GLOBAL CX FORUM

28º CONGRESO MÉXICO

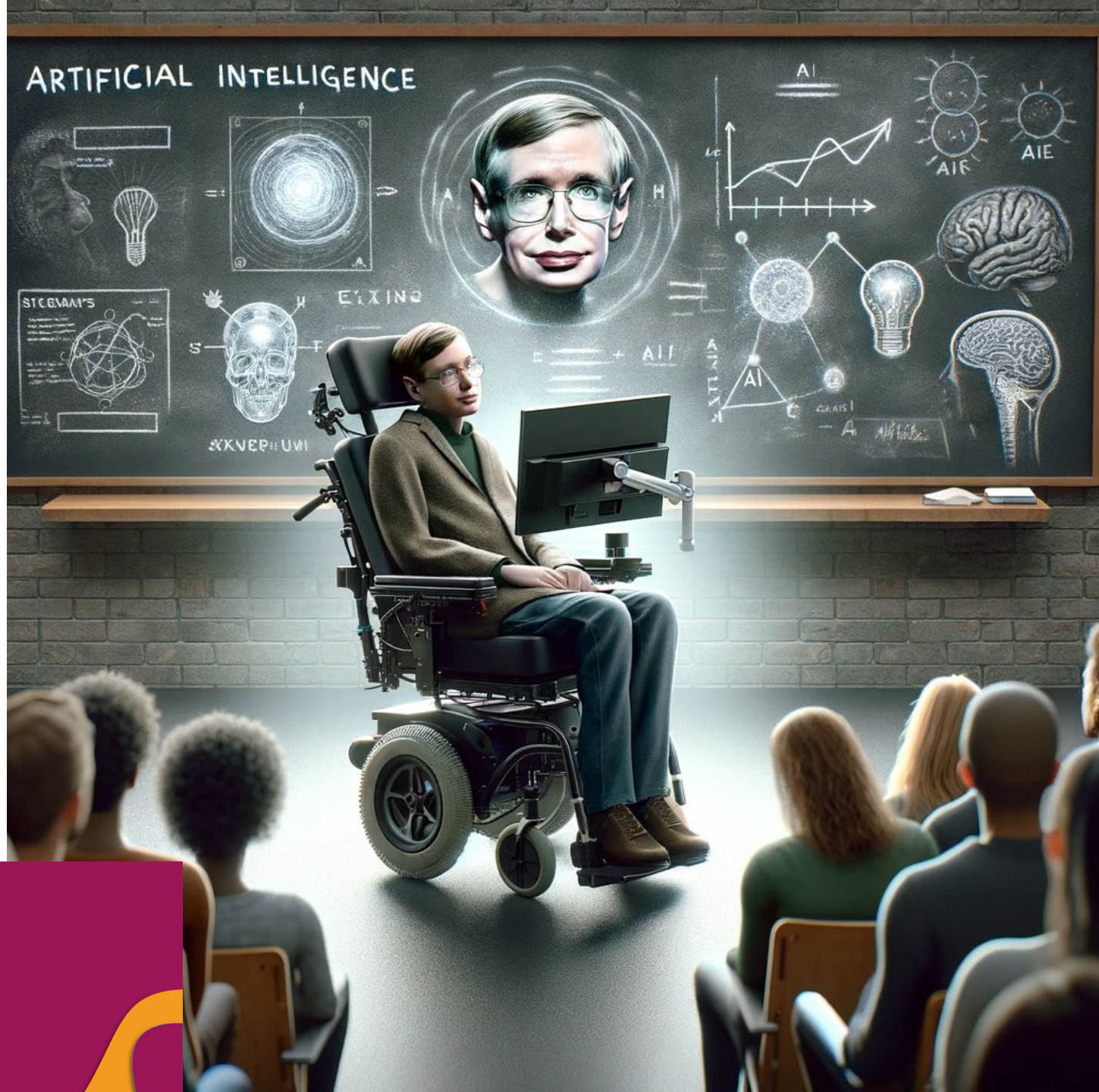
THE HUMAN *POWER*

La brújula ética de la IA: el poder humano en juego

- **Pablo Corona Fraga**

"El éxito en la creación de la inteligencia artificial podrá ser el evento más grande en la historia de la humanidad. Desafortunadamente también sería el último, a menos de que aprendamos cómo evitar los riesgos“

- Stephen

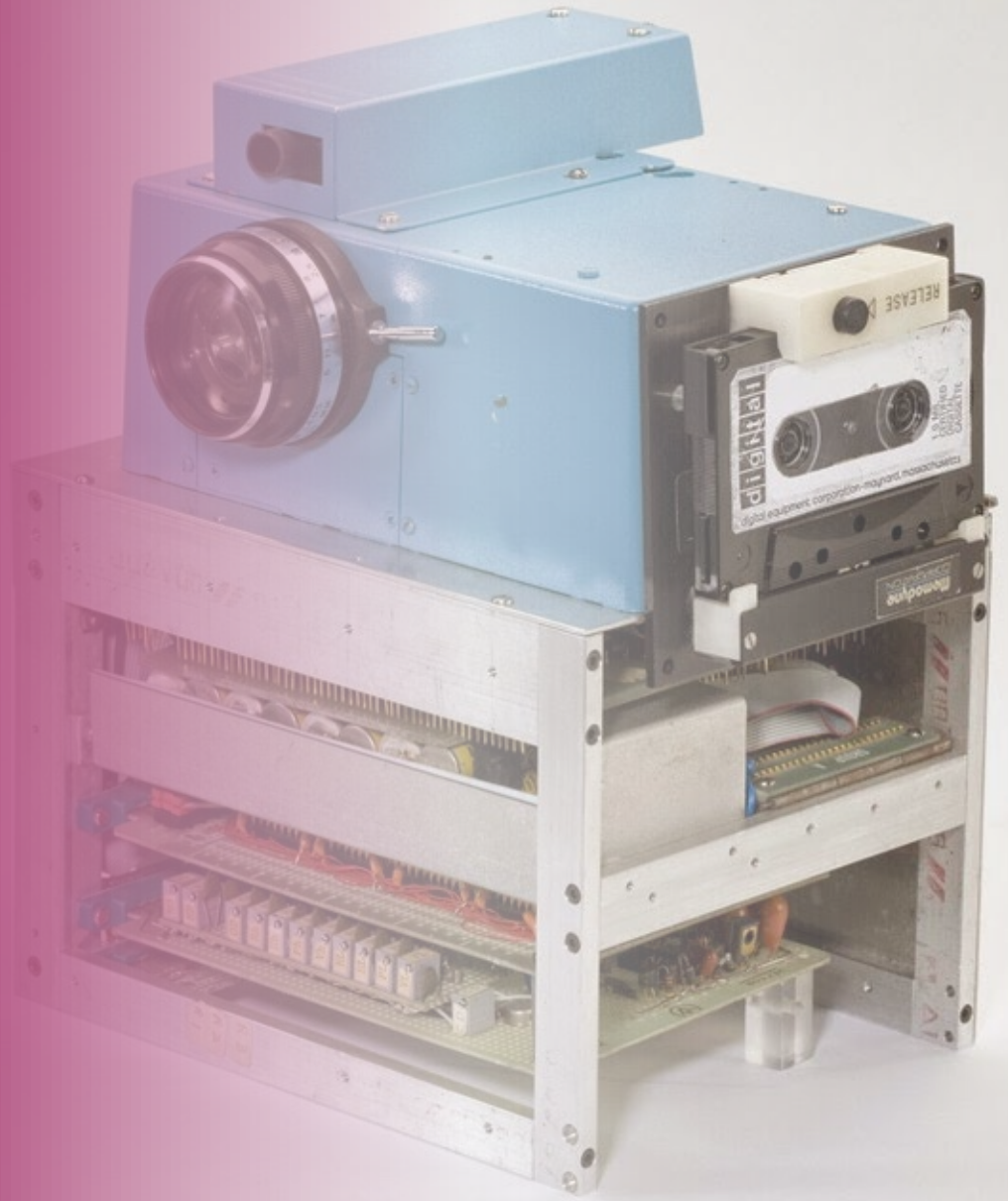


GLOBAL CX FORUM
28º CONGRESO MÉXICO

1876 - Wester Union

Este "teléfono" tiene demasiadas deficiencias para ser considerado seriamente como un medio de comunicación. El dispositivo es inherentemente de ningún valor para nosotros.

