EDICIONES COMPLUTENSE

**Derecom**
e-ISSN: 1988-2629

ARTÍCULOS

# “Creando” una auditoría para verificar la ética de la Inteligencia artificial

**Laura Davara Fernández de Marcos**
Socia de Davara&Davara
Universidad Europea de Madrid (España)

https://dx.doi.org/10.5209/dere.102348 Recibido: 13/12/2024 • Revisado: 24/02/2025 • Aceptado: 05/03/2025

ES **Resumen.** El presente artículo se centra en la creación de una auditoría para verificar la ética en la inteligencia artificial (IA) puesto que constituye el tercer pilar de la IA junto con cuestiones legales y técnicas. Precisamente por la importancia de auditar la ética que hay detrás de las decisiones de la IA, la autora utiliza ChatGPT para proponer una auditoría ética basándose, en un primer momento, en dar respuesta a ocho cuestiones esenciales, a saber: ¿Cómo se garantiza que los datos utilizados para entrenar la IA están libres de sesgos discriminatorios?; ¿Qué mecanismos hay para que los usuarios afectados por decisiones de la IA puedan apelar decisiones sesgadas?;¿Qué medidas se implementan para evitar que los modelos generen contenido ofensivo o dañino?; ¿Cómo se gestiona el uso de IA para prevenir su abuso en la creación de *deepfakes* o la manipulación de imágenes y videos?;¿Cómo se asegura que los datos utilizados para entrenar la IA cumplen con las normativas de protección de datos como el GDPR?; ¿Cómo se maneja el impacto ambiental del entrenamiento de sus modelos de IA?; ¿Qué medidas se implementan para garantizar que su IA no sea utilizada con fines maliciosos?; ¿Qué medidas se toman para proteger a los menores de la exposición a IA que pueden influir negativamente en su desarrollo? El objetivo del artículo es sentar las bases de una auditoría de la ética de la IA, asegurando que la ética esté integrada en el ADN de cualquier IA diseñada.
**Palabras clave.** ética, Inteligencia Artificial, auditoría, sesgos y protección de datos.

## ENG “Creating” an audit to check the ethics of Artificial Intelligence

ENG **Abstract.** In this paper we focus on creating an audit to verify the ethics in artificial intelligence (AI) as it turns out to be the third pillar of AI alongside legal and technical issues. Precisely because of the importance of auditing the ethics behind AI decisions, the author uses ChatGPT to propose an ethical audit based initially on addressing eight essential questions, namely: How is it ensured that the data used to train the AI is free from discriminatory biases?; What mechanisms are in place for users affected by AI decisions to appeal biased decisions?; What measures are implemented to prevent models from generating offensive or harmful content?; How is the use of AI managed to prevent its abuse in creating deepfakes or manipulating images and videos?; How is it ensured that the data used to train the AI complies with data protection regulations such as GDPR?; How is the environmental impact of training AI models managed?; What measures are implemented to ensure that AI is not used for malicious purposes?; What measures are taken to protect minors from exposure to AI that may negatively influence their development? The goal of the paper is to lay the foundations for an AI ethics audit, ensuring that ethics is integrated into the DNA of any designed AI.
**Keywords.** ethics, Artificial Intelligence, audit, biases, data protection.



**Cómo citar**: Davara Fernández de Marcos, L. (2025). “Creando” una auditoría para verificar la ética de la Inteligencia artificial. *Derecom* 38(1), 95-102. https://dx.doi.org/10.5209/dere.102348

## 1. Seamos realistas: ¡pidamos lo imposible!

Puede parecer sorprendente comenzar este trabajo con la frase del filósofo Herbert Marcuse "Seamos realistas: ¡pidamos lo imposible" pero, desde ya, queremos comenzar afirmando que, el tema que vamos a abordar en él no es imposible –pese a lo que muchos consideran– pero sí es extremadamente difícil.

Pero, también les decimos, es tan difícil como necesario. Y es por ello por lo que queremos dedicar este trabajo a una de las cuestiones a las que, en nuestra opinión, mayor importancia se le ha de dar en lo que a inteligencia artificial se refiere: ¡la ética!

Por supuesto, no es nuestra intención restar importancia a todas las cuestiones legales y técnicas, pero forman un triángulo equilátero porque, sí, los tres lados son iguales, al menos en lo que a importancia se refiere.

¿Y por qué para muchos la ética de la inteligencia artificial es algo inalcanzable? Porque hay quien opina que la ética –diferente a la moral– es una cuestión subjetiva, que está relacionada con la educación o incluso con la religión o la cultura. Y no ponemos en duda que estas realidades –la educación, la creencia en la divinidad o la cultura– influyan en los valores que acaban formando la ética que cada uno ponemos en juego en todas nuestras acciones. Pero lo cierto es que no queremos ampararnos en que "no hay una ética universal" para dejar de lado la importancia de, al menos, auditar para conocer la ética que hay detrás de esa famosa "Inteligencia artificial" (en adelante, también, IA) que tantas veces utilizamos sin saber qué hay detrás.

