2.1 Introducción a la ética de la IA

Una perspectiva existencial y pedagógica

Este capítulo, dedicado a la ética en torno a la inteligencia artificial, es, como no podía ser de otra forma, un paseo por algunos de los conceptos que consideramos claves en esta transformación. Pretende ser una reflexión abierta, desde una perspectiva múltiple, que facilite a los lectores formarse su propio criterio desde sus valores personales.

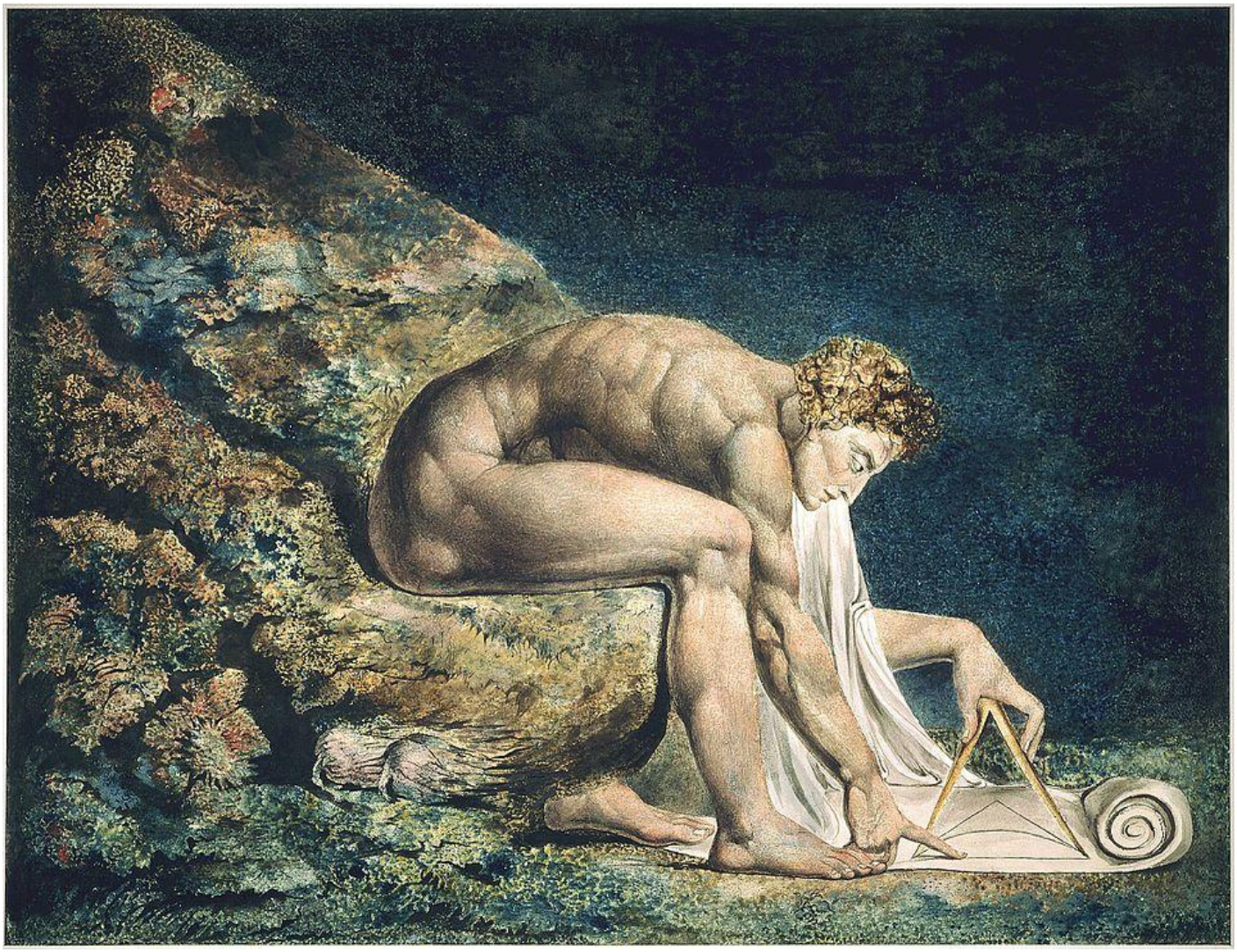
La irrupción global de la inteligencia artificial no es simplemente un cambio en el paradigma del procesamiento de la información, sino que puede suponer una reconfiguración de las bases mismas de la identidad humana y de nuestra autoimagen.

