Inteligencia Artificial

¿Qué queremos?

Fernando Esponda

ITAM División Académica de Ciencias de la Computación

Qué es Inteligencia Artificial

- Inteligencia Artificial
  - Robótica
  - Visión
  - Demostración automática de teoremas, verificación de código
  - Aprendizaje de máquina *
- IA débil vs fuerte
    - Acotada vs abierta, general

Riesgos y Beneficios

Beneficios:

- Potenciar nuestras habilidades
- Liberarnos de trabajo tedioso o peligroso

Riesgos:

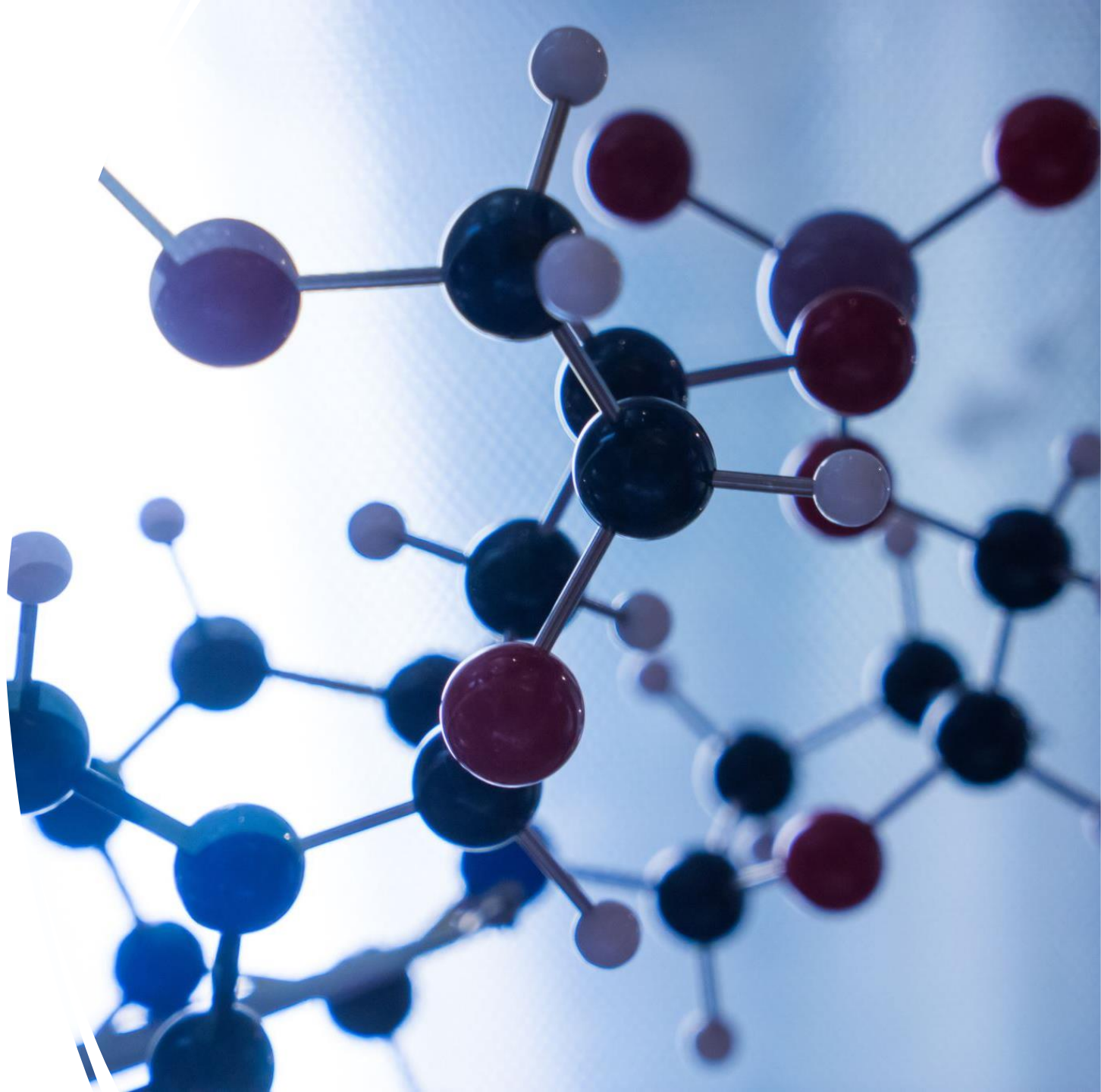
- Que funcionen mal
- Que funcionen bien