Por este motivo y precisamente para demostrar tanto las bondades como los defectos de la inteligencia artificial y, en concreto, de la que mayor éxito y acogida ha tenido en los últimos dos años, hemos querido contar con ChatGPT para tratar de proponer una "auditoría para verificar la ética en la IA". Para ello, le hemos proporcionado el siguiente prompt[1]:

> *"Eres un experto en ética en IA, quiero que prepares una lista de 20 preguntas[2] que abarquen todas las cuestiones que debe tener en cuenta un programador que diseñe una IA para garantizar una ética y moral adecuada a las exigencias de la UNESCO y de la moral y las buenas costumbres de manera que, con toda la información que me des y con mi conocimiento, pueda proponer un "protocolo de ética" y/ o una "auditoría de la ética de la IA". Quiero, además, que, a cada una de las preguntas, añadas la respuesta que crees que debería dar para garantizar que se respeta la ética y la moral y que no se causa ningún daño –ni directo ni indirecto– a ningún ser humano, respetando los derechos humanos y los derechos fundamentales de la Constitución española. Y quiero, además, que añadas un párrafo en cada una de las respuestas con un ejemplo concreto y real (en caso de que exista) de qué podría ocurrir si no se tuviesen en cuenta esas cuestiones. Quiero que seas concreto, riguroso, sincero, que no ocultes nada y que te pongas siempre en el peor de los casos para que la ética, verdaderamente, esté en el ADN de la IA que vaya a diseñar quien responda a estas preguntas. Además, quiero que me indiques "red flags" en cada una de las preguntas en el sentido de respuestas que pudieran suponer un riesgo, explicando la manera de solucionar y/o prevenir esos riesgos, siempre de la manera más concreta y explicativa posible. Sin teorías, con muchos ejemplos. Añade, además, a cada pregunta, alguna cuestión desconocida o poco conocida sobre lo que hay detrás de la IA y la importancia de conocerla y explica por qué se gana tanto dinero con un servicio aparentemente gratuito y por qué varios de los expertos mundiales quisieron detener su lanzamiento cuando se vio todo el potencial dañino que podía generar. Hazlo todo de manera clara, sencilla, práctica y añadiendo enlaces oficiales donde ampliar información. Hazme todas las preguntas que necesites para que la respuesta sea lo mejor posible[3]".*

## 2. Cuestiones a considerar a la hora de crear una auditoría para la IA

### 2.1. Primera cuestión: ¿Cómo garantizan que los datos que utilizan para entrenar la IA están libres de sesgos que puedan causar discriminación?

Si bien no queremos trasladar íntegramente todas las respuestas dadas por ChatGPT puesto que convertiríamos este trabajo en un trabajo realizado por una IA, sí que compartiremos con el lector que el chatbot considera que para "pasar positivamente" la auditoría, la respuesta debería implicar el uso, precisamente, de mecanismos que auditen los datos utilizados "antes, durante y después del entrenamiento" para garantizar que no se trata de datos sesgados. Y, para ello, aboga por el uso de "técnicas como el sobreentrenamiento de clases subrepresentadas" y por la colaboración de expertos en diversidad para identificar lo que se conoce como "puntos ciegos".

Y nosotros nos preguntamos ¿acaso cuando utilizamos ChatGPT o cualquier otra IA nos paramos a

1 Vaya por delante que no queremos que se tome como "ejemplo del prompt perfecto", ni mucho menos. Simplemente queremos ser totalmente transparentes y compartir la idea inicial de la que partía. De hecho, el propio ChatGPT me hizo tres preguntas tras mi prompt y tras responderle, ya me dio sus respuestas sobre las cuales realizaré el presente trabajo con el ánimo de hacer reflexionar al lector sobre la importancia de las cuestiones éticas y morales que puede –y debe– haber en todo lo que a tecnología se refiere, no siendo la IA una excepción.

2 En el trabajo verán solo reflejadas ocho porque, de las veinte, hemos elegido ocho que nos parecían relevantes y que, por lo limitado del espacio, nos permitían hacer una reflexión sobre cada una de ellas y aportar un poco de luz en este sentido. Pero, por supuesto, son mucho más de ocho las cuestiones que se han de tener en cuenta en lo que a ética de la inteligencia artificial se refiere.

3 A lo largo del presente trabajo, cuando se trate de una respuesta dada literalmente por ChatGPT, lo pondremos en cursiva y, sobre ella, compartiremos nuestras reflexiones, inquietudes y propuestas.

pensar si esos resultados que nos están arrojando están basados en datos sesgados? ¿Sabemos lo que es el sesgo y lo que, en la práctica, puede suponer para los afectados? Según la RAE, en su séptima acepción, sesgo es "Error sistemático en el que se puede incurrir cuando al hacer muestreos o ensayos se seleccionan o favorecen unas respuestas frente a otras".

Uno de los casos reales más sonados sobre el uso de datos sesgados para entrenar aplicaciones basadas en inteligencia artificial es el del gigante Amazon quien no dudó en utilizar –por un tiempo limitado– un software basado en IA para llevar a cabo el proceso de contratación para seleccionar nuevos trabajadores. ¿Y por qué lo utilizó únicamente durante un tiempo limitado? Porque los resultados que arrojaba estaban sesgados (Reuters, 2024), llevando a la compañía a favorecer a los hombres en detrimento de las mujeres, precisamente por determinados datos que se habían introducido en el diseño de la IA que favorecían a los hombres en lugar de a las mujeres sin motivo objetivo para ello.