La historia de los avances científicos y los cambios en nuestra cosmovisión que vienen asociados son una constante cura de humildad para la humanidad. El paso del geocentrismo al heliocentrismo, propuesto en el siglo III AC por Aristarco de Samos se consolida como cosmovisión dominante tras "De Revolutionibus Orbium Coelestium" publicada por Copernico en 1543. En este primer salto abandonamos el centro del universo para pasar a ser viajeros sobre un planeta más.

Posteriormente tras El origen de las especies (1854) de Charles Darwin dejamos de ser la cima de un plan de diseño, la más perfecta de las especies, para ser un afortunado producto de las variaciones en la replicación unidas a la presión selectiva.

¿Es posible que en un futuro no muy lejano tengamos que asumir dejar de ser la única inteligencia que conocemos? Nuestra mente singular está dotada de autoconsciencia y es un milagro de la casualidad que nos permite, entre otras cosas más profanas, comunicarnos, amar, admirar la belleza del mundo e incluso crear belleza a través de la expresión artística.

Copyright 2026



William Blake (1795–1805). *Newton* [Impresión en color, tinta y acuarela sobre papel]. Tate Britain. https://en.wikipedia.org/wiki/Newton_(Blake)

> *Para ver un mundo en un grano de arena*
> *y un paraíso en una flor silvestre,*
> *sostén el infinito en la palma de la mano*
> *y la eternidad en una hora.*
> William Blake

El auge de la Inteligencia artificial ya ha destronado a los seres humanos en algunos campos en los que, hasta donde nosotros sabíamos, teníamos el monopolio. La victoria de Deep Blue frente a Kasparov en 1997 marcó el inicio de la amenaza de los algoritmos sobre nuestra primacía intelectual. La confirmación indudable llega en el duelo en el complejísimo juego oriental de "Go"

Copyright 2026

en el que Alpha Go vence Lee Sedol en 2016. Estos juegos son altamente algoritmizables y no dejan de tener unas reglas definidas y lógicas que permiten una mayor facilidad para el funcionamiento de una máquina. Se puede decir que en 2016 la mente humana fue superada por los ordenadores que jugaban en su propio terreno.

La llegada de la Inteligencia Artificial Generativa nos hace sentir la amenaza en otros terrenos más indefinidos, como la creación artística, o muchos trabajos intelectuales. Es la primera vez en la historia de la humanidad que un avance tecnológico pone en cuestión el trabajo de las élites intelectuales, hasta ahora todo avance en la mecanización iba orientado a la sustitución del trabajo físico.

El empleo de la IA va a convertirnos, usando la metáfora de **Cory Doctorow** explicada en el punto 1.4 de este curso**,** bien en centauros, bien en centauros inversos. Una de las claves es identificar aquellas actividades, experiencias y vivencias que consideramos valiosas y en las que, bien porque forman parte de algo intrínsecamente humano, bien por somos superiores a la Inteligencia Artificial en su resolución, pueden constituir, en el futuro, un santuario humano.

https://www.youtube.com/embed/Q8CsSvGffjg

Charla TedX de Javier Recuenco en Málaga. 20 de abril de 2024

La libertad humana y la tentación de delegar decisiones existenciales

Las dos primeras acepciones de "Libertad en la RAE" son:

> 1. f. Facultad natural que tiene el hombre de obrar de una manera o de otra, y de no obrar, por lo que es responsa-ble de sus actos.
> Sin.:
> - voluntad, albedrío, autodeterminación.
>
> 2. f. Estado o condición de quien no es esclavo.
> Sin.:
> - liberación.
>
> Ant.:
> - esclavitud.