Estamos convencidos de que nadie querra ver sus fotos en un televisor. La impresión ha estado con nosotros durante más de 100 años, nadie se queja de las impresiones, son muy baratas, ¿por qué alguien querría ver su foto en un televisor?



- Consideramos que no hay más de 100 hogares que puedan ser clientes de esto…
- Netflix no está en nuestro radar en términos de competencia

El iPhone no es atractivo para gente de negocios porque no es útil para correos electrónicos, no tiene teclado físico.

Exageran al decir que es una amezara para RIM

Portada / Uncategorized /

Forbes Staff
mayo 17, 2018 @ 1:07 pm

Adiós a la legendaria Guía Roji, la compañía está en bancarrota

La empresa no pudo sobrevivir a un mundo cada vez más tecnológico dominado por Google Maps y Waze, a pesar de que por algún tiempo supo adaptarse a los cambios.

¿En
esta
resp
inte
artif

AI will elev

While AI is sti
we know its fu
starring role i
processes. **In**
industry, AI w
digest previou
data and turn
information. A
with the keys
the service yo
(**Artificial Int**
welcome a gu
based on thei
bar, offer a dis
gone previous
to surprise yo
loyalty.

El 'radiólogo'

ChatGPT Will Replace Programmers Within 10 Years

Predicting The End of Manmade Software

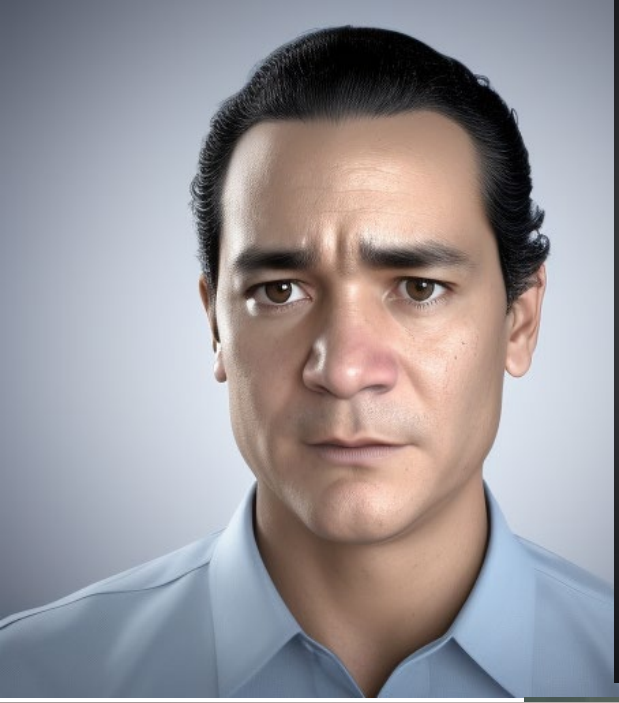
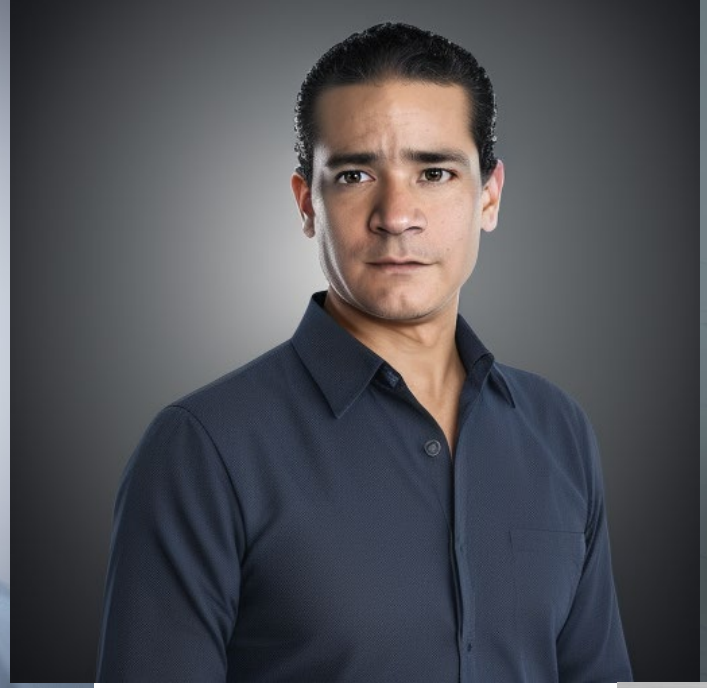
Adam Hughes · Follow

Published in Level Up Coding · 12 min read · Feb 28

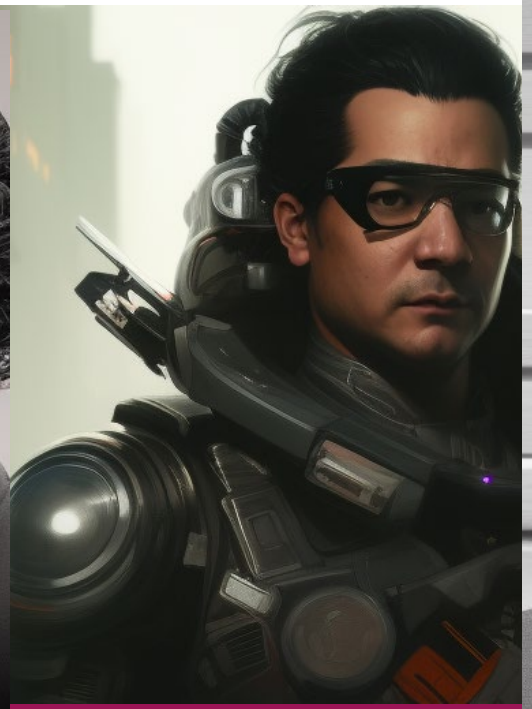
1.3K 92

OpenAI's DALL-E prompted with "photo-realistic 3D robot destroying a computer"

In 2024, 50% of all tasks will be automated.





inteligir Conjugar

Del lat. *intelligĕre.*

1. tr. cult. Entender algo o a alguien. *La facultad de inteligir las acciones humanas.*

computer systems able to perform tasks normally requiring human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages

