Cómo aprenden la máquinas

Te proponemos una lectura sobre cómo aprenden la máquinas que no no es evaluable, pero que si creemos que deberías conocer en caso de que no hayas hecho la formación sobre IA. Este texto está extraído del curso " IA en el Diseño Curricular" del Itinerario de IA que se oferta en Aularagon

Aprender: una palabra común para procesos muy distintos

Hablar de aprendizaje parece, a primera vista, algo sencillo. Aprendemos desde que nacemos: a hablar, a caminar, a resolver problemas, a convivir con otros. En el aula, aprender significa comprender un contenido, practicarlo, equivocarse, corregir y, poco a poco, hacerlo propio. Cuando trasladamos esta palabra al ámbito de la Inteligencia Artificial, el término es el mismo, pero el proceso es muy distinto.

Cuando decimos que una máquina “aprende”, no lo hace porque entienda lo que está haciendo ni porque tenga una intención de mejorar. Aprende porque **ajusta su comportamiento a partir de ejemplos**, del mismo modo que un alumno mejora cuando recibe práctica y retroalimentación, aunque en el caso de la máquina todo ocurre a nivel numérico y estadístico. La IA no trabaja con ideas, trabaja con datos; no interpreta significados, detecta patrones.

Un ejemplo sencillo ayuda a aclararlo. Si un alumno aprende a reconocer tipos de texto leyendo muchos ejemplos y recibiendo correcciones del profesor, la máquina aprende de forma parecida: analiza miles de textos, compara sus predicciones con el resultado esperado y va ajustando sus respuestas. La diferencia es que el alumno comprende por qué algo es un texto narrativo o argumentativo, mientras que la máquina solo aprende qué combinaciones de palabras suelen aparecer en cada caso.

A partir de esta idea general, no todas las máquinas aprenden del mismo modo. Algunas lo hacen con ejemplos corregidos, otras explorando datos sin respuestas previas y otras aprendiendo a base de ensayo y error, según las consecuencias de sus acciones. Curiosamente, estas formas de aprendizaje tienen muchos paralelismos con la educación humana: aprendizaje guiado, aprendizaje por descubrimiento y aprendizaje por experiencia.

Comprender **cómo aprenden las máquinas** no solo nos permite usar mejor la Inteligencia Artificial, sino que también nos invita a reflexionar sobre **cómo enseñamos y cómo aprenden nuestros alumnos y alumnas**. Con esta base, podemos ahora describir los principales tipos de

Copyright 2026

aprendizaje que utilizan los sistemas de IA y entender en qué contextos se aplica cada uno.

Tipos de aprendizaje

Cuando hablamos de *aprendizaje* en Inteligencia Artificial conviene detenerse un momento y ajustar expectativas, especialmente en el ámbito educativo. Las máquinas **no aprenden como aprendemos las personas**. No sienten curiosidad, no tienen intención de mejorar ni conciencia de lo que hacen. Y, sin embargo, aprenden.

En IA, aprender significa algo muy concreto: **ajustar un modelo matemático a partir de datos** para que, con el tiempo, cometa menos errores en una tarea concreta. No hay comprensión del contenido, sino detección de regularidades numéricas. Donde una persona entiende un concepto, una máquina identifica patrones estadísticos. Esta diferencia es esencial para comprender tanto el potencial como los límites de la IA en educación.

El aprendizaje automático como núcleo de la IA moderna

El corazón de la Inteligencia Artificial actual es el **aprendizaje automático o Machine Learning (ML)**. A diferencia de la programación clásica —en la que el ser humano escribe reglas explícitas—, en el aprendizaje automático **no se le indica a la máquina cómo resolver el problema**, sino que se le muestran ejemplos para que descubra por sí misma qué funciona mejor.