Como toda palabra desgastada por el uso, la polisemia es muy grande y la RAE recoge hasta 12 acepciones.

Los humanos podemos elegir y no podemos no elegir. Es frecuente la tentación de delegar nuestras elecciones, nuestra libertad, ante la incertidumbre que siempre existe respecto a las consecuencias de nuestros actos. La angustia que surge de esta libertad absoluta, "Angst" en Søren Kierkegaard, es el "vértigo de la libertad", un estado previo al pecado y a la acción, donde el ser humano se enfrenta al abismo de sus propias posibilidades.

Es el peso de saber que no existen excusas externas para nuestros actos. En la era digital, la IA se presenta como el refugio ideal contra esta angustia. Delegar decisiones críticas a un sistema algorítmico, desde la selección de personal hasta diagnósticos médicos o sentencias judiciales, permite al sujeto humano incurrir en lo que Sartre denominó "mala fe" (*mauvaise foi*): el intento de escapar de la responsabilidad fingiendo que uno es un objeto determinado por fuerzas externas. Esta "mauvaise foi" no es mas que una versión refinada de "la noche me confunde".

La mala fe se manifiesta cuando las instituciones o individuos afirman que una decisión es "objetiva" o "neutral" simplemente porque ha sido generada por un algoritmo. Esta ilusión de neutralidad ignora que los sistemas de IA están impregnados de los sesgos y las intenciones de sus creadores, así como de las desigualdades presentes en los datos de entrenamiento. El Reglamento Europeo de IA identifica este riesgo y establece que la IA debe ser una tecnología centrada en el ser humano, actuando como una herramienta que aumente el bienestar y no como un sustituto de la voluntad humana.

Elegir es una cualidad esencialmente humana, problemática y en ocasiones dolorosa, pero es imprescindible hacerse cargo para una vida plena.

La interacción entre la humanidad y la IA puede ser analizada a través de la dialéctica del amo y el esclavo de Hegel. En este conflicto por el reconocimiento, el amo es quien somete al otro, pero al delegar todo trabajo y contacto con la realidad en el esclavo, el amo termina por atrofiarse, volviéndose dependiente y pasivo. El esclavo, en cambio, mediante el trabajo formativo (*bildung*), transforma la materia y desarrolla su autoconciencia, encontrando en la actividad laboriosa el camino hacia su propia liberación, aquí la metáfora pierde poder explicativo pues una IA liberada es, de momento, ciencia ficción.

En cualquier caso, y para lo que nos ocupa, el ser humano adopta la posición del "amo" que delega todas sus facultades intelectuales, creativas y de toma de decisiones en la IA, que haría el papel de un nuevo "esclavo" cognitivo. De este modo se estanca, y pierde la capacidad de juicio crítico y de acción autónoma.

La indefensión aprendida ante los cambios tecnológicos.

Copyright 2026

> La **indefensión aprendida** (también, **desesperanza aprendida** o **impotencia aprendida**) es un tecnicismo acuñado por Martin Seligman que se refiere a la condición de un ser humano o de un animal no humano que ha "aprendido" a comportarse pasivamente, con la sensación subjetiva de que no tiene la capacidad de hacer nada y que no responde a pesar de que existen oportunidades reales de cambiar la situación aversiva, evitando las circunstancias desagradables u obteniendo recompensas positivas.
>
> Wikipedia. Indefensión aprendida.

La aceptación pasiva de los cambios sociales y educativos que producirá la IA no es solo un fenómeno psicológico individual, sino una construcción social. No necesariamente la humanidad tiene que elegir desarrollar y emplear todas las posibilidades y avances científicos y tecnológicos. Esta posición, lejos de ser ingenua, puede sostenerse históricamente, puesto que existen ejemplos, como el Protocolo de Montreal, donde hubo un acuerdo unánime para evitar usar ciertos productos químicos y revertir el agujero de la capa de ozono.