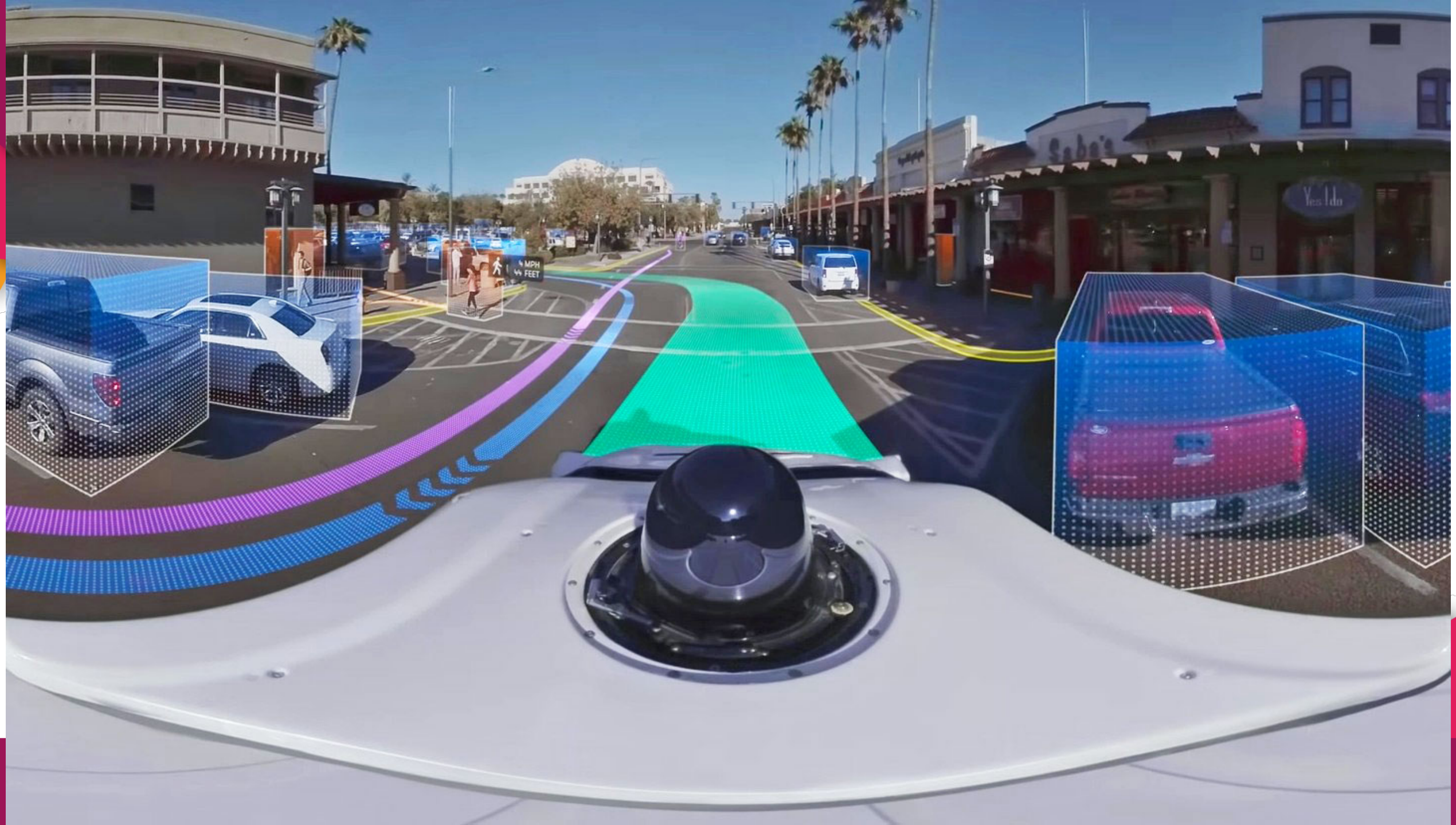
etymonline Buscar Colu

Publicidad

intelligence (n.)

"Tardío 14c., 'la facultad más alta de la mente, capacidad para comprender verdades generales'; c. 1400, 'facultad de comprensión, comprensión', del Antiguo Francés *intelligence* (12c.) y directamente del Latín *intelligentia, intellegentia* 'comprensión, conocimiento, poder de discernir; arte, habilidad, gusto', del *intelligentem* (nominativo *intelligens*) 'discernir, apreciar', participio presente de *intelligere* 'entender, comprender, llegar a conocer', del forma asimilada de *inter* 'entre' (ver **inter-**) + *legere* 'elegir, seleccionar, leer', del raíz PIE ***leg-** (1) 'recoger, reunir', con derivados que significan 'hablar (para 'seleccionar palabras')'."

El significado de "comprensión superior, sagacidad, cualidad de ser inteligente" proviene de principios del siglo XV. El sentido de "información recibida o impartida, noticias" se registra por primera vez a mediados del siglo XV, especialmente "información secreta de espías" (1580s). El significado de "un ser dotado de entendimiento o inteligencia" es de finales del siglo XIV. *Intelligence quotient* se registra por primera vez en 1921 (ver **I.Q.**).


MPH
FEET

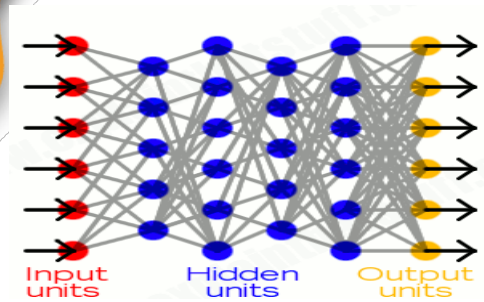
DARPA la clasifica en tres “olas”

Cada ola tiene algunas similitudes, sus propias capacidades y limitaciones.
De las tres, la tercera es la más nueva y poderosa.

Fuente: Defense Advanced Research Projects Agency Perspective on Artificial Intelligence.
John Launchbury


Inteligencia Artificial
...Machine Learning...
... Estadística...
if if if if if if if if

