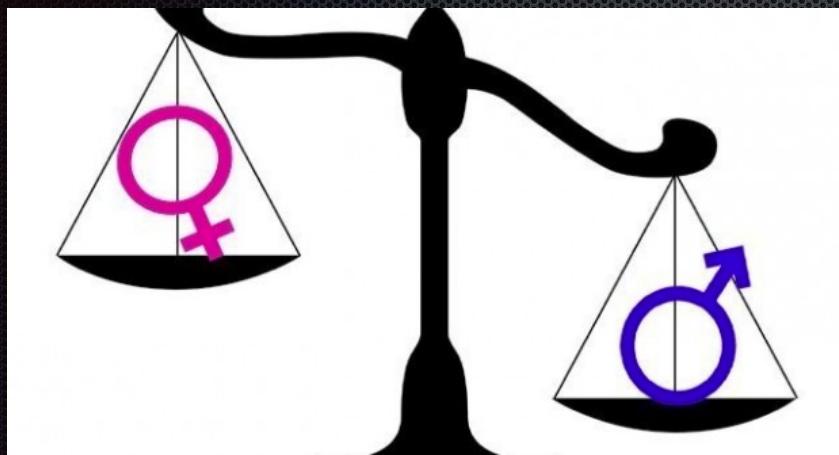
Riesgos reales de la IA: ¿Cómo tomar decisiones difíciles?



Riesgos reales de la IA: Preocupaciones sobre la privacidad



Riesgos reales de la IA: Discriminación algorítmica



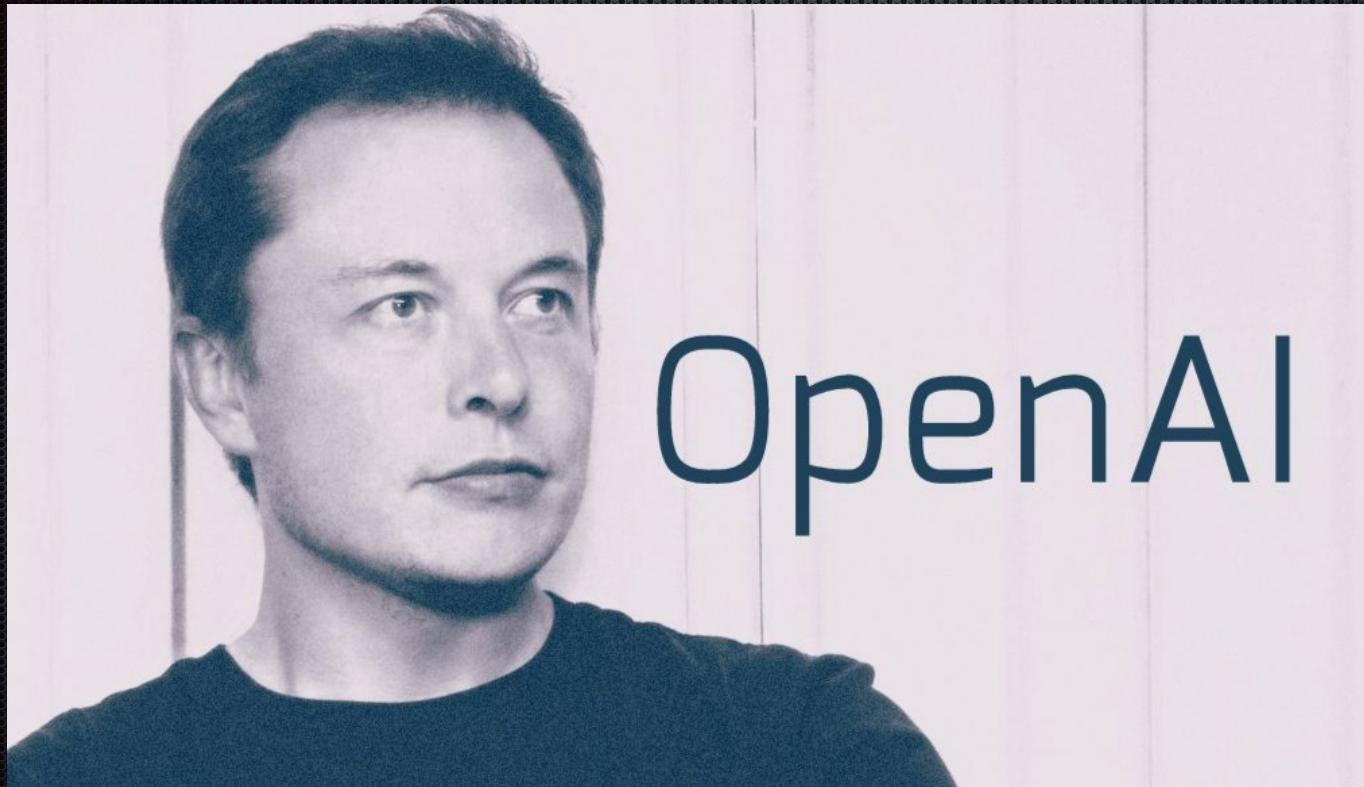
Riesgos reales de la IA: Manipulación emocional poco ética