Es un acto de valentía hacernos cargo de nuestro presente para construir un futuro deseable y acogedor. Podemos dirigir, hasta cierto punto, la forma en que las transformaciones tecnológicas transforman nuestra sociedad y, en particular, nuestra educación, el porvenir vendrá dado por la suma de nuestras elecciones.

La puntuación social o ciudadana

Una de las formas más extremas de control social de la conducta de los individuos es la "puntuación ciudadana" o "puntuación social" (*social scoring*).

> “Un sistema de "puntuación social" asigna puntos positivos o negativos a los ciudadanos en función de sus comportamientos y los recompensa o castiga en sus relaciones con el estado y otros agentes sociales. Por ejemplo, cruzar un semáforo en rojo puede restarte puntos. Tener un balance negativo puede dificultar tu acceso al crédito o a una beca de estudios"

Estos sistemas eran inviables hasta que nuestra capacidad para recopilar, almacenar y procesar datos se ha multiplicado exponencialmente. El Reglamento Europeo de IA prohíbe explícitamente el uso de sistemas de IA para clasificar la fiabilidad de las personas físicas basándose en su comportamiento social en contextos inconexos, ya que esto conduce a un trato discriminatorio e injustificado.

Copyright 2026

Este tipo de prácticas elimina la presunción de inocencia y la posibilidad de que el individuo se defina a sí mismo fuera de su huella digital. Es una forma oscura de gamificación de la participación ciudadana donde nuestra vida pública pasa a ser un dato a optimizar. Si la escuela es un entorno controlado donde experimentar cara a la vida adulta, debemos plantearnos, al emplear sistemas similares, que tenemos que tener en cuenta que hay muchos factores que no podemos evaluar, que no todo en el aula es un dato, y debemos ser abiertos y compasivos.

Artículo 5.1 c del Reglamento Europeo de Inteligencia Artificial

> Quedan prohibidas las siguientes prácticas de IA: c) La puesta en el mercado, la puesta en servicio o la utilización de sistemas de IA para la evaluación o clasificación de personas físicas o grupos de personas durante un determinado período de tiempo en función de su comportamiento social o de sus características personales o de personalidad conocidas, deducidas o previstas, con la puntuación social que conduzca a una de las siguientes situaciones o a ambas:
> (i) trato perjudicial o desfavorable de determinadas personas físicas o grupos de personas en contextos sociales que no guardan relación con los contextos en los que se generaron o recopilaron originalmente los datos;
> (ii) trato perjudicial o desfavorable a determinadas personas físicas o grupos de personas, injustificado o desproporcionado en relación con su comportamiento social o su gravedad;

Dimensiones no computables de la experiencia humana

Otro riesgo ético de la expansión de los algoritmos y de la IA es alimentar la idea de que la realidad humana es totalmente capturable mediante los datos y la lógica matemática y reducible a ellos. Sin embargo, existen dimensiones de la existencia que son intrínsecamente no computables, pues carecen de los patrones repetitivos y las estructuras lógicas que requieren los algoritmos y trascienden esa dimensión lógica y fáctica. En esas dimensiones no computables podemos incluir los deseos, la espiritualidad, la corporeidad, la estética, la ternura, la búsqueda del placer... Una descripción de la realidad humana en la que solo existen hechos, acciones y relaciones entre ellos, es una humanidad amputada para ser algoritmizable,

Catherine L'Ecuyer defiende la "pedagogía del asombro" como la base de todo aprendizaje auténtico. El asombro es la respuesta interna ante la belleza y el misterio de la realidad, y requiere condiciones que la IA suele destruir: silencio, calma, tiempo y respeto por los ritmos naturales. La IA opera en el ámbito de la inmediatez y el ruido informativo; sus resultados son respuestas, nunca preguntas que abran al misterio.