Un paralelismo muy útil para el profesorado es el siguiente: programar un sistema tradicional se parece a dar una receta paso a paso; entrenar un sistema de IA se parece más a poner muchos ejemplos corregidos y dejar que el sistema extraiga conclusiones. El resultado suele ser más flexible y potente, aunque también más difícil de explicar.

A partir de esta idea general, podemos distinguir **varios tipos de aprendizaje**, según cómo se presentan los datos y qué tipo de información recibe el sistema durante su entrenamiento.

Aprendizaje supervisado: aprender con corrección

El **aprendizaje supervisado** es probablemente el tipo de aprendizaje automático que resulta más fácil de entender para un profesor, porque su lógica se parece bastante a la forma tradicional en que enseñamos en el aula. En este enfoque, el sistema aprende a partir de ejemplos en los que ya conocemos la respuesta correcta. Dicho de otra manera, el modelo recibe datos de entrada junto con la solución esperada, y a partir de muchos ejemplos aprende la relación entre ambos para poder hacer predicciones sobre casos nuevos.

Podemos imaginarlo como un proceso muy parecido al de **aprender con ejercicios resueltos o corregidos**. El sistema analiza muchos ejemplos y, cuando comete un error, ajusta sus parámetros internos para mejorar la próxima vez. Con el tiempo, ese proceso le permite detectar patrones en

los datos y aplicar lo aprendido a situaciones que no había visto antes. En términos técnicos, el objetivo es que el modelo aprenda la relación entre unas variables de entrada y una salida conocida, de manera que pueda predecir esa salida en datos nuevos.

Dentro del aprendizaje supervisado suelen distinguirse dos grandes tipos de tareas. La primera es la **clasificación**, que consiste en asignar una categoría a un elemento. Por ejemplo, decidir si un correo es spam o no, o clasificar una imagen como perro o gato. La segunda es la **regresión**, que se utiliza cuando el objetivo es predecir un valor numérico continuo, como estimar la temperatura, el precio de una vivienda o las ventas futuras de un producto.

Desde el punto de vista educativo, este tipo de aprendizaje tiene aplicaciones cada vez más interesantes. En investigación educativa se utilizan algoritmos de aprendizaje supervisado para **analizar datos académicos y predecir el rendimiento de los estudiantes**, lo que permite detectar con antelación posibles dificultades o riesgo de abandono escolar. También se emplea en sistemas de evaluación automática, en plataformas de aprendizaje adaptativo o en herramientas que analizan grandes conjuntos de datos educativos para comprender mejor cómo progresa el alumnado.

La principal ventaja del aprendizaje supervisado es su **precisión y capacidad predictiva**, especialmente cuando se dispone de buenos datos. Sin embargo, su mayor debilidad es precisamente esa dependencia de los datos. El modelo solo puede aprender a partir de los ejemplos que recibe. Si los datos están mal etiquetados, contienen errores o reflejan sesgos, el sistema aprenderá esos mismos problemas y los reproducirá en sus predicciones.

Para el profesorado de educación secundaria, comprender esta idea es especialmente útil porque permite explicar al alumnado uno de los principios fundamentales de la inteligencia artificial actual: **las máquinas no aprenden por comprensión, sino por comparación de ejemplos**. En cierto sentido, el aprendizaje supervisado se parece mucho a un estudiante que practica con muchos ejercicios resueltos hasta que empieza a reconocer patrones y a resolver problemas nuevos por sí mismo.

https://www.ceupe.com/images/easyblog_articles/3266/aprendizaje-supervisado-ejemplo.png
Esquema de aprendizaje supervisado en el que se presentan ejemplos categorizados o etiquetados

Aprendizaje no supervisado: descubrir sin que nadie explique

El **aprendizaje no supervisado** es un tipo de aprendizaje automático que funciona de una forma muy distinta al aprendizaje supervisado. En este caso, el sistema no recibe ejemplos con la respuesta correcta ni etiquetas que indiquen qué debe aprender. En lugar de eso, el algoritmo analiza grandes cantidades de datos sin clasificar e intenta descubrir por sí mismo **patrones, estructuras o relaciones internas** dentro de esa información.