Riesgos reales de la IA: Uso poco ético: ¿guerra con drones?



Riesgos reales de la IA: IA en las "manos equivocadas"



Beneficios realistas de la IA

La pregunta central:

¿Podemos asegurarnos de que los beneficios de la IA superen los riesgos potenciales?

Beneficios realistas de la IA

Reducción significativa de las muertes por conducir



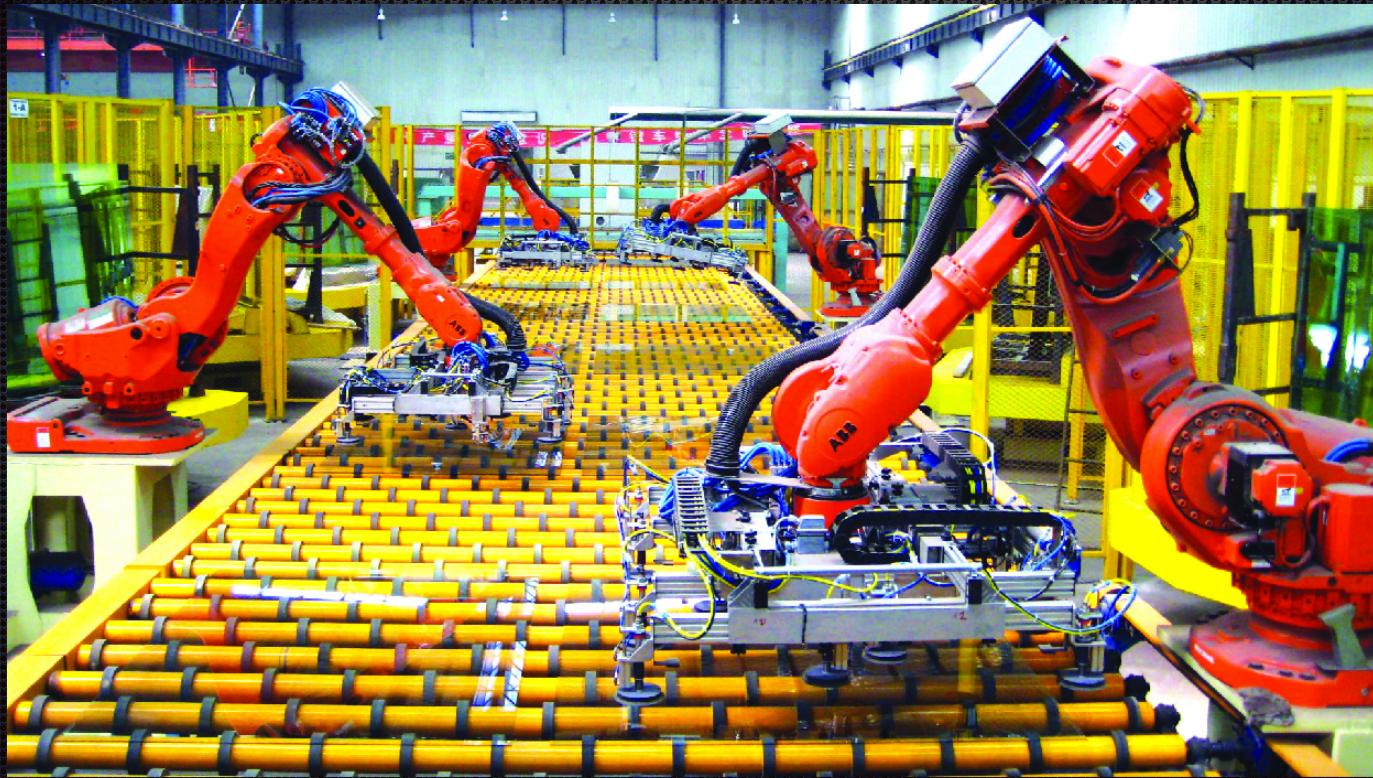