Copyright 2026

El conocimiento verdadero no es una acumulación de información (datos), sino un estado que transforma al sujeto. Mientras que la IA puede procesar millones de terabytes, es incapaz de experimentar el "asombro ante el universo" que define la curiosidad humana. Educar en la era de la IA implica proteger la inocencia del niño frente al bombardeo de estímulos digitales que anulan su capacidad de asombrarse por lo real.

El amor, la ternura y la compasión son experiencias que involucran la vulnerabilidad física y emocional, dimensiones que escapan a cualquier modelo predictivo. Un sistema de IA puede simular una conversación empática, pero no puede "sentir" la compasión, pues no conoce el dolor ni la finitud ni la corporeidad. La ética de la IA debe reconocer estos límites, evitando la suplantación de vínculos humanos por interfaces artificiales.

Respecto a la estética, la IA generativa puede producir imágenes estéticamente agradables mediante la combinación de patrones ya existentes, pero carece de la capacidad de crear una obra que nazca de una necesidad existencial o de un conflicto espiritual, así como tampoco puede crear un nuevo estilo.

El arte es una forma de resistencia porque introduce la negatividad, el dolor y la ruptura en un mundo de positividad algorítmica. Mantener espacios para la creación humana no asistida por IA es una defensa de la singularidad de nuestra especie. La verdadera innovación no es hacer las cosas más rápido, sino darles un sentido profundo que trascienda la lógica del rendimiento.

Mas allá de esta perspectiva de creadores de belleza, de seres que necesitan expresarse, son valiosas también por si mismas las experiencias del disfrute de la belleza o del asombro ante las creaciones artísticas. ¿Cuántos puntos nos daria en un Class Dojo algoritmizado malgastar 1 h contemplando una escultura? ¿Y una puesta de sol?

> Al ladrón
> se le olvidó la luna
> en la ventana


Daigu Ryokan

Principios éticos

UNESCO

En noviembre de 2021, la UNESCO elaboró la primera norma mundial sobre la ética de la IA: la "Recomendación sobre la ética de la inteligencia artificial". Los cuatro valores en los que se basa son:

1. Derechos humanos y dignidad humana
2. Vivir en sociedades pacíficas
3. Garantizar la diversidad y la inclusión
4. Florecimiento del medioambiente y los sistemas

Establece un **enfoque de la ética de la IA centrado en los derechos humanos,** establecido por 10 principios:

1. Proporcionalidad e inocuidad: el uso de sistemas de IA no debe ir más allá de lo necesario para alcanzar un objetivo legítimo. La evaluación de riesgos debe utilizarse para prevenir los daños que puedan derivarse de usos ilegítimos.

2. Seguridad y protección: los daños no deseados (riesgos de seguridad) y las vulnerabilidades a los ataques (riesgos de protección) deberían ser evitados y tomados en consideración.

3. Derecho a la intimidad y protección de datos: la privacidad debe protegerse y promoverse a lo largo de todo el ciclo de vida de la IA. También deben establecerse marcos adecuados de protección de datos.

4. Gobernanza y colaboración adaptativas y de múltiples partes interesadas: en el uso de datos, deben respetarse el derecho internacional y la soberanía nacional. La participación de diversas partes interesadas a lo largo del ciclo de vida de los sistemas de IA es necesaria para el desarrollo de enfoques inclusivos de gobernanza.

5. Responsabilidad y rendición de cuentas: los sistemas de IA deben ser auditables y trazables. Deben existir mecanismos de supervisión, evaluación de impacto, auditoría y diligencia debida para evitar conflictos con las normas de derechos humanos y amenazas al bienestar medioambiental.

6. Transparencia y explicabilidad: el despliegue ético de los sistemas de IA depende de su transparencia y explicabilidad (T&E). El nivel de T&E debe ser adecuado al contexto, ya que puede haber tensiones entre T&E y otros principios como la privacidad, la seguridad y la protección.

7. Supervisión y decisión humanas: los Estados Miembros deberían velar por que siempre sea posible atribuir la responsabilidad ética y jurídica a personas físicas o a entidades jurídicas existentes.