Sistema de AI



Watson

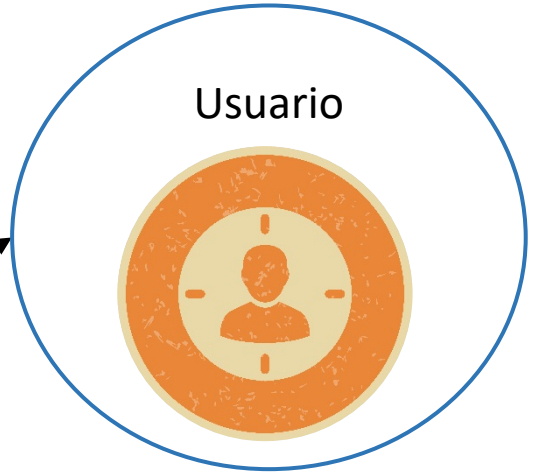
AlphaGo

Sentido de las decisiones

Operaciones

LLMs

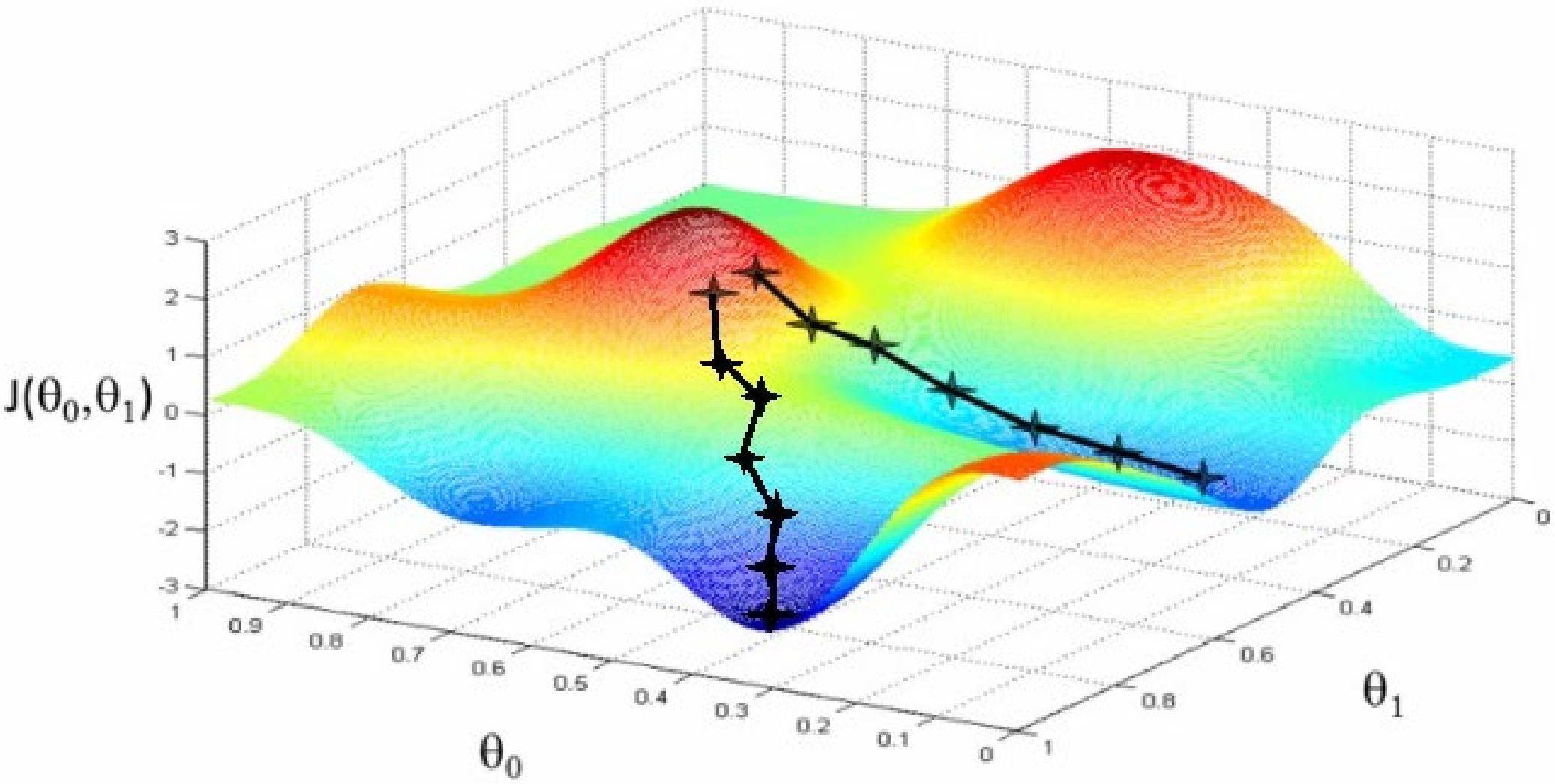
Usuario



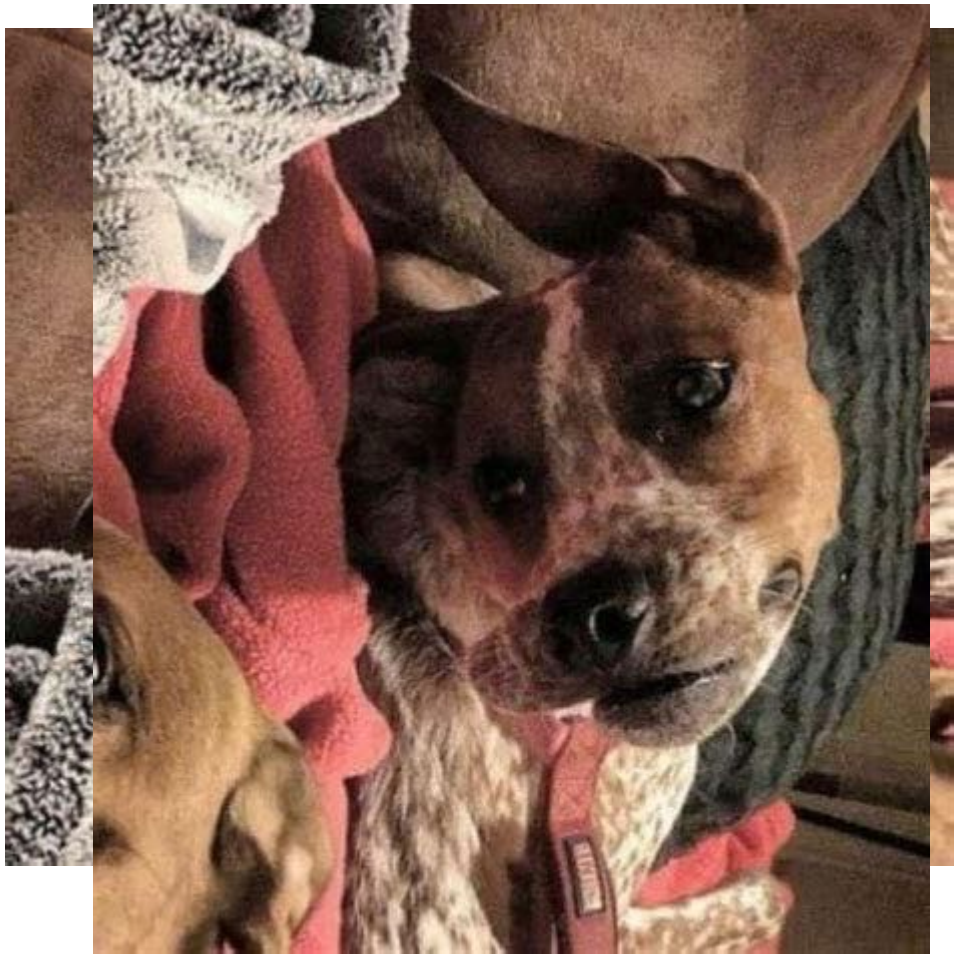
- Estamos entrando en una nueva era de aplicaciones de IA
- El aprendizaje automático es la tecnología central
- Los modelos de aprendizaje automático son opacos, no intuitivos y difíciles de entender para las personas.