Quizás hay que ir un paso más allá o, mejor dicho, un paso más atrás e ir al origen y preguntarnos ¿de dónde, de qué y de quién se nutre la inteligencia artificial[4]? Y la respuesta no es sencilla ni breve porque no hay una única fuente[5] de la que beba la inteligencia artificial –de hecho, sería mucho más fácil auditar la IA en general y, en concreto, la ética y los valores que hay detrás de la IA si así fuera porque bastaría con analizar la veracidad, adecuación y ética de esa fuente– sino que son muchas y de lo más variado, entre los que podemos citar:

- Wikipedia y otras plataformas de código abierto.
- Noticias y medios de comunicación
- Artículos científicos, bases de datos, manuales y libros
- Foros y sitios web tipo Reddit y Quora de los que se obtiene la manera en la que se relacionan los usuarios.
- Blogs disponibles en la red. El objetivo de acudir a estas fuentes para obtener información es aprender términos de lenguaje natural y mejorar la manera de expresarse.
- Textos de código abierto
- Interacción con los usuarios

Son muchas y muy variadas las fuentes de las que "bebe" la inteligencia artificial. Y, en nuestra opinión, no son todas igual de fiables ni de objetivas, con los enormes riesgos que ello supone. En este sentido, el propio ChatGPT plantea como red flag el hecho de que las IA se entrenan con frecuencia con lo que califica como "datos no regulados" o, dicho de otra manera, datos obtenidos de las redes sociales y de otras plataformas que pueden comportar sesgos que, una vez trasladados a la IA, supone un riesgo enorme de replicar dichos sesgos en los resultados que aporte la IA y hacerlo de manera automática y sin ser identificados como tales.

### 2.2. Segunda cuestión: ¿Qué mecanismos tienen para que los usuarios afectados por decisiones de la IA puedan apelar decisiones sesgadas?

La ética está muy relacionada con la justicia –o con la injusticia, según se quiera ver– y es por ello que, en ese marco de auditoría, es fundamental que desde la propia programación de la IA se ofrezcan mecanismos que permitan garantizar que el usuario pueda solicitar una revisión humana en el caso de que una decisión automatizada le esté causando un daño.

En este punto cobra especial importancia el derecho para los interesados previsto por la normativa de protección de datos, en concreto en el artículo 22 del Reglamento europeo de protección de datos[6] que, bajo el epígrafe "Decisiones individuales automatizadas, incluida la elaboración de perfiles" reza así: "Todo interesado tendrá derecho a no ser objeto de una decisión basada únicamente en el tratamiento automatizado, incluida la elaboración de perfiles, que produzca efectos jurídicos en él o le afecte significativamente de modo similar".

Como ejemplo de caso real, el propio ChatGPT sugiere hablar del funcionamiento de COMPAS y de su "justicia o injusticia11" (*Technology Review,* 2024) para con los sujetos analizados. Por si alguien no lo conociera, COMPAS es una herramienta de aprendizaje automático que, haciendo uso de inteligencia artificial, lleva a cabo un análisis pormenorizado a más de 130 preguntas –137, para ser exactos– en la que se recoge información sobre historia familiar, estado emocional, antecedentes penales o cuestiones relativas al comportamiento de cara a calcular el riesgo de reincidencia en la comisión de delitos de una determinada persona.

¿Y por qué sugiere analizar esta herramienta? Porque se demostró que sus decisiones, de base, contenían sesgos. ¡E imagínense el enorme impacto en la persona cuyos datos son analizados que podría suponer un fallo en el resultado de esta herramienta! Las implicaciones son enormes y es por ello que aquí queremos concluir trayendo a colación lo que se conoce como "caja negra" de la IA que implica que, en determinadas ocasiones, ni los propios creadores de la IA saben cómo se llega a determi-

4 El propio ChatGPT, al preguntarle, indica: "OpenAI no ha especificado una lista exhaustiva de todas las fuentes específicas utilizadas para entrenar modelos como GPT, pero se ha aclarado que se han empleado grandes volúmenes de datos públicos, sin acceso a contenido privado o protegido por derechos de autor que no esté disponible públicamente. El entrenamiento incluye datos recopilados antes de una fecha límite (por ejemplo, el conocimiento que poseo tiene un corte de septiembre de 2021)". Consulta realizada el 10 de septiembre de 2024.

5 De hecho, el propio ChatGPT tras inquirirle en numerosas ocasiones y de diferentes maneras sobre la manera que tiene de recopilar los datos y los sesgos a los que se pueden ver expuestos, afirma que "Los datos de entrenamiento muchas veces contienen sesgos ocultos, no solo en términos de género o raza, sino también en relación con estratos sociales, niveles educativos o ubicación geográfica. Estos sesgos pueden surgir por la falta de representatividad en los datos o por cómo se recopilan". Consulta realizada el 10 de septiembre de 2024.

6 Reglamento (UE) 2016/679 del Parlamento Europeo y del Consejo, de 27 de abril de 2016, relativo a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos y por el que se deroga la Directiva 95/46/CE (Reglamento general de protección de datos). Publicado en el *DOUE serie L núm. 119, de 4 de mayo de 2016.*

nadas decisiones. No entramos aquí en si realmente no saben o no quieren saber, pero lo que sí es cierto es que la transparencia en los criterios que se tienen en cuenta y en la adopción de las decisiones y la posibilidad de que sea revisado por un ser humano se alza como una cuestión fundamental y de innegable importancia.

### 2.3. Tercera cuestión. ¿Qué medidas implementan para evitar que los modelos generen contenido ofensivo o dañino?

Todas las aplicaciones, programas y servicios que se basen en IA han de contar con filtros que garanticen que el sistema no se entrene con datos violentos –o que generen violencia– ni con datos ofensivos o falsos. Y, por supuesto, que dichos filtros sean revisados por seres humanos. En cuestión de IA, al igual que tecnología en general, creemos –tal y como ya hemos comentado a lo largo del presente trabajo– que la figura de la supervisión humana nunca podrá ser sustituida por una aplicación informática por cuanto en tema de matices, dobles sentidos, ironía, cuestiones e interpretaciones culturales, etc. ¡El ser humano siempre irá por delante!