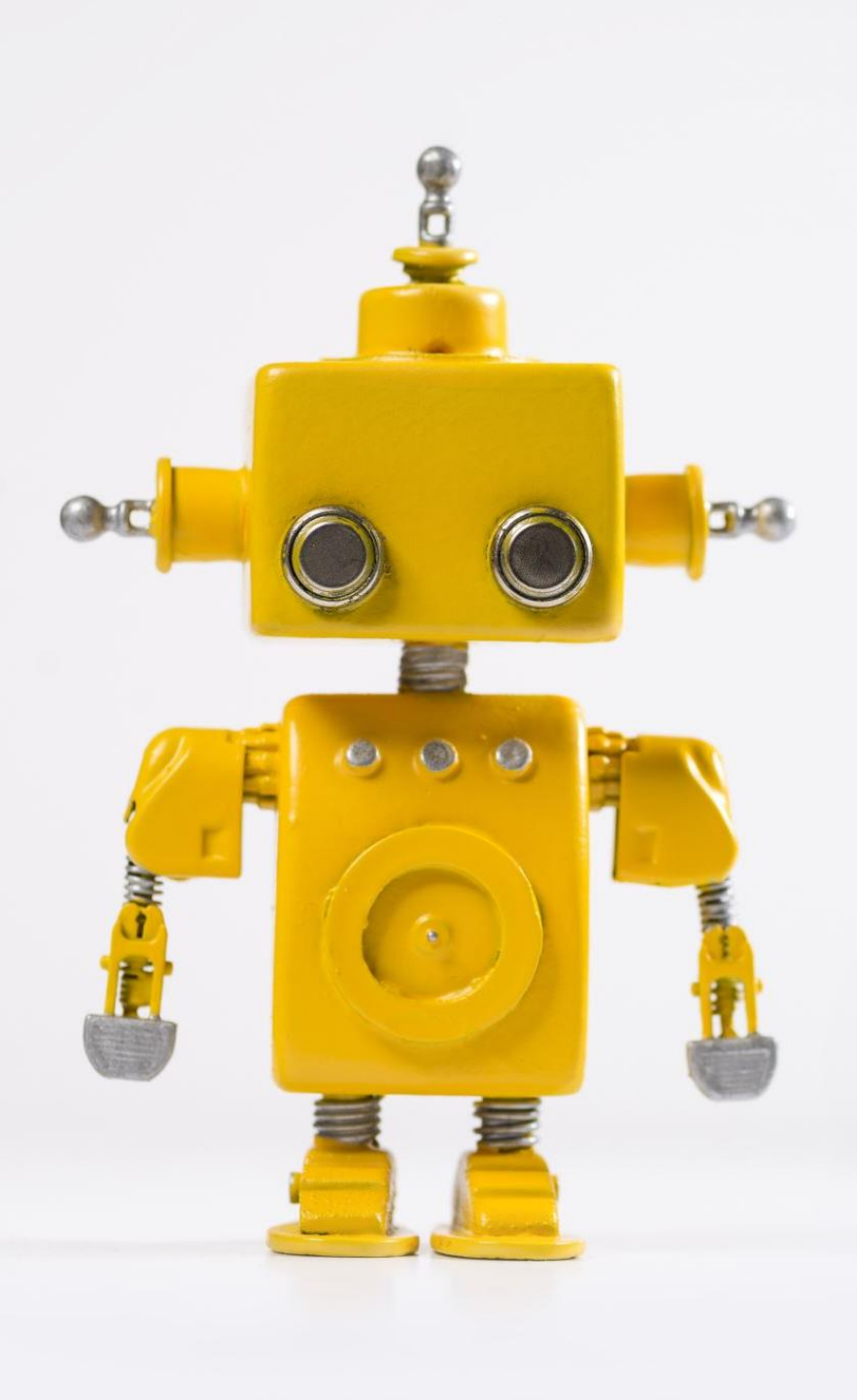
IA Segura

- ”Que los sistemas de IA promuevan los objetivos y valores deseados de los seres humanos” parafraseado a Norbert Wiener (en los 60’s)
  - De cuáles seres humanos?
- Por una parte tenemos que asegurarnos que la IA haga lo que teníamos la intención que hiciera
- Por otra, pensar en si esa intención es la adecuada y qué consecuencias puede tener
  - La escala que tienen las aplicaciones de IA hace esto muy importante