Dicho de una manera sencilla, la máquina observa los datos y busca similitudes entre ellos. Si encuentra elementos que se parecen entre sí, los agrupa automáticamente en conjuntos o **clusters**. Este proceso se conoce como agrupamiento o *clustering* y es una de las técnicas más comunes del aprendizaje no supervisado. La idea es que los datos que pertenecen a un mismo grupo compartan características similares, mientras que los que pertenecen a grupos distintos sean más diferentes entre sí.

Para un profesor de educación secundaria, puede resultar útil imaginar este proceso como cuando un docente observa una clase durante varias semanas y empieza a identificar patrones en el comportamiento o en la forma de trabajar de los estudiantes. Sin necesidad de que nadie lo indique explícitamente, puede detectar grupos de alumnado que tienen hábitos de estudio similares, ritmos de aprendizaje parecidos o formas de participación semejantes. El aprendizaje no supervisado funciona de manera parecida: el sistema observa los datos y descubre regularidades que no estaban previamente definidas.

En el ámbito educativo, este tipo de aprendizaje no suele utilizarse tanto para calificar o evaluar directamente al alumnado, sino para **comprender mejor fenómenos complejos dentro de los datos educativos**. Por ejemplo, puede servir para identificar perfiles de estudiantes con comportamientos similares en una plataforma digital de aprendizaje, detectar patrones de uso de recursos educativos o descubrir relaciones entre variables académicas que no se habían considerado previamente.

Sin embargo, hay una idea fundamental que conviene subrayar desde una perspectiva pedagógica. Aunque el algoritmo pueda encontrar agrupaciones o patrones en los datos, **no entiende el significado educativo de esos resultados**. El sistema puede detectar que ciertos estudiantes presentan comportamientos parecidos, pero no puede interpretar por qué ocurre eso ni qué implicaciones pedagógicas tiene. Esa interpretación siempre corresponde al profesorado.

Por esta razón, el aprendizaje no supervisado no sustituye el papel del docente. Más bien funciona como una herramienta que puede ofrecer nuevas perspectivas sobre grandes volúmenes de datos. En lugar de reemplazar la mirada pedagógica del profesor, amplía su capacidad para analizar información compleja y detectar patrones que de otro modo podrían pasar desapercibidos. En última instancia, la máquina puede identificar estructuras en los datos, pero **el sentido educativo de esas estructuras lo aporta siempre la interpretación humana**.

https://uk.mathworks.com/discovery/unsupervised-learning/_jcr_content/thumbnail.adapt.1200.med
Esquema aprendizaje no supervisado en el que se forman agrupaciones de datos sin etiquetar

https://codideep.com/img/blogpost/imagenportada/20250605000001.p

Aprendizaje por refuerzo

El **aprendizaje por refuerzo** es un tipo de aprendizaje automático que se basa en una idea muy sencilla y bastante intuitiva: **aprender actuando y observando las consecuencias**. En lugar de recibir ejemplos con la respuesta correcta —como ocurre en el aprendizaje supervisado— el sistema interactúa con un entorno, toma decisiones y recibe una señal que le indica si esa acción ha sido buena o mala. Esa señal suele llamarse **recompensa o penalización**, y el objetivo del sistema es aprender a actuar de manera que, con el tiempo, consiga la mayor recompensa posible.

Podemos imaginar este proceso de forma parecida a cómo aprenden muchas personas o animales: probando cosas, observando el resultado y ajustando su comportamiento. Si una acción produce un resultado positivo, el sistema tenderá a repetirla; si produce un resultado negativo, intentará evitarla en el futuro. Con el tiempo, mediante miles o millones de intentos, el modelo aprende qué decisiones son más adecuadas en cada situación.