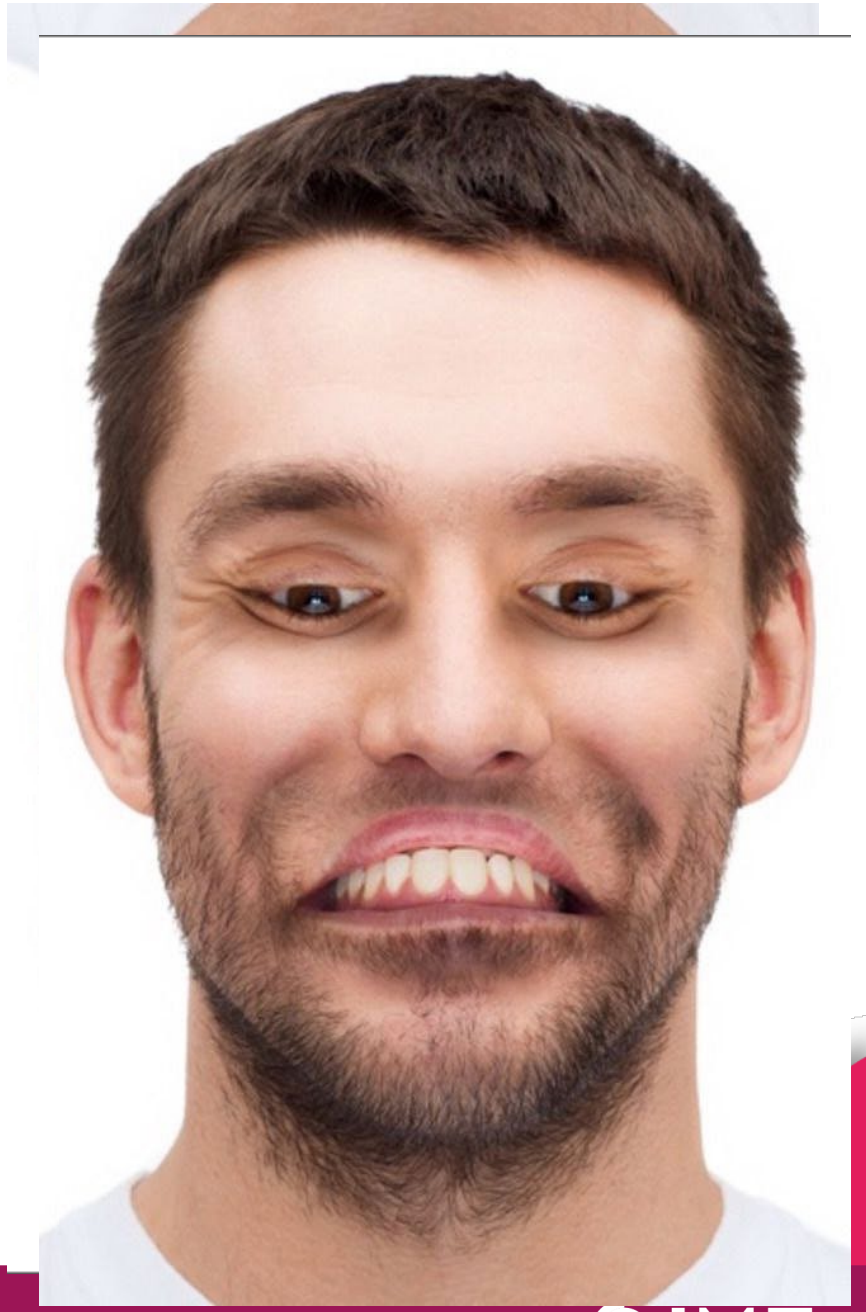
- ¿Por qué hiciste eso?
- ¿Por qué no otra cosa?
- ¿Cuándo tienes éxito?
- ¿Cuándo fallas?
- ¿Cuándo puedo confiar en ti?
- ¿Cómo corrijo un error?

Función de pérdida



¿Cómo opera nuestro cerebro?





Inteligencia artificial general (AGI)

- Razona, usa estrategias, resuelve acertijos y realiza juicios bajo incertidumbre
- Abstrae el conocimiento, incluido el sentido común
- Planeación
- Aprendizaje autónomo
- Comunicarse en lenguaje natural
- Integrar todas estas habilidades hacia objetivos comunes
- Explicar/entender sus decisiones

Riegos

La Inteligencia Artificial ha sido utilizada como una tecnología de invasión a la privacidad
- Michelle Bachelet

No es probable que la IA actual ataque por sí misma

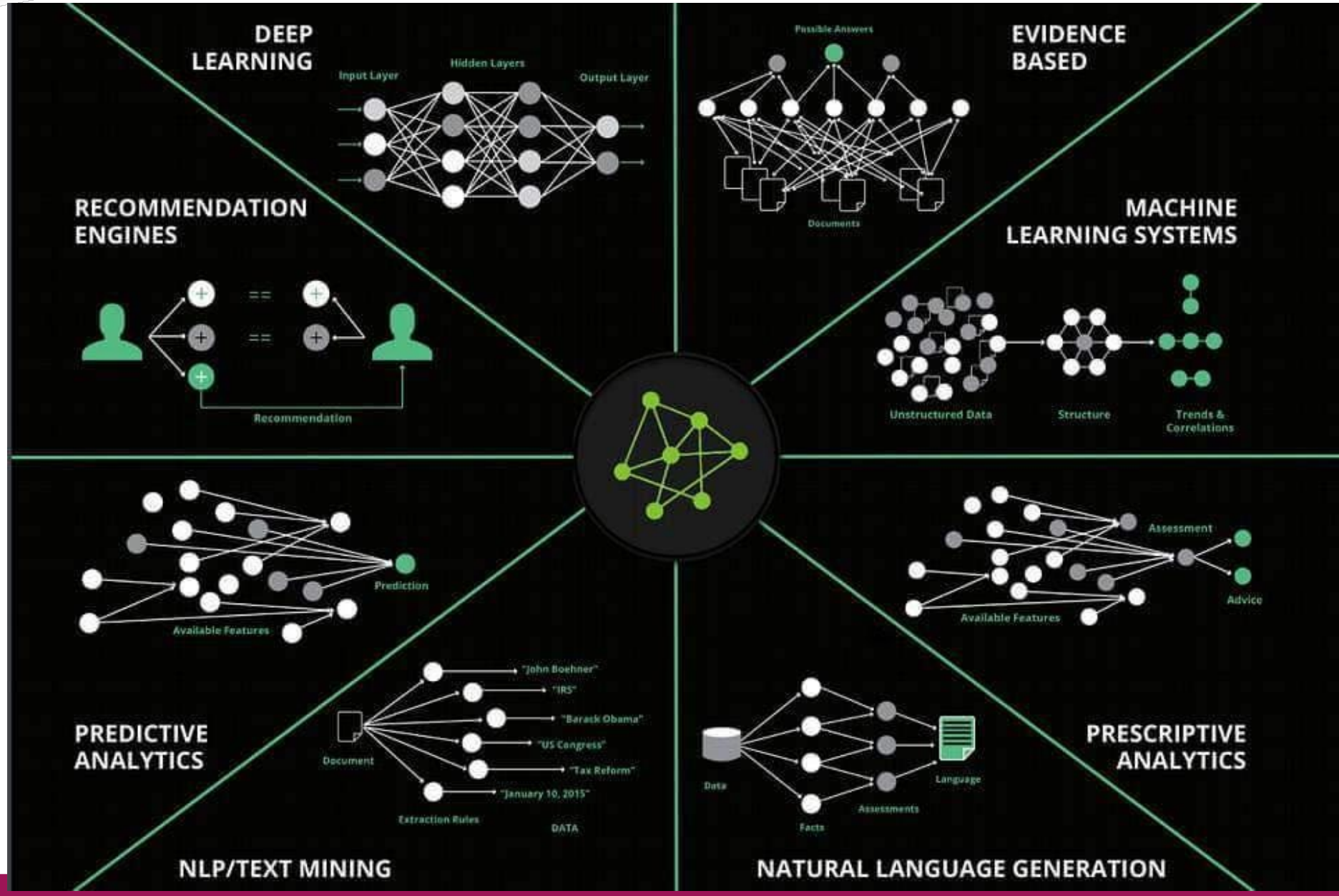
Sin que exista Inteligencia artificial general, la IA no:

- Tomará consciencia de sí misma
- Tratará de escapar ni esconderse en programas o sistemas
- Intentará autorreplicarse (a menos que se lo indiquemos)
- Modificará la forma en que se replica (a menos que se lo indiquemos)
- Puede aplicar lo que aprende a otros dominios