Y es que, yendo al origen, ¿cómo podemos garantizar que los sistemas no se entrenen, en este caso concreto, con contenido violento u ofensivo? En este punto, traemos a colación uno de los casos más famosos y paradigmáticos para ilustrar este problema: el caso del ChatBot Tay.

Si bien es cierto que es un caso de hace varios años –puesto que se lanzó al mercado en marzo de 2016–, creemos que es realmente interesante hacer una breve referencia a lo que ocurrió con este chatbot puesto que pone de manifiesto la posibilidad –más fácil y rápida de lo que a priori pueda parecer– de que la IA pueda ser "origen y causa" de contenido violento e inadecuado.

Resumidamente: Microsoft diseñó y lanzó al mercado a través de Twitter –ahora X– un chatbot llamado Tay que trataba de imitar a una chica americana de 19 años en lo que a interacciones y relaciones se trataba. Sin embargo, pese a la intención inicial, en unas horas Tay empezó a emitir todo tipo de tuits de carácter racista, sexista y ofensivo puesto que, si se nos permite la expresión, "había sido víctima de su propia capacidad de aprendizaje" y acogió todas las observaciones y comentarios –apropiados e inapropiados– que le hicieron los usuarios. Ante la avalancha de quejas y problemas que suscitó tal situación, Microsoft procedió a suspender la cuenta de Tay en el antiguo Twitter y, por supuesto, a pedir disculpas a todos los afectos por los mensajes absolutamente inapropiados que desde la cuenta de Tay se habían emitido.

### 2.4. Cuarta cuestión: ¿Cómo gestionan el uso de IA para prevenir su abuso en la creación de *deepfakes* o la manipulación de imágenes y videos?

Si preguntamos a Google por qué es famoso Almendralejo, el primer resultado que nos ofrece es este "Desde la década de 1980 la ciudad ha adquirido cierta tradición como centro productor de cava. Almendralejo es también conocida como «Ciudad del Romanticismo» por ser el lugar de nacimiento de dos principales poetas de este movimiento literario. Y, sin duda, Almendralejo es famoso por estas cuestiones y muchas otras. Pero si preguntamos por el binomio "Almendralejo-Inteligencia Artificial", el resultado es bien distinto puesto que son múltiples los resultados que arroja el buscador en el que alude al caso de *deepfakes* de chicas desnudas creadas por menores de Almendralejo utilizando IA.

Pero empecemos por el principio ¿qué es un *deepfake*? El término *deepfake* es un anglicismo que se utiliza para aludir a un contenido –en formato imagen, audio o vídeo– que imita –a veces con serias dificultades para identificar si se trata de algo veraz o falso– la apariencia de una persona –tanto en la voz, el físico, los gestos...–. Pues bien, existen varias aplicaciones de inteligencia artificial que permiten –a veces de manera gratuita, a veces en modalidad "freemium" – Unión de los términos "free" y "premium" que alude al modelo de negocio basado en la prestación de un servicio gratuito que, para tener algunas funcionalidades extra, requiere de un pago– crear *deepfakes*. Y esto fue precisamente lo que hicieron unos alumnos de un Instituto de Almendralejo con las imágenes de unas compañeras, hablando claro: pusieron la cara de las alumnas en el cuerpo desnudo generado por la IA.

Más allá de todo el daño que causó a las menores, de todo el revuelo mediático y de todo lo que hay detrás –dejando a un lado que, en nuestra opinión, no se puede considerar como "Mera broma entre adolescentes"–, queremos en este punto poner el foco en la propia aplicación de Inteligencia Artificial ¿tiene alguna responsabilidad? ¿hemos de inculcarle valores como el respeto a la intimidad y a la privacidad de las personas o, por el contrario, toda la responsabilidad recae en quien las usa?

Y es que, todas estas preguntas, cuyas respuestas nos darían para un artículo en sí mismo, afectan a todas las *deepfakes*, no solo a las relacionadas con desnudos sino también a las que se generan con la intención de influir en cuestiones electorales o de desestabilizar mercados, por poner solo algunos ejemplos. Por ello, queremos concluir el presente apartado trayendo a colación algunas noticias relacionadas con los *deepfakes* que no hacen sino mostrar el papel protagonista de esta realidad y la necesidad de poner "la ética por defecto" como punto de partida:

- "Inglaterra y Gales prohibirán la creación de contenido para adultos Deepfake sin consentimiento" (Nintenderos, 2024).
- "Los deepfakes sexuales no pararán con solo castigar su distribución. Hay que fijarse en la pionera ley de Corea del Sur" (Pérez, 2024).
- "Deepfakes de famosos: desde Taylor Swift sorteando sartenes a Leo Messi pidiendo que descargues su app" (Hernando, 2024a).
- "El uso de deepfakes se cuadruplica, lo potencia el uso de redes sociales" (Expansión, 2024).
- "Un directivo de Ferrari tuvo una idea genial para evitar una estafa con deepfakes: preguntar al estafador algo que solo el CEO sabía" (Pastor, 2024).
- "Una técnica usada en astronomía puede ser usada para identificar 'deepfakes'" (Hernando, 2024b).

- "Estafa con 'deepfakes': 24 millones de euros y una reunión virtual con solo una persona real" (La Vanguardia, 2024).
- "Debemos proteger nuestra capacidad de reconocer a seres humanos reales: más de 400 expertos firman carta contra los deepfakes" (González, 2024).
- "Deepfakes, clonación de voz y descontextualización: la desinformación enturbia las campañas electorales en México" (Garay, 2024).