En este tipo de aprendizaje suelen aparecer tres elementos básicos. Por un lado está el **agente**, que es el sistema que toma decisiones. Luego está el **entorno**, que es el mundo o situación donde actúa el agente. Y finalmente está la **recompensa**, que es la señal que indica si la acción tomada ha sido favorable o no. El agente va probando distintas acciones y aprende gradualmente qué estrategias le permiten maximizar la recompensa a largo plazo.

Uno de los ejemplos más famosos de aprendizaje por refuerzo es **AlphaGo**, el sistema desarrollado por DeepMind para jugar al juego de mesa Go. Este programa comenzó aprendiendo a partir de partidas humanas, pero después siguió mejorando jugando millones de partidas contra sí mismo. En ese proceso fue ajustando su comportamiento en función de si ganaba o perdía, lo que le permitió descubrir estrategias nuevas que incluso sorprendieron a jugadores profesionales.

Este ejemplo es interesante porque muestra que el aprendizaje por refuerzo no se limita a imitar comportamientos existentes. Cuando el sistema tiene libertad para explorar muchas posibilidades, puede encontrar soluciones eficaces que los humanos no habían considerado.

En el ámbito educativo, este enfoque resulta especialmente interesante cuando se trabaja con **entornos interactivos o simulaciones**. Por ejemplo, en simuladores de aprendizaje, videojuegos educativos o laboratorios virtuales, un agente puede aprender progresivamente cómo actuar dentro del entorno. También se investiga su uso en sistemas educativos inteligentes que aprenden **cuándo intervenir para ayudar al estudiante y cuándo dejar que explore por sí mismo**.

En definitiva, el aprendizaje por refuerzo introduce una idea muy poderosa dentro del Machine Learning: en lugar de aprender solo a partir de ejemplos, los sistemas pueden **aprender actuando, experimentando y ajustando su comportamiento en función de las**

Copyright 2026

consecuencias. Esto abre la puerta a aplicaciones muy interesantes en ámbitos como la robótica, los videojuegos, la optimización de procesos y también en entornos educativos basados en simulación y experimentación.

El resultado: los modelos

Cuando entrenamos un sistema de **Machine Learning**, el objetivo final es construir un **modelo**. Ese modelo es, en esencia, una función matemática o estadística que ha aprendido a relacionar ciertos datos de entrada con un resultado. Una vez entrenado, ese modelo puede utilizar lo aprendido para analizar datos nuevos y producir una respuesta. Esa respuesta suele tomar dos formas principales: **clasificar algo en una categoría o predecir un valor**.

En el primer caso hablamos de **modelos de clasificación**. Su objetivo es asignar un elemento a una categoría concreta. El modelo analiza las características de un dato y decide a qué grupo pertenece. Este tipo de modelos se utiliza, por ejemplo, para determinar si un correo es spam o no, para identificar si una imagen contiene un gato o un perro, o para clasificar opiniones en positivas o negativas. En este tipo de problemas el resultado no es un número continuo, sino una **etiqueta o clase** previamente definida.

En el segundo caso hablamos de **modelos de predicción o regresión**. Aquí el objetivo no es asignar una categoría, sino estimar un valor numérico. El modelo intenta aprender la relación entre distintas variables y producir una predicción. Por ejemplo, puede estimar el precio de una vivienda a partir de su tamaño y ubicación, predecir la temperatura del día siguiente o calcular cuántos productos se venderán la próxima semana. En estos casos el resultado es una **cantidad continua**, no una categoría.

Ambos tipos de modelos se entrenan normalmente con **datos históricos**. El sistema analiza muchos ejemplos del pasado, detecta patrones y aprende relaciones entre variables. Una vez finalizado el entrenamiento, ese conocimiento se utiliza para analizar situaciones nuevas. Por ejemplo, si un modelo ha aprendido cómo se comportan ciertos datos financieros, podrá estimar tendencias futuras; si ha aprendido a distinguir imágenes de distintos objetos, podrá identificar nuevas imágenes que nunca había visto.

