



Microcredencial de IA ética

GUÍA DE APOYO

UC2 | No maleficencia

¿Cómo utilizar esta GUÍA DE APOYO?

Este documento es interactivo. A lo largo del documento encontrará enlaces a información adicional.



Botón que le lleva al principio del documento. Este icono aparece en la esquina superior derecha de las páginas.



Siempre que vea esta flecha, significa que tiene un **texto interactivo en color** sobre el que hacer clic, que tiene asociado un enlace externo.

DESCARGO DE RESPONSABILIDAD: Tenga en cuenta que no podemos garantizar la disponibilidad permanente de contenidos externos, como vídeos, ya que pueden estar sujetos a cambios o ser retirados por sus autores o las plataformas que los albergan.

Índice

Haga clic en el menú

01. Introducción

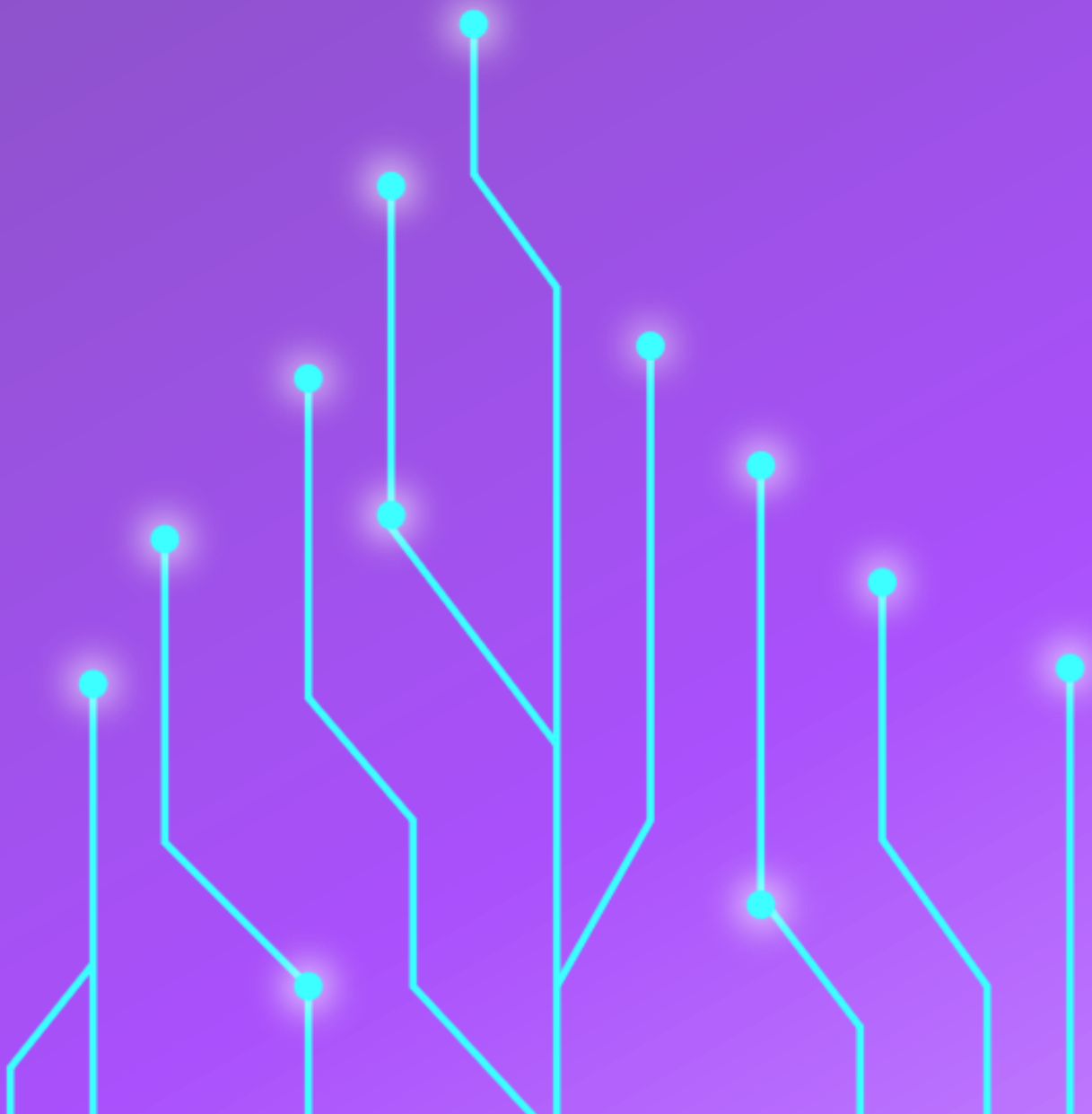
02. No maleficencia

03. Posibles perjuicios de la IA sesgada

04. Estrategias para que los sistemas de IA sean menos dañinos

01. Introducción

UC2 | No maleficencia