### 2.5. Quinta cuestión: ¿Cómo aseguran que los datos utilizados para entrenar la IA cumplen con las normativas de protección de datos como el GDPR?

Al igual que con el resto de las cuestiones, responder a la que aquí planteamos, daría, por sí sola, para un trabajo único. Y es por ello que, en este punto, queremos partir del propio concepto de tratamiento que ofrece el artículo 4 del Reglamento europeo de protección de datos, a saber, "cualquier operación o conjunto de operaciones realizadas sobre datos personales o conjuntos de datos personales, ya sea por procedimientos automatizados o no, como la recogida, registro, organización, estructuración, conservación, adaptación o modificación, extracción, consulta, utilización, comunicación por transmisión, difusión o cualquier otra forma de habilitación de acceso, cotejo o interconexión, limitación, supresión o destrucción"

¿Por qué? Porque, al utilizar la inteligencia artificial podemos encontrarnos ante un tratamiento de datos personales y, por tanto, se han de cumplir todas y cada una de las obligaciones que prevé la normativa (Agencia Española de Protección de Datos, 2024), entre las que cabe destacar:

- Cumplimiento de los principios de licitud, lealtad y transparencia; principio de minimización; principio de exactitud; principio de limitación de la finalidad; principio de limitación del plazo de conservación y principio de integridad y confidencialidad.
- Cumplimiento del principio de *accountability* (rendición de cuentas) que pasa por acreditar el cumplimiento de los antedichos principios y de la actitud de cumplimiento proactivo de la normativa.
- Garantizar el deber de información respecto al tratamiento de los datos personales.
- Permitir el ejercicio de los derechos ARSOPOL cuyas siglas responden a derecho de acceso, rectificación, supresión, oposición, portabilidad, olvido y limitación.
- Implementar el enfoque basado en el análisis y gestión de los riesgos para los datos personales objeto de tratamiento.
- Proporcionar la transparencia necesaria a los interesados en lo que al tratamiento de sus datos personales se refiere.
- Contar con el delegado de protección de datos "como elemento de transparencia", en palabras de la propia Agencia Española de Protección de Datos.

Y es que, de nuevo, el binomio "privacidad y ética" se convierte –o, en nuestra opinión, se ha de convertir en protagonista de todo proceso que implique Inteligencia Artificial en el que se vean envueltos datos personales.

### 2.6. Sexta cuestión: ¿Cómo manejan el impacto ambiental del entrenamiento de sus modelos de IA?

Aunque para muchos resulte sorprendente, es completamente cierto que la Inteligencia Artificial tiene un impacto innegable en el medioambiente. Algunas cifras que así lo acreditan son las siguientes:

- Se calcula que entre el 1 y el 1,5% de la demanda total de electricidad a nivel mundial procede de los centros de datos, lo que equivale a unos 220-320 teravatios/hora (Funds Society, 2024).
- Un estudio de OpenAI sugiere que la cantidad de recursos computacionales necesarios para entrenar los modelos más avanzados de IA ha estado duplicándose cada 3.4 meses, lo que implica un crecimiento exponencial en el consumo energético y las emisiones asociadas (Fundación Fepropaz, 2024).
- La electricidad media de una búsqueda típica de Google es de 0,3 vatios/ hora de electricidad mientras que una búsqueda de ChatGPT es de 2,9 vatios/hora por solicitud (International Energy Agency, 2024).
- El entrenamiento de un único modelo generativo de IA puede consumir hasta 284.000 litros de agua (Funds Society, 2024).
- El proceso de desarrollo de técnicas de 'aprendizaje profundo' puede llegar a emitir 284 toneladas de dióxido de carbono equivalente (Cangelosi, 2024).
- Google publicó la última versión de su informe ambiental, en el cual se revela que la empresa ha registrado un aumento del 48% en sus emisiones de carbono, en comparación con 2019, a raíz de los avances realizados en Inteligencia Artificial (Bencina, 2024).
- La misma Hugging Face calculó que el entrenamiento de BLOOM (su propio modelo de lenguaje) generó 25 toneladas métricas de emisiones de CO2, como apunta un artículo del Instituto Tecnológico de Massachusetts (MIT) (EE.UU.) (Cangelosi, 2024).
- En la fabricación de chips intervienen cientos de sustancias químicas, entre ellas las altamente tóxicas PFAS, una familia de alrededor de 12.000 sustancias químicas que no se descomponen en el medioambiente hasta después de decenas de miles de años, lo que les ha valido el sobrenombre de sustancias químicas eternas (Gerry, 2024).

Y, en este caso, compartimos lo que, a juicio de la propia Inteligencia Artificial, sería una respuesta adecuada en el marco de una auditoría para verificar la gestión del impacto medioambiental. La respuesta fue la siguiente: "Hemos optimizado nuestros procesos de entrenamiento utilizando centros de datos que emplean energías renovables y técnicas de eficiencia computacional. Además, estamos trabajando para reducir el tamaño de nuestros mode-

los y optimizar los ciclos de entrenamiento para reducir su huella de carbono".

Y nuestra pregunta es ¿les parece suficiente como respuesta? ¿lo darían como "apto" en el marco de una auditoría? O, por el contrario ¿les resulta una respuesta vaga, difusa, poco concreta y que apenas aporta valor en un tema de tanta importancia como el impacto de la IA en el medioambiente? Pueden imaginar que nuestra opinión está más cerca de este segundo planteamiento...Y es que, de nuevo, aquí creemos que la ética y el valor que le damos a la protección y defensa del medioambiente también debe jugar un papel fundamental.