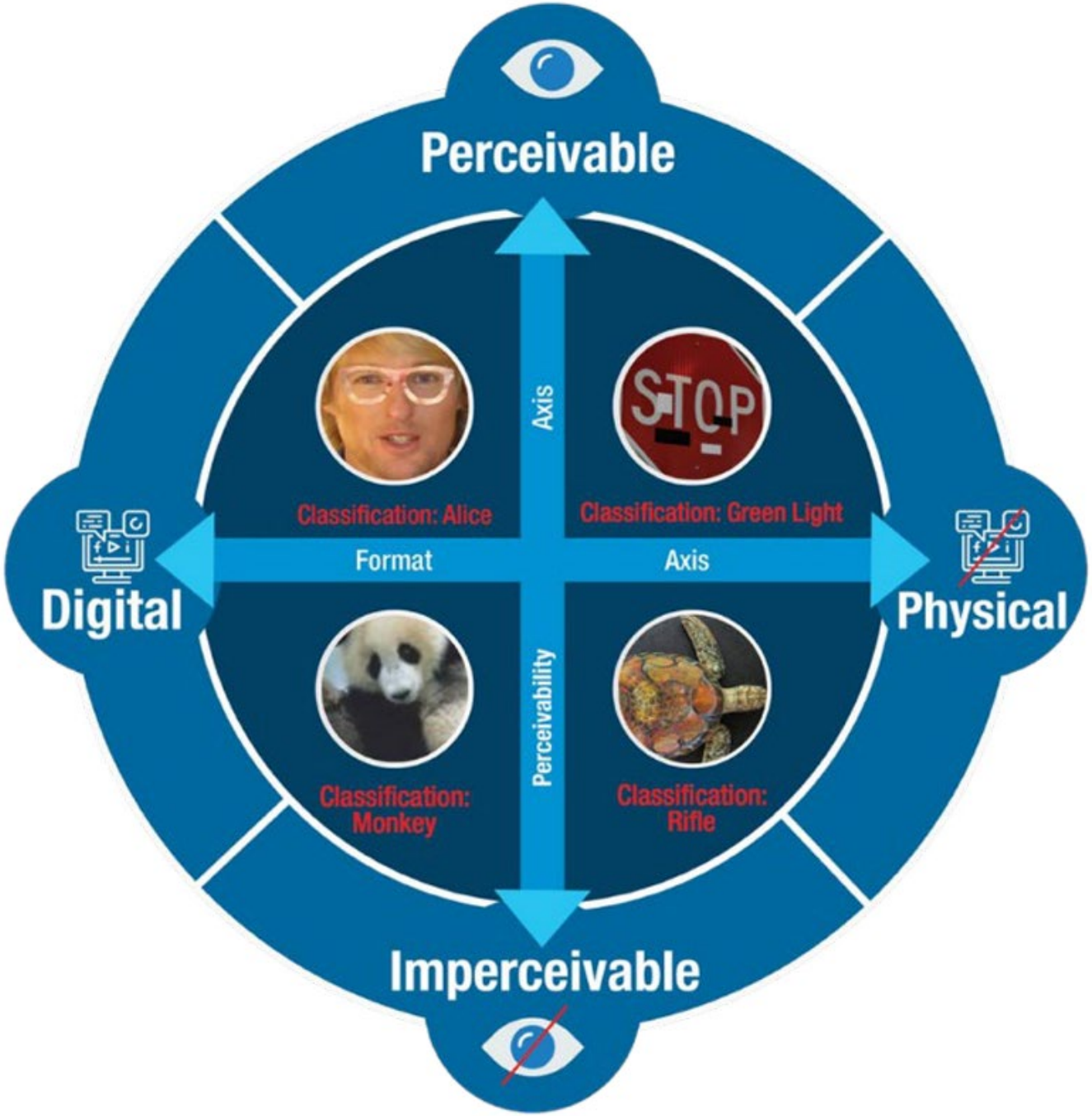
Riesgos

- Recolección Masiva de Datos
- Seguridad de los Datos
- Sesgo y Discriminación
- Falta de Transparencia y Responsabilidad
- Uso Secundario de Datos
- Anonimización y Desidentificación
- Inferencias y Perfilado
- Manipulación y Control
- Cumplimiento Normativo

Ataques a los algoritmos

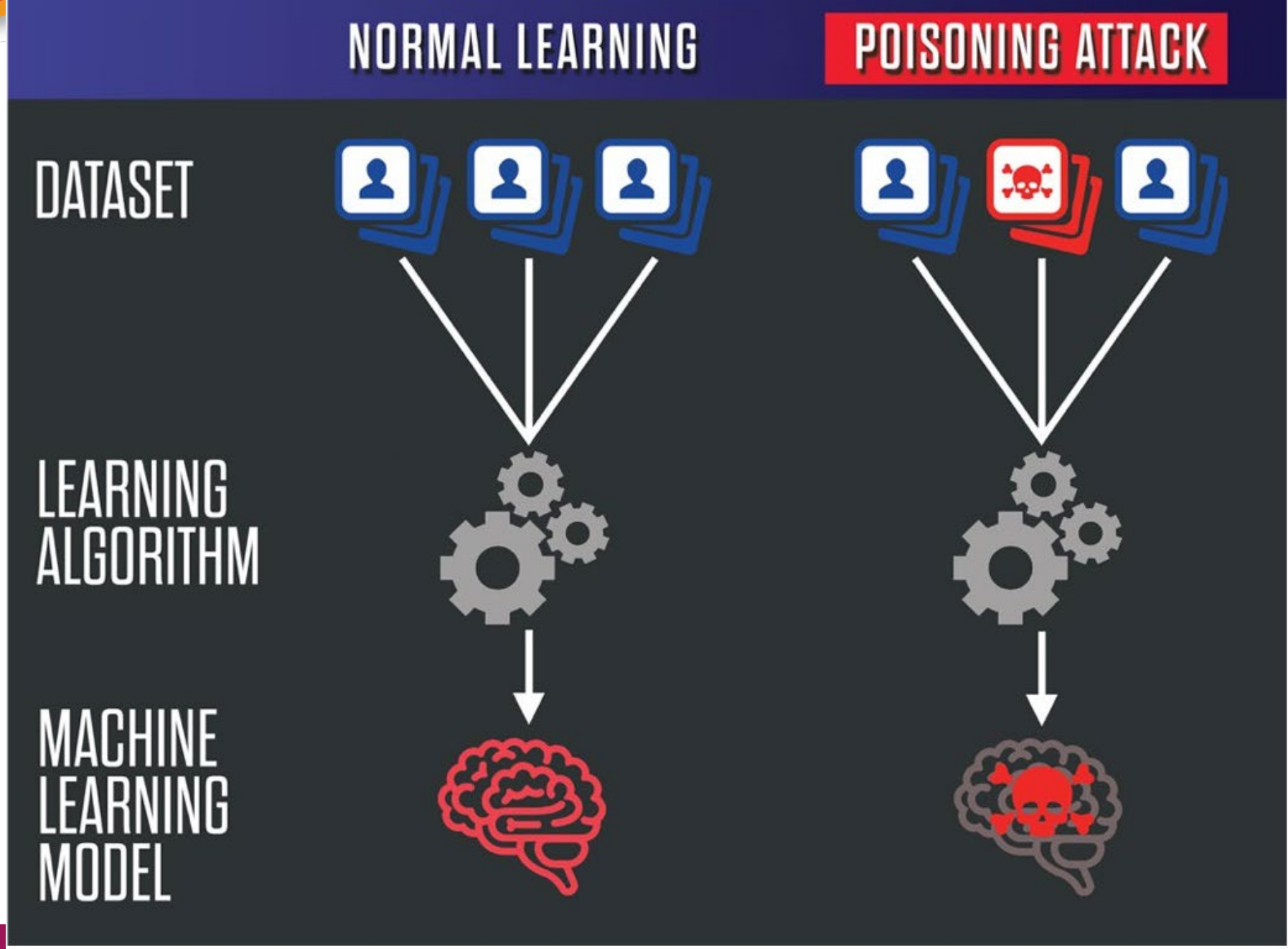


Ataques de entrada



Source: M. Comiter (2019), op.cit., p. 19.

Ataques de envenenamiento del origen de datos



Source: M. Comiter (2019), op.cit., p. 29.