Beneficios realistas de la IA

Vidas más felices y saludables



Beneficios realistas de la IA

Mayor productividad y prosperidad



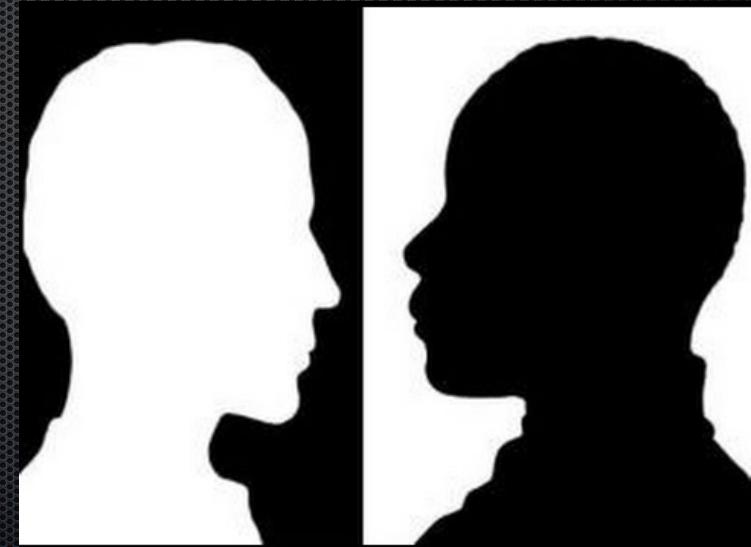
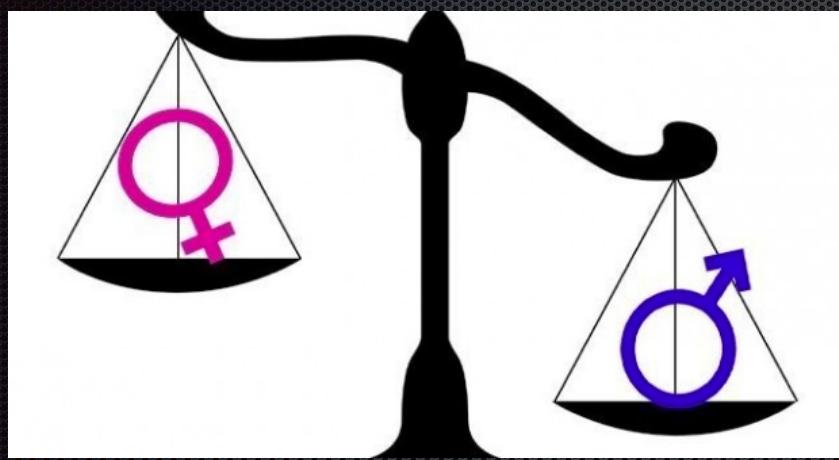
Beneficios realistas de la IA

Sucio, peligroso y aburrido



Beneficios realistas de la IA

Mayor justicia social



Beneficios realistas de la IA

Más allá de las capacidades humanas