8. Sostenibilidad: las tecnologías de IA deben evaluarse en función de su impacto en la "sostenibilidad", entendida como un conjunto de objetivos en constante evolución, incluidos los establecidos en los Objetivos de Desarrollo Sostenible (ODS) de Naciones Unidas.

9. Sensibilización y educación: la sensibilización y la comprensión del público respecto de la IA y el valor de los datos deberían promoverse mediante una educación abierta y accesible, la participación cívica, las competencias digitales y la capacitación, y la alfabetización mediática e información.

10. Equidad y no discriminación: los actores de la IA deberían promover la justicia social, salvaguardar la equidad y luchar contra todo tipo de discriminación, adoptando un enfoque inclusivo para garantizar que los beneficios de la IA sean accesibles para todos.

OCDE

Por otro lado, la **OCDE** (Organización para la Cooperación y Desarrollo Económicos) propone la adopción de una serie de principios para una gestión responsable y de confianza de la **Inteligencia Artificial (IA)**. Se trata de un instrumento jurídico no vinculante pero que representa un compromiso político de los países adherentes. Los principios de la OCDE sobre la IA se basan los valores del respeto a los derechos humanos y los valores democráticos, la inclusión, la diversidad,

la equidad, la innovación y el bienestar.

Estos principios son complementarios y deben considerarse en su conjunto para promover una IA en la que se pueda confiar.

1.1. Crecimiento inclusivo, desarrollo sostenible y bienestar: Los actores interesados deben participar proactivamente en la gestión responsable de una IA fiable para buscar resultados beneficiosos para las personas y el planeta. Esto incluye potenciar las capacidades humanas, mejorar la creatividad, avanzar en la inclusión de poblaciones subrepresentadas y reducir las desigualdades económicas, sociales y de género. Asimismo, se debe proteger el medio ambiente para vigorizar el crecimiento inclusivo y la sostenibilidad ambiental.

1.2. Respeto al estado de derecho, los derechos humanos y los valores democráticos: Los actores de la IA deben respetar los derechos humanos y los valores centrados en el ser humano durante todo el ciclo de vida del sistema, incluyendo la libertad, la dignidad, la privacidad y la justicia social. Es fundamental abordar la **desinformación y la información errónea** amplificadas por la IA, respetando siempre la libertad de expresión. Para ello, se deben implementar mecanismos de **supervisión humana** y salvaguardias contra el uso indebido, ya sea intencionado o no.

1.3. Transparencia y explicabilidad: Los actores deben comprometerse con la divulgación responsable proporcionando información que fomente la comprensión general de las capacidades y limitaciones de los sistemas. Se debe informar a las personas cuando estén interactuando con sistemas de IA y, siempre que sea posible, ofrecer información clara sobre los datos y procesos que conducen a un resultado o decisión específica. Esto permite que quienes se vean afectados negativamente por un sistema puedan **impugnar sus resultados**.

1.4. Robustez, seguridad y protección: Los sistemas de IA deben ser robustos y seguros para que, en condiciones de uso normales o ante usos indebidos previsibles, no supongan riesgos de seguridad irrazonables. Deben existir mecanismos para anular, reparar o desactivar de forma segura los sistemas si presentan comportamientos no deseados o riesgos de daño. Además, se deben aplicar medidas para reforzar la integridad de la información respetando la libertad de expresión.

1.5. Responsabilidad: Los actores de la IA son responsables del correcto funcionamiento de los sistemas y del cumplimiento de los principios mencionados según su rol y contexto. Deben garantizar la **trazabilidad** de los conjuntos de datos y los procesos de decisión para permitir su análisis posterior. Finalmente, deben aplicar un enfoque sistemático de **gestión de riesgos** en cada fase del ciclo de vida de la IA para abordar sesgos perjudiciales y proteger los derechos humanos y la propiedad intelectual.

Revision #19
Created 2026-01-01 20:09:45 CET by Maria
Updated 2026-03-15 07:00:14 CET by Chefo Cariñena

Copyright 2026