Una definición

- Los sistemas de aprendizaje de máquina son **modelos** que a través de **datos** ajustan sus parámetros para optimizar una **función objetivo**
- Cada uno de estos puntos debe ser estudiado a detalle, monitoreado y en ocasiones regulado
- Qué no es la IA?
  - Totalmente objetiva
  - Libre de errores
  - Mágica





La función objetivo

- Es en ocasiones difícil especificar o codificar exacta y completamente el objetivo que intenta lograr un algoritmo
  - Maximizar el número de clicks a contenido puede favorecer la creación de contenido escandaloso y falso
  - Yo robot: Proteger al ser humano lleva a que los robots nos encarcelen pues somos nuestro peor enemigo
  - Efecto del “evil genie” y el toque de midas. Nick Bostrom
- Cómo alineamos los objetivos que le damos a la IA con los objetivos del ser humano?
  - Redes sociales
  - No ven el ”big picture”

Los datos

Sesgos

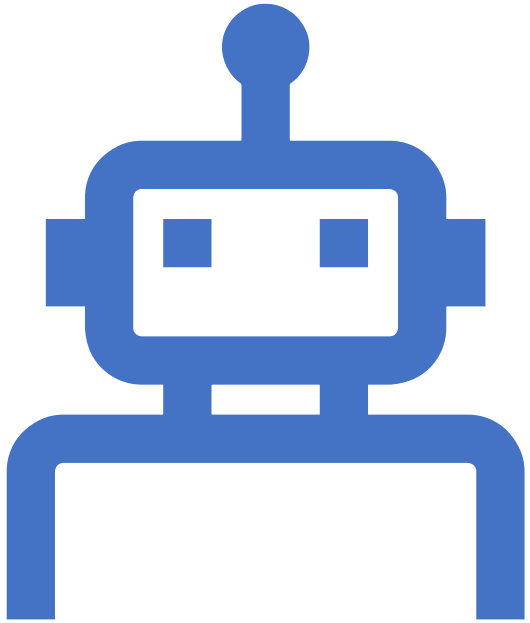
- Muestras incompletas
- Los datos contienen nuestros sesgos por lo tanto los modelos los replican
  - Racismo
    - Google photo
  - Antisemitismo
    - Chatbot Tay
  - Ideología política
    - ChatGPT

Remedios

- Estar conscientes de los sesgos
  - Usar técnicas para medirlos
  - Tomar medidas para reducirlos
- ”Fairness through blindness” No funciona
  - Caso: Sexismo en la orquesta de Boston
  - Código postal y color de piel

Privacidad

- ¿Qué datos se pueden usar?
- ¿Qué inferencias se vale hacer?



El modelo

- Complejidad correcta, que generalice bien
- Que el procedimiento de creación sea auditable y correcto
- Transparente
  - Capaz de explicar una decisión
  - Documentado
- Que sea mantenido (reentrenado) regularmente

El caso de COMPAS

- **Función objetivo**
  - Probabilidad de un individuo a cometer un crimen (reincidir)
- **Modelo**
  - Sin trasparencia: Algoritmo propietario
  - Usado para determinar si se deja en libertad bajo fianza
  - Usado para determinar sentencias
- **Datos**
  - No se sabe, pero se infiere: Historia criminal y aspectos socio-demográficos
- Un estudio de ProPublica sugiere la existencia de sesgo racial
  - Los afroamericanos que no reinciden tiene una calificación de riesgo más alta que los blancos que no reinciden
- Consideraciones importantes
  - El modelo esta calibrado (de 100 individuos con calificación de 80% de riesgo 80 reinciden)
  - Resultados teóricos demuestran que no es posible tener un modelo calibrado y sin sesgo en los errores de falsos positivos y negativos (a menos de que las distribuciones sean idénticas)
  - **Cuál es el concepto de justicia que queremos codificar**?
    - Qué errores importan más?