En la práctica existen muchos algoritmos diferentes que permiten construir modelos de **Machine Learning**, como los árboles de decisión, las regresiones estadísticas o los métodos de vecinos más cercanos. Sin embargo, en las últimas décadas ha ganado un protagonismo especial un tipo concreto de modelos: las **redes neuronales profundas**, que constituyen el núcleo del **Deep Learning**.

https://codificandobits.com/media/red-neuronal-profunda-1.png

Copyright 2026

Esquema de funcionamiento de una red neuronal profunda, formada por muchas capas que procesan cada aspecto de la imagen o de la información con la que se entrena

El Deep Learning es un subcampo del Machine Learning que utiliza redes neuronales con muchas capas capaces de aprender representaciones cada vez más complejas de los datos. Estas redes permiten detectar patrones muy sofisticados en información difícil de describir mediante reglas simples, como imágenes, sonido, vídeo o lenguaje natural. Gracias a este enfoque, los sistemas pueden aprender directamente a partir de grandes volúmenes de datos sin necesidad de programar manualmente las características que deben analizar.

Este desarrollo de redes profundas ha sido uno de los factores clave que ha impulsado los avances recientes de la inteligencia artificial. Dentro de este contexto apareció una arquitectura especialmente influyente: los **transformers**, introducidos en 2017 en el artículo científico *Attention Is All You Need*. Esta arquitectura se basa en mecanismos de atención que permiten al modelo analizar relaciones entre distintas partes de un texto o de una secuencia de datos de manera muy eficiente.

Los transformers han supuesto un cambio importante porque permiten entrenar modelos mucho más grandes y manejar contextos más amplios que las arquitecturas anteriores. Gracias a ello se han convertido en la base tecnológica de muchos sistemas actuales de inteligencia artificial, especialmente de los **modelos de lenguaje de gran tamaño (LLM)** y de gran parte de las aplicaciones de **IA generativa** que hoy conocemos, como los asistentes conversacionales, los generadores de texto o los sistemas que crean imágenes y otros contenidos.

En definitiva, aunque el Machine Learning incluye muchos algoritmos diferentes, el desarrollo del **Deep Learning y de las arquitecturas basadas en transformers** ha sido el elemento que ha permitido dar el salto hacia los sistemas actuales capaces de generar texto, analizar lenguaje natural o crear contenido digital de forma cada vez más sofisticada.

Desde una perspectiva educativa, lo interesante es entender que estos modelos no “piensan” ni comprenden los fenómenos que analizan. Lo que hacen es **aprender relaciones estadísticas entre datos**. Si esas relaciones son suficientemente claras y los datos son de calidad, el modelo puede llegar a producir resultados muy útiles. Pero su funcionamiento sigue siendo, en última instancia, el de un sistema que aprende patrones para **clasificar o predecir**, no el de una inteligencia que entiende realmente el mundo.

https://upload.wikimedia.org/wikipedia/commons/7/78/Esquema_aprendizaje.jpg
Esquema de funcionamiento de los modelos automáticos que predicen un valor (supervisado) o un grupo (no supervisado)

Copyright 2026

Cómo aprenden las máquinas y los humanos: el paralelismo educativo

Llegados a este punto, resulta casi inevitable comparar el aprendizaje de las máquinas con el aprendizaje humano. Aunque los mecanismos sean muy diferentes, el paralelismo resulta muy útil desde el punto de vista educativo, porque permite explicar la inteligencia artificial con conceptos que el profesorado conoce bien.

Gran parte del aprendizaje humano se parece al **aprendizaje supervisado**. En el aula, el alumnado aprende a partir de explicaciones, ejemplos y correcciones del profesorado. Cuando un estudiante resuelve un ejercicio y se equivoca, ese error no es simplemente un fallo, sino una señal que permite ajustar el razonamiento y mejorar en el siguiente intento. Algo muy parecido ocurre en muchos sistemas de aprendizaje automático: el modelo recibe ejemplos con la respuesta correcta, compara su predicción con el resultado esperado y ajusta sus parámetros internos para mejorar.