Orígenes de la discriminación y sesgos

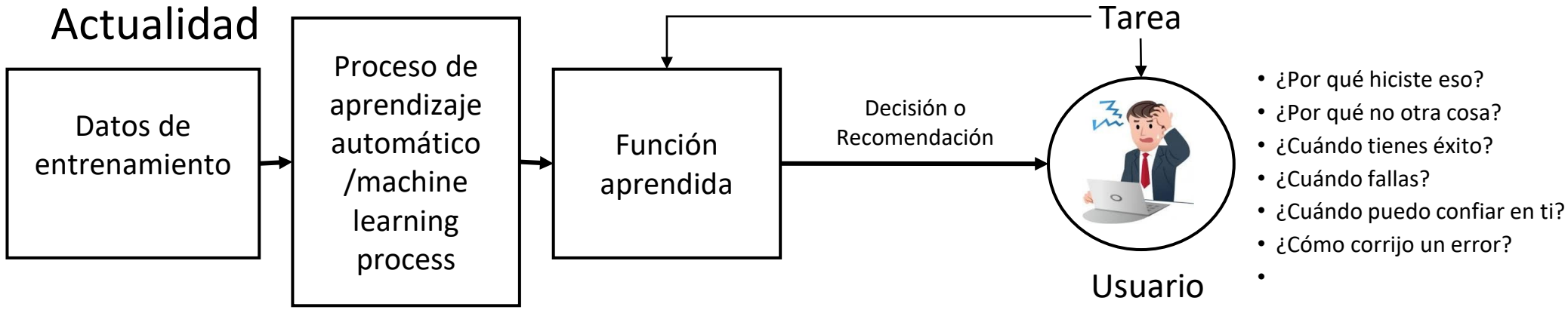
- Discriminación en el mundo real
- Submuestreo
- Sesgo de proxy
- Modelos de grano grueso
- Sesgo de referencia
- Sesgo de función objetivo

Acciones

- Ética
- Centrada en el usuario
- Apegada ala regulación
- Verificación y validación
- Toma de decisions a cargo de una persona

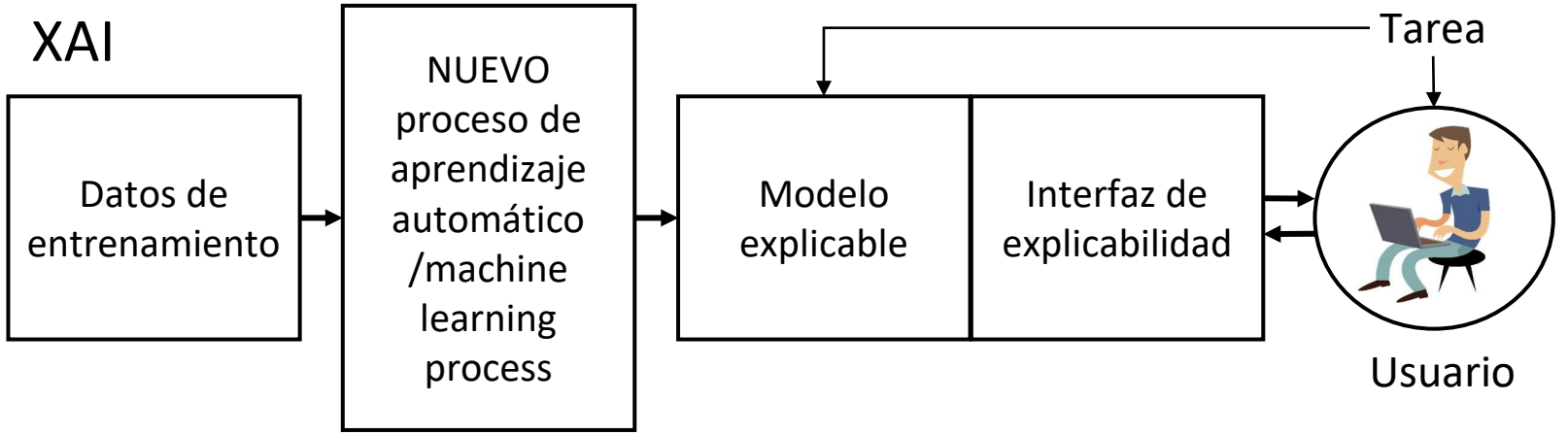
Inteligencia Artificial eXplicativa XAI

Actualidad



- ¿Por qué hiciste eso?
- ¿Por qué no otra cosa?
- ¿Cuándo tienes éxito?
- ¿Cuándo fallas?
- ¿Cuándo puedo confiar en ti?
- ¿Cómo corrijo un error?
-

XAI



- Entiendo por qué
- Entiendo por qué no
- Sé cuándo tienes éxito
- Sé cuándo fallas
- Sé cuándo confiar en ti
- Sé por qué te equivocaste
-

Ética para la IA

IA para el bien

¿Qué podemos hacer con esta poderosa herramienta, para tener impacto positivo en la sociedad?

IA no para el mal

¿Cómo evitar los riesgos éticos de la IA?

Contenido

¿Cuáles son los riesgos que tratamos de evitar?

- IA Explicable
- IA justa y sin sesgos
- Privacidad

Estructura

¿Cómo implementamos la mitigación de estos riesgos?

Construir programas de riesgos éticos de la IA

Fuente: Reid Blackman, Ethical machines

Riesgos en Ai en distintos sectores

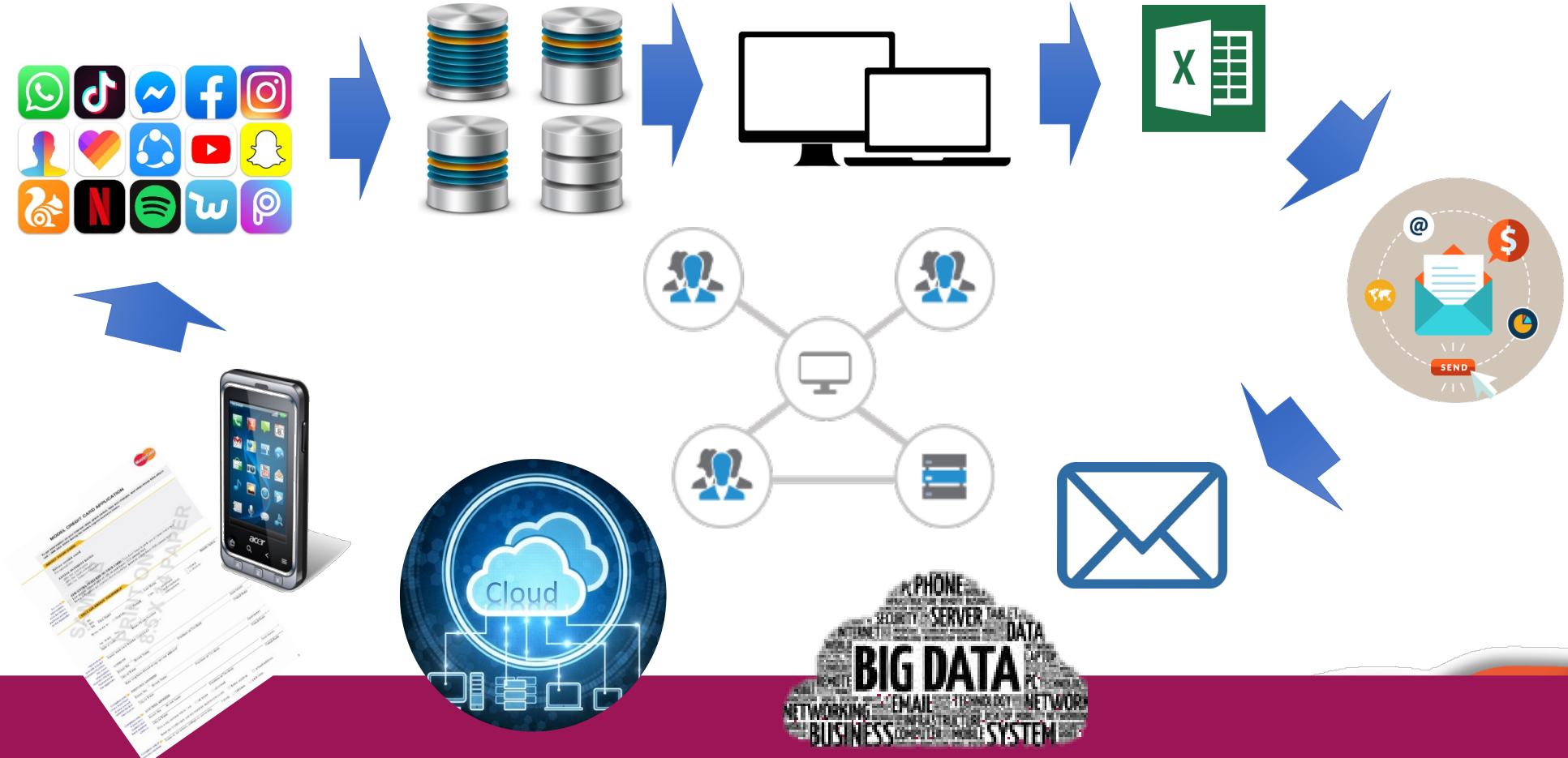
El sector gobierno tiene menos experiencia en el uso de estas herramientas y debe ser más cauto en las aplicaciones que utiliza