Otras consideraciones

Riesgos cuando funcionan bien

Desigualdad

- Entre los que emplean, manejan o comandan la tecnología y quienes no

Realidad

- La distinción entre textos escritos o no por humanos, entre fotos y videos reales se vuelve más complicado (más si nos encapsulamos en grupos que piensan igual a nosotros)

Agencia

- Delegar nuestras decisiones a un algoritmo inescrutable
- Lo que vemos en la tele, con quién salimos, qué comparamos
- Corremos el riesgo de convertirnos en agentes pasivos, espectadores de nuestra propia vida

Gracias

Lo Bueno

Avances en ciencia, medicina, tecnología

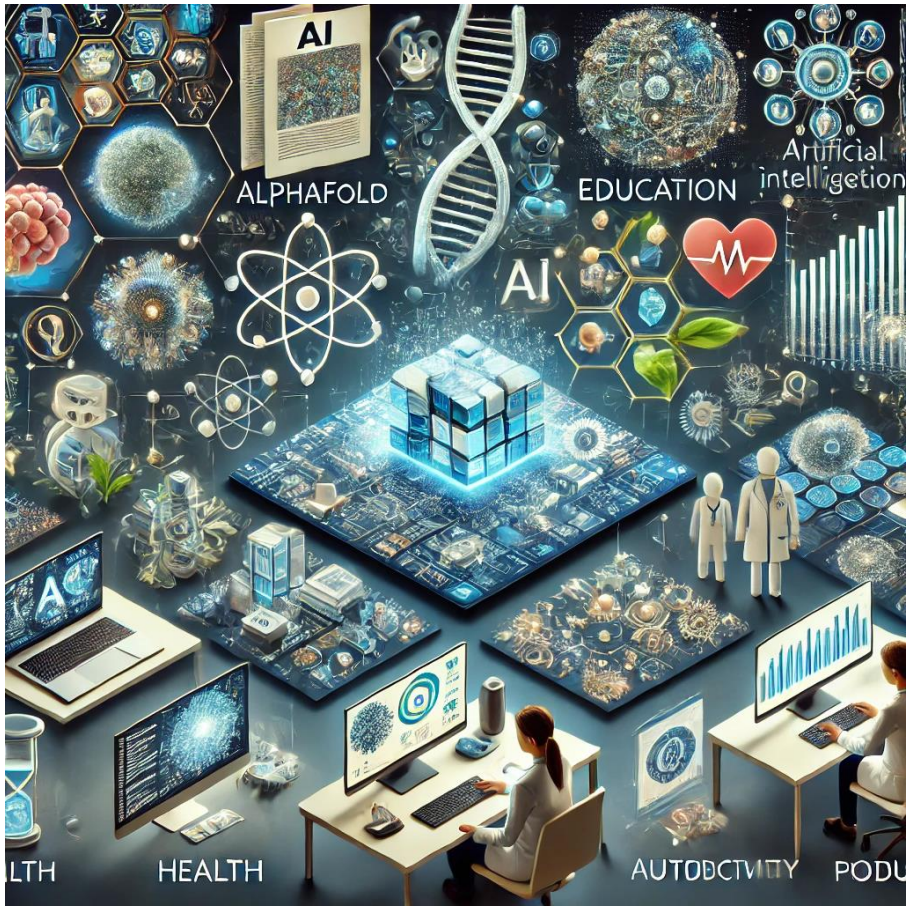
- AlphaFold, “drug repurposing”
- Cambio climático (predicciones y optimizaciones)

Educación y asistencia

- A través de herramientas como GPT se aumenta el potencial para aprender
  - La misma información, pero administrada de manera diferente. Personalizada
- Es como tener un tutor experto en la disciplina que te interesa, bueno para enseñar y paciente
- Asistencia médica, emocional y creativa

Productividad

- Asistentes inteligentes
- Ir más lejos. Puedo yo solo ahora quizá diseñar una enzima para limpiar el mar. Trabajar en equipo sin equipo
- Trabajos tardados se pueden eficientar (correos, reportes, presentaciones)
- Análisis de datos para tomar mejores decisiones



Para adelante

- La IA llegó para quedarse
- Agentes
- La historia de cómo usarse, de cuáles son nuestros derechos, cuáles las obligaciones de los proveedores y las cualidades que deben tener los modelos está por escribirse