Pero las personas también aprendemos de otras formas. A menudo descubrimos patrones sin que nadie nos los explique directamente. Observamos cómo funcionan las cosas, detectamos regularidades y construimos significado a partir de la experiencia. Este tipo de aprendizaje recuerda al **aprendizaje no supervisado**, en el que los algoritmos analizan datos sin respuestas previas y buscan agrupaciones, estructuras o relaciones internas.

Existe además un tipo de aprendizaje muy habitual en educación que podríamos llamar **semisupervisado**. Ocurre cuando el alumnado recibe algunas orientaciones iniciales, pero tiene que investigar, experimentar o resolver problemas abiertos. Es lo que sucede en muchos proyectos, investigaciones o situaciones de aprendizaje donde no hay una única solución correcta. Curiosamente, este enfoque también está ganando importancia en inteligencia artificial, donde los sistemas combinan pequeñas cantidades de datos etiquetados con grandes volúmenes de datos sin etiquetar.

Por último, también aprendemos mediante **refuerzo**. Las consecuencias de nuestras acciones influyen en cómo actuamos en el futuro. El éxito, el reconocimiento, la frustración o la satisfacción son señales que guían nuestro comportamiento. En el aprendizaje por refuerzo de las máquinas ocurre algo parecido: los algoritmos ajustan sus decisiones en función de recompensas o penalizaciones. Sin embargo, existe una diferencia fundamental. En las personas, esas recompensas están cargadas de significado emocional, social y cultural.

En el fondo, las máquinas y las personas aprenden de maneras que, en apariencia, pueden parecer similares, pero que tienen una diferencia esencial. Las máquinas optimizan funciones matemáticas y buscan patrones estadísticos en los datos. Las personas, en cambio, aprendemos construyendo

sentido: interpretamos, relacionamos ideas, damos significado a lo que aprendemos y lo conectamos con nuestra experiencia.

Comprender cómo aprenden las máquinas es importante para poder utilizarlas correctamente, pero también es necesario entender sus límites. Uno de los más conocidos es el problema del **sesgo**. Los modelos de inteligencia artificial aprenden a partir de datos humanos y, por tanto, pueden reproducir errores, prejuicios o desigualdades presentes en esos datos. Si los datos de entrenamiento están sesgados, el sistema aprenderá esos mismos sesgos y los reproducirá en sus decisiones.

Otro límite importante es la **opacidad**. Muchos modelos avanzados, especialmente los basados en deep learning, funcionan como una especie de “caja negra”. Esto significa que pueden producir resultados muy precisos, pero no siempre es fácil comprender exactamente cómo han llegado a esa conclusión.

Existe además un riesgo más sutil: la **apariencia de inteligencia**. Los sistemas de IA pueden expresarse con gran seguridad incluso cuando se equivocan, lo que puede generar una falsa sensación de fiabilidad. Entender que estos sistemas no comprenden el mundo, sino que predicen patrones, es una de las claves para desarrollar pensamiento crítico en el alumnado.