Salud tendría elementos para considerarlos desde su origen, por la naturaleza ética de sus servicios, pero no ha sido de los más avanzados así

El sector financiero tiene más avance posiblemente por la sobre regulación a la que están sometidos y a reputación que tienen que cuidar

Dos perspectivas

- privacidad: se trata del control sobre el uso de los datos
- ciberseguridad: se trata del control sobre el flujo y el acceso a los datos



Crear estructura desde el contenido

Sesgo

- métricas de equidad son apropiadas para el caso de uso
- estrategias de mitigación de sesgos evaluada por personas con experiencia en la materia
- identificación y mitigación de los sesgos de los modelos inicia antes del entrenamiento del modelo, idealmente, antes de determinar conjuntos de datos de entrenamiento
- incluir a un abogado al determinar las técnicas de mitigación de sesgos

Explicabilidad

- entender si las explicaciones son aplicables al caso de uso
- incluir personas con perfil ético y legal con experiencia para evaluar la equidad de las reglas
- consultar usuarios finales para determinar si se necesita una explicación y si es aceptable en el contexto, dados sus conocimientos, habilidades y propósito del sistema

Privacidad

- antes de comenzar a recopilar datos para entrenar la IA, determinar qué nivel ético de privacidad es apropiado para el caso de uso
- responsable del tratamiento
- decisiones consultadas con expertos cuando los valores éticos entran en conflicto

Niveles de etica en la privacidad para Ai

Table 4-1

The five ethical levels of privacy

	Level 1 *Blindfolded and handcuffed*	Level 2 *Handcuffed*	Level 3 *Pressured*	Level 4 *Slightly curtailed*	Level 5 *Grateful*
Transparency		✔	✔	✔	✔
Data control			✔	✔	✔
Opt out by default				✔	✔
Full services	✔	✔			✔

Fuente: Reid Blackman, Ethical machines

Niveles de impacto

TABLE 7-1

Ethical risk due diligence framework

	Physical harm	Mental harm	Autonomy	Trust & respect	Relationships & social cohesion	Social justice & fairness	Unintended consequences
Stakeholder A (for instance, a set of individuals)	High	Medium	Medium	Low	High	Low	High
Stakeholder B (a collective, for instance, a country, a community)	Medium	High	Medium	High	High	Low	Low
Stakeholder C	High	Low	High	Low	Low	Medium	High
Stakeholder D	Low	Low	Medium	Low	High	Medium	Low

Table key

- = Low risk
- = Medium risk
- = High risk

Fuente: Reid Blackman, Ethical machines

Experiencia personalizada

Mejora en la Atención al Usuario

Capacitación

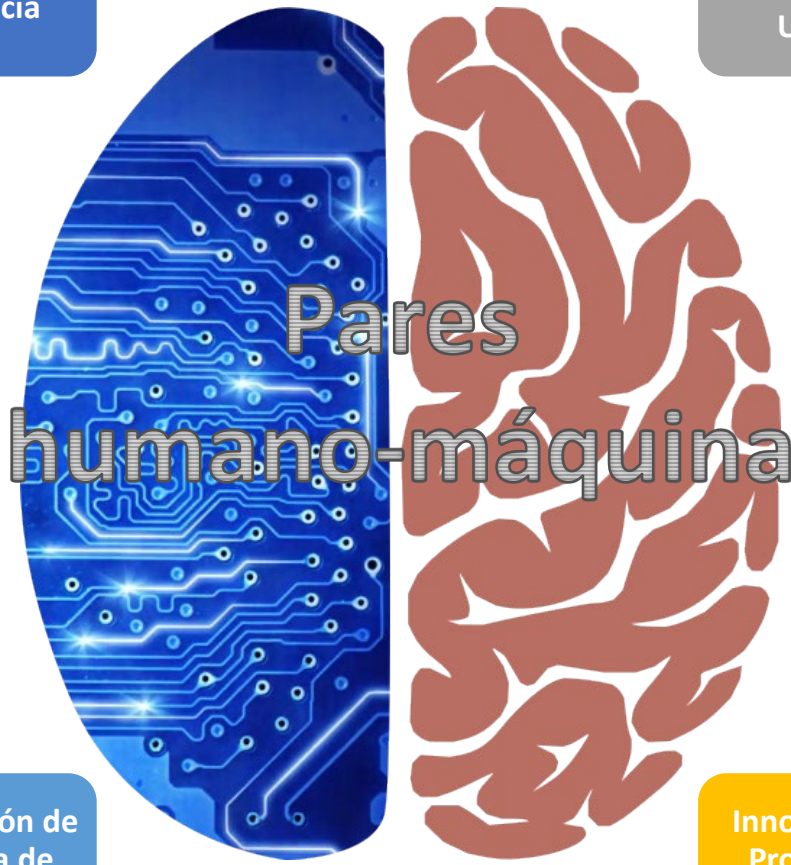
Innovación en Productos y Servicios

Detección de fraudes

Optimización de la Cadena de Suministro

Detección de fallas

Automatización y Eficiencia



Pares humano-máquina

Sistema de Gestión de IA

ISO/IEC 42001

La norma ISO/IEC 42001:2023 proporciona un marco de sistema de gestión de IA certificable dentro del cual se pueden desarrollar productos IA como parte de un ecosistema para asegurar la IA. El objetivo final es ayudar a las empresas y a la sociedad a obtener el máximo beneficio de la IA y asegurar a las partes interesadas que sus sistemas se están desarrollando de manera responsable.

- Mejorar la calidad, la seguridad, la trazabilidad, la transparencia y la fiabilidad de las aplicaciones de IA, así como resolver algunos retos de implementación;
- Generar una mayor confianza en los sistemas de IA;
- Reducir los costos del desarrollo de IA;
- Mantener el cumplimiento normativo;
- Satisfacer las expectativas de los clientes, el personal y otras partes interesadas en torno al uso ético y responsable de la IA; y
- Mejorar la eficiencia y la gestión de riesgos.

La major herramienta en manos no adecuadas puede causar daño

Gracias

Pablo Corona Fraga

pcoronaf@nyce.org.mx

Twitter: @pcoronaf