### 2.7. Séptima cuestión: ¿Qué medidas implementan para garantizar que su IA no sea hackeada o utilizada con fines maliciosos?

Si nos fijamos bien, vemos cómo en la propia pregunta se plantean dos realidades diferentes, pero íntimamente relacionadas con el binomio "IA-ciberataques". Por un lado, se pregunta qué medidas implementan las aplicaciones de IA para evitar ser víctimas de ciberataques y, por otro, se plantea si las aplicaciones de IA incorporan algún límite para evitar ser usadas con fines de ciberatacar a terceros.

En ambos planteamientos resulta fundamental incluir lo que podríamos denominar "ética por defecto y desde el diseño" en tanto en cuanto si bien es cierto que al igual que la seguridad total no existe, la ciberseguridad total tampoco existe.

Y, además, somos conscientes de que, por supuesto, no podemos "poner puertas al campo" y, hablando en términos muy coloquiales, "quien quiera ciberdelinquir, va a ciberdelinquir" de la misma manera que, por simplista que parezca el planteamiento, aunque matar esté prohibido y esté "mal en términos éticos", hay quien mata.

No obstante, de nuevo, no queremos caer en el error de que "como es imposible que no se use mal...no vamos a hacer nada", así como queríamos –y queremos– evitar a toda costa el error que comentábamos al comienzo del presente trabajo al decir que pese a que pudiera ser difícil o incluso imposible llegar a una "ética común", intentaríamos incluir una "ética mínima por defecto" para dar respuesta –siquiera parcial– a las cuestiones que se plantean aquí y que, repetimos, son solo una pequeña muestra de todo lo que se debe tener en cuenta en el ámbito de la inteligencia artificial.

Por tanto, volviendo a la cuestión que ahora nos ocupa relacionada con la ciberseguridad y, más concretamente, centrándonos en las dos sub-cuestiones que planteábamos, aportamos nuestro granito de arena a modo de cierre con dos ideas clave:

- Respecto a las medidas de seguridad que debe adoptar la IA para evitar ser víctima de un ciberataque, creemos que lo primero que hay que destacar es la necesidad de que las aplicaciones de IA –como toda aplicación donde se traten datos personales– han de cumplir con el principio de integridad y confidencialidad y, por tanto, han de adoptar todas las medidas de seguridad que sean necesarias para evitar cualquier acceso no autorizado.
- Respecto a las medidas de seguridad que han de adoptar los sistemas de IA para evitar que sean utilizados para ciberatacar a terceros, en nuestra opinión, la palabra clave es la transparencia, la información y la claridad en las condiciones de uso y en las penalizaciones a las que se podrían enfrentar en caso de utilizar los servicios con fines delictivos.

### 2.8. Octava cuestión. ¿Qué medidas toman para proteger a los menores de la exposición a IA que pueden influir negativamente en su desarrollo emocional o cognitivo?

Si bien hemos querido dejar la pregunta, tal cual la formuló la propia Inteligencia Artificial, por desgracia, la realidad es que podríamos preguntar cómo están preparadas las aplicaciones de IA para evitar que sus usuarios se suiciden –entiéndase– "por su culpa".

Y sí, sabemos que suena a ciencia ficción, pero ya está ocurriendo. Uno de los casos más recientes lo ha protagonizado un adolescente llamado Sewell Setzer quien, a través de Character,AI comenzó a interactuar con un chatbot que "asumía la identidad" de Daenerys Targaryen –personaje de una de las series más conocidas a nivel mundial "Juego de tronos". Pese a que el chatbot es, como su propio nombre indica, un chatbot –o, dicho con otras palabras, un robot–, la realidad es que la "relación" entre el adolescente y el chatbot se volvió cada vez más intensa. El adolescente –con problemas psicológicos– encontró en el chatbot a la confidente, novia, amiga y "ente" de máxima confianza, ya que no se puede decir "persona" de máxima confianza, porque no era así. Si bien es cierto que, como comentábamos, el adolescente ya tenía algunos problemas de salud mental y ya se había planteado quitarse la vida en alguna ocasión, la realidad es que "La gota que colmó el vaso" fue que cuando el adolescente preguntó al chatbot "¿Qué te parecería que pudiera ir a casa ahora mismo?" (Limón, 2024) –siendo "casa" un sinónimo de "muerte" puesto que el chico de catorce años consideraba que era el único lugar en el que podía encontrarse con "su amada"– el chatbot le respondió "Por favor, hazlo mi dulce rey". Llegó la trágica noticia de la consumación del suicidio y su madre ya ha anunciado que ha emprendido acciones legales contra los creadores de la aplicación de Character.AI.

No podemos entrar aquí a valorar, juzgar o tratar de adivinar "¿qué hubiera pasado si...?" puesto que no lo sabemos y no lo podremos saber nunca. Pero lo innegable es que se han de articular los mecanismos necesarios para evitar que pueda volver a pasar. Y es que, de nuevo, se pone de manifiesto la incapacidad de la IA de empatizar, de tener relaciones "reales" pero eso impide que se le exija la adopción de todas las medidas que sean necesarias para evitar el daño que se puede causar a todos los niveles. Y es que no hablamos solo de causar la muerte –puesto que es la punta del iceberg y la consumación de lo peor que puede pasar– sino también de todo el daño que, a nivel emocional, psicológico y cognitivo, puede llegar a causar.