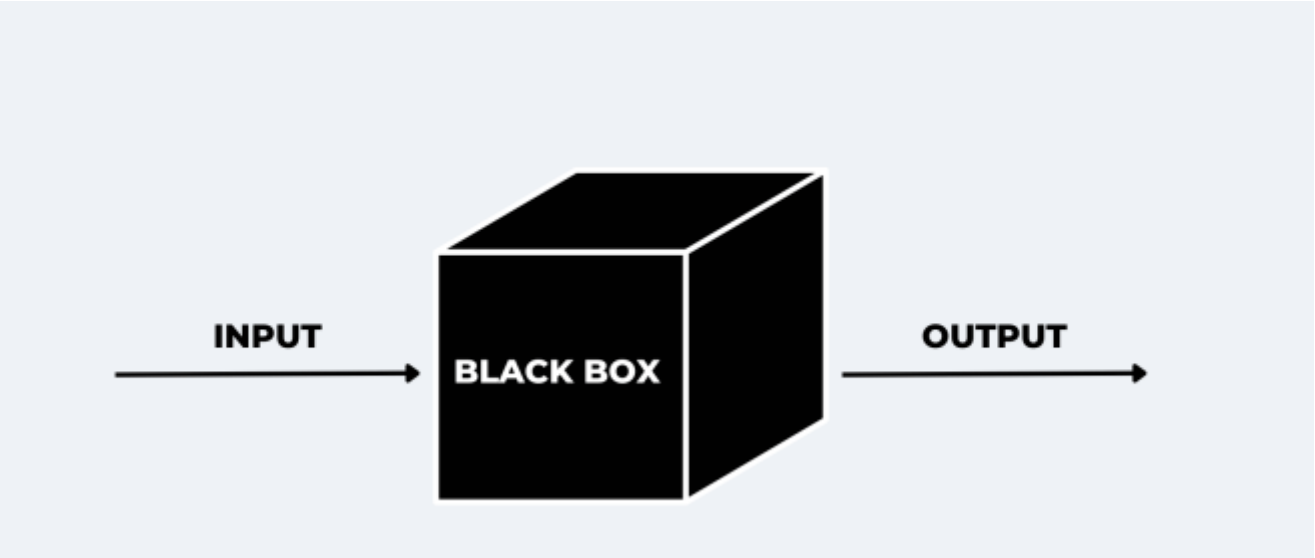
Por eso, en el contexto educativo, la inteligencia artificial debe entenderse siempre como **una herramienta**, no como un sustituto del pensamiento humano ni del criterio pedagógico. Puede ayudar a analizar información, generar materiales o personalizar el aprendizaje, pero la interpretación de los resultados, la transmisión de valores y la construcción de significado siguen siendo tareas profundamente humanas.

En última instancia, podría resumirse con una idea sencilla para el aula:

Las máquinas pueden aprender patrones.
Las personas, además de aprender, **comprenden por qué y para qué aprenden**.

Y precisamente ahí es donde el papel del profesorado sigue siendo insustituible.

Copyright 2026



Comprender cómo aprenden las máquinas es imprescindible para usarlas bien, pero **no es suficiente**. Igual de importante es entender sus límites.

El primero es el **sesgo**. La IA aprende de datos humanos y, por tanto, hereda errores, prejuicios y desigualdades presentes en esos datos. Un sistema entrenado con información parcial o injusta reproducirá esas mismas distorsiones, a menudo de forma invisible.

El segundo límite es la **opacidad**. Muchos modelos, especialmente los basados en Deep Learning, funcionan como una “caja negra”: ofrecen resultados muy precisos, pero difíciles de explicar. En educación, esto obliga a extremar la prudencia cuando se utilizan para evaluar, clasificar o tomar decisiones relevantes.

El tercero es la **apariencia de inteligencia**. Los sistemas de IA pueden expresarse con seguridad incluso cuando se equivocan, lo que puede generar una falsa sensación de fiabilidad. Comprender que la IA no entiende, sino que **predice**, es clave para desarrollar pensamiento crítico en el alumnado.

Copyright 2026

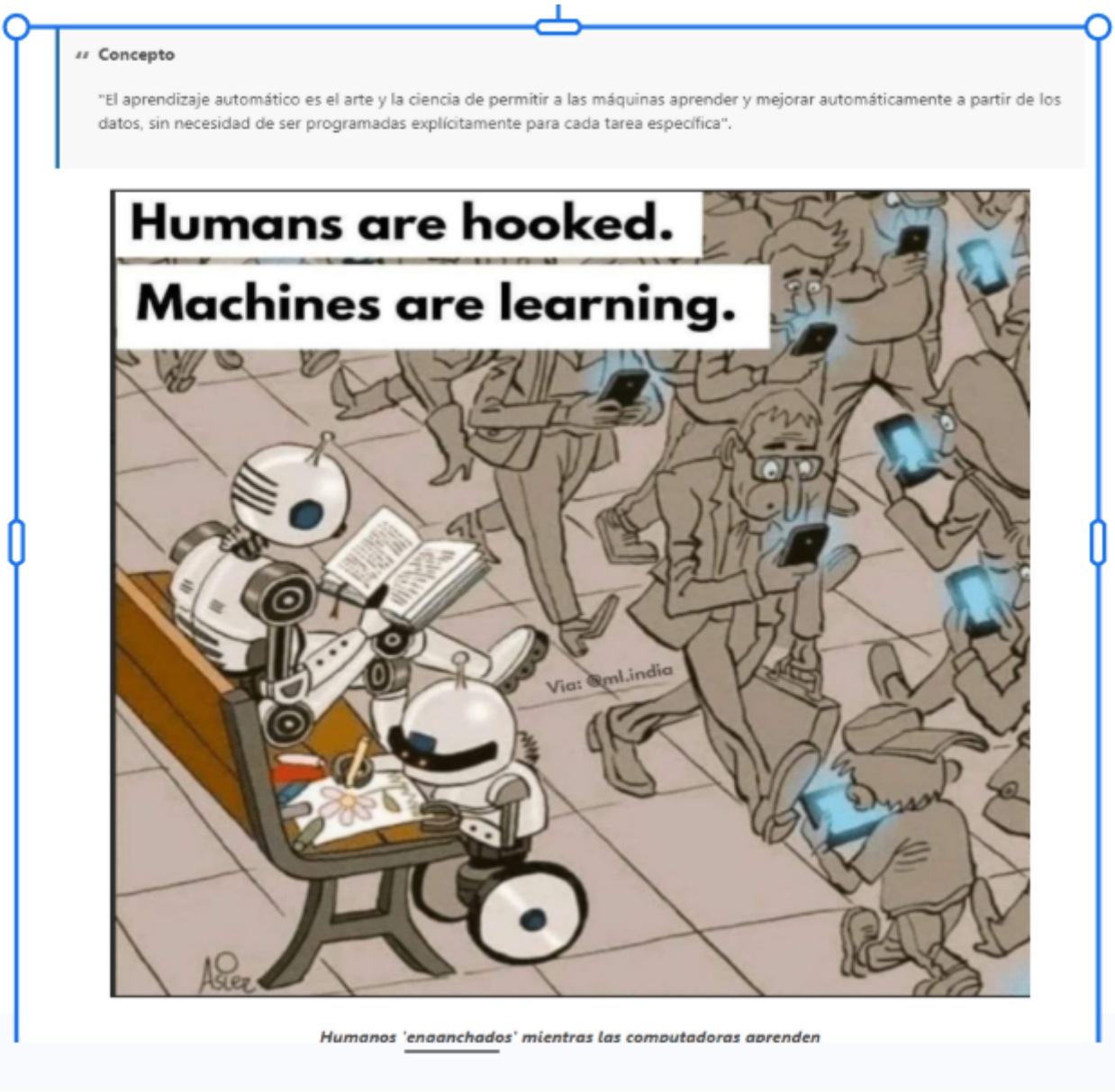


Imagen autoexplicativa que demuestra el estado actual de las cosas. Humanos consumiendo información y máquinas aprendiendo

> **Las máquinas aprenden con datos; las personas aprenden con sentido, valores y acompañamiento.**

Revision #1
Created 2026-09-11 12:23:27 CEST by Silvia Coscolin Sanchez
Updated 2026-09-11 12:27:47 CEST by Silvia Coscolin Sanchez