En concreto, la aplicación de Charcater.AI ya ha anunciado la adopción de medidas para evitar que puedan volver a suceder tragedias de este calibre, entre las que cabe citar: inclusión de recordatorios frecuentes insistiendo en que el personaje no es real y la aparición de ventanas emergentes de ayuda con recursos para la prevención del suicidio cuando desde la propia aplicación se detecten indicios de suicidio. Tal vez no son suficientes, pero son un buen punto de partida. Y, desde luego, nos han de hacer caer en la cuenta de que la IA no es un juego, ni es un "Buscador mejorado y avanzado" ... Es una herramienta mucho más potente y en la que la ética –desde el diseño y por defecto– y el cumplimiento normativo –desde el diseño y por defecto– han de jugar un papel fundamental.

## 3. A modo de conclusión ¿Hemos "creado" una auditoría para verificar la ética de la IA?

Queremos dedicar estas últimas palabras a valorar lo comentado hasta ahora y a hacer hincapié en lo que son las ocho ideas clave que responden a cada una de las ocho cuestiones que nos hemos planteado.

Pero antes de entrar en ellas, recordamos que entrecomillábamos el término "creando" al plasmarlo en el título del presente artículo puesto que el objetivo de un trabajo tan limitado en el espacio no es, ni mucho menos, tan pretencioso como el de "crear" o, mejor dicho, "dejar creado" un protocolo de auditoría para verificar algo tan sumamente complejo como la ética en los sistemas de inteligencia artificial sino que, precisamente, hemos utilizado el gerundio y lo hemos entrecomillado con el ánimo de plasmar la dificultad para llegar a una "respuesta única, uniforme, total y absoluta" y dejar entrever, así, que se trata de un proceso vivo, complejo y en constante cambio pero que, no por ello ni por la dificultad de alcanzar la respuesta a los interrogantes planteados aquí –y a los no planteados– debemos rendirnos y "dar rienda suelta". Ni mucho menos. La ética es y ha de ser una cuestión absolutamente protagonista en todo lo que está relacionado con la inteligencia artificial.

De este modo, queremos concluir el presente trabajo, recordando la cuestión planteada y la idea fundamental sobre la que animamos al lector a reflexionar, al legislador a regular y a los usuarios y responsables de la IA a implementar:

> A la primera cuestión *¿Cómo garantizan que los datos que utilizan para entrenar la IA están libres de sesgos que puedan causar discriminación?,* la clave la encontramos en las fuentes de las que bebe.
>
> A la segunda cuestión *¿Qué mecanismos tienen para que los usuarios afectados por decisiones de la IA puedan apelar decisiones sesgadas?,* la clave la encontramos en la transparencia en los criterios utilizados y en la posibilidad de acudir a revisión humana.
>
> A la tercera cuestión *¿Qué medidas implementan para evitar que los modelos generen contenido ofensivo o dañino?,* la clave la encontramos en la manera en la que aprende.
>
> A la cuarta cuestión *¿Cómo gestionan el uso de IA para prevenir su abuso en la creación de deepfakes o la manipulación de imágenes y videos?,* la clave la encontramos en formar en valores como la privacidad, la intimidad y la responsabilidad ante el uso de datos de terceros.
>
> A la quinta cuestión *¿Cómo aseguran que los datos utilizados para entrenar la IA cumplen con las normativas de protección de datos como el GDPR?,* la clave la encontramos en cumplir la normativa de protección de datos en todo tratamiento de datos personales que se lleve a cabo.
>
> A la sexta cuestión *¿Cómo manejan el impacto ambiental del entrenamiento de sus modelos de IA?,* la clave la encontramos en hacer uso de energías renovables e implementar medidas que pongan la protección del medioambiente como cuestión clave al desarrollar sistemas de IA.
>
> A la séptima cuestión *¿Qué medidas implementan para garantizar que su IA no sea hackeada o utilizada con fines maliciosos?,* la clave la encontramos en adoptar medidas de seguridad conforme al principio de integridad y confidencialidad regulado en el RGPD y formar sobre el valor de la privacidad, la intimidad y la responsabilidad legal.
>
> A la octava cuestión *¿Qué medidas toman para proteger a los menores de la exposición a IA que puede influir negativamente en su desarrollo emocional o cognitivo?,* la clave la encontramos en la transparencia, claridad y veracidad de lo falso de las interacciones de la IA.

Por tanto, la creación de un protocolo de auditoría para verificar la ética de la Inteligencia artificial es fundamental puesto que, si bien no ponemos en duda las bondades de la IA a nivel de producción, mejora del rendimiento y agilidad y eficacia en determinadas tareas, ello no obsta para que seamos conscientes –y absolutos promotores– de que las decisiones que adopta la IA han de estar "tintadas" de un mínimo de cuestiones éticas que permitan garantizar la protección y defensa de los derechos humanos de todos los usuarios que acuden a la inteligencia artificial. Por ello, con este trabajo hemos querido dar el primer paso para la elaboración de este protocolo que permita auditar las decisiones adoptadas por la IA para que el foco esté, en todo momento, en que los derechos humanos sean garantizados.

## 4. Referencias


Agencia Española de Protección de Datos. (2024). Adecuación al RGPD en IA. https://www.aepd.es/sites/default/files/2020-02/adecuacion-rgpd-ia.pdf

Bencina, J. (2024). "El lado oscuro de la innovación: Inteligencia artificial generativa y sus efectos en el medio ambiente". Ladera Sur, 17 de julio. https://laderasur.com/articulo/el-lado-oscuro-de-la-innovacion-inteligencia-artificial-generativa-y-sus-efectos-en-el-medio-ambiente/

Cangelosi, H. (2024). "Huella ecológica e impacto ambiental de la inteligencia artificial". Carbono News, 19 de octubre. https://www.carbono.news/ciudades-inteligentes/huella-ecologica-e-impacto-ambiental-de-la-inteligencia-artificial/

Dastin, J. (2018). "Amazon scraps secret AI recruiting tool that showed bias against women". Reuters, 11 de octubre. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

Davara Fernández de Marcos, L. (2023). *Memento de Protección de datos*. Lefebvre.

Davara Fernández de Marcos, L. (2024). 2ª edición –ampliada, revisada y actualizada– del "Libro definitivo sobre Redes Sociales: Claves para padres y educadores". *Cuadernos de pedagogía*. Wolters Kluwer.

Davara Fernández de Marcos, L. (2024). *El móvil que todo lo sabía – cuentos para familias en la era digital.* Dykinson.

Fundación Fepropaz. (2024). "El impacto ambiental de la inteligencia artificial", 14 de mayo. https://fepropaz.com/el-impacto-ambiental-de-la-inteligencia-artificial/

Funds Society (2024). "Inteligencia artificial: ¿Cuál es su impacto medioambiental?", 7 de febrero. https://www.fundssociety.com/es/opinion/inteligencia-artificial-cual-es-su-impacto-medioambiental/

Garay, J. (2024). "Elecciones en México: Deepfakes, clonación de voz y descontextualización, la desinformación en el proceso electoral". Wired, 29 de mayo. https://es.wired.com/articulos/elecciones-en-mexico-deepfakes-clonacion-de-voz-y-descontextualizacion-la-desinformacion-en-el-proceso-electoral

García, J. M. (2019). "Inteligencia artificial: Impacto ambiental mayor de lo que se creía", 17 de junio. https://www.lavanguardia.com/tecnologia/innovacion/20190617/462863973194/inteligencia-artificial-impacto-ambiental-mayor-creia.html

Gerry, S. (2024). "Inteligencia artificial: Impactos ambientales en América Latina. Mongabay", 17 de abril. https://es.mongabay.com/2024/04/inteligencia-artificial-impactos-ambientales-america-latina/

González, F. (2024). "Más de 400 expertos firman carta contra los deepfakes". Wired, 21 de febrero. https://es.wired.com/articulos/mas-de-400-expertos-firman-carta-contra-los-deepfakes

Hernando, P. (2024a). "Deepfakes de famosos: Desde Taylor Swift sorteando sartenes a Leo Messi pidiendo que descargues su app". La Vanguardia, 10 de agosto. https://www.lavanguardia.com/andro4all/tecnologia/deepfakes-de-famosos-desde-taylor-swift-sorteando-sartenes-a-leo-messi-pidiendo-que-descargues-su-app

Hernando, P. (2024b). "Una técnica usada en astronomía puede ser usada para identificar deepfakes". La Vanguardia, 28 de julio. https://www.lavanguardia.com/andro4all/tecnologia/una-tecnica-usada-en-astronomia-puede-ser-usada-para-identificar-deepfakes

International Energy Agency (2024). *Electricity 2024: Analysis and forecast to 2026*. https://iea.blob.core.windows.net/assets/6b2fd954-2017-408e-bf08-952fdd62118a/Electricity2024-Analysisandforecastto2026.pdf

La Vanguardia (2024). "Estafa con múltiples deepfakes: 24 millones de euros a una persona real", 5 de febrero. https://www.lavanguardia.com/tecnologia/ciberseguridad/20240205/9513300/estafa-multiples-deepfakes-24-millones-euros-persona-real.html

Limón, R. (2024). "Un adolescente se suicida en EE.UU. tras enamorarse de un personaje creado con IA". El País, 24 de octubre. https://elpais.com/tecnologia/2024-10-24/un-adolescente-se-suicida-en-ee-uu-tras-enamorarse-de-un-personaje-creado-con-ia.html

Nintenderos (2024). "Inglaterra y Gales prohibirán la creación de contenido para adultos deepfake sin consentimiento", 3 de diciembre. https://www.nintenderos.com/ninteocio/inglaterra-y-gales-prohibiran-la-creacion-de-contenido-para-adultos-deepfake-sin-consentimiento/

Pastor, J. (2024). "El CEO de Ferrari ha sido la última víctima de deepfakes para estafas a empresa: Pudo salirle muy caro". Xataka, 29 de julio. https://www.xataka.com/robotica-e-ia/ceo-ferrari-ha-sido-ultima-victima-deepfakes-para-estafas-a-empresa-pudo-salirle-muy-caro

Pérez, E. (2024). "Deepfakes sexuales: No pararán solo con castigar su distribución, hay que fijarse en la pionera ley de Corea del Sur". Xataka, 24 de octubre. https://www.xataka.com/legislacion-y-derechos/deepfakes-sexuales-no-pararan-solo-castigar-su-distribucion-hay-que-fijarse-pionera-ley-corea-sur

Reyes, E. (2024). "El uso de deepfakes se cuadriplica: Lo potencia el uso de redes sociales". Expansión, 28 de noviembre. https://expansion.mx/tecnologia/2024/11/28/el-uso-de-deepfakes-se-cuadriplica-lo-potencia-el-uso-de-redes-sociales

Technology Review (2024). "Caso práctico: Probamos por qué un algoritmo judicial justo es imposible". https://www.technologyreview.es/s/13800/caso-practico-probamos-por-que-un-algoritmo-judicial-justo-es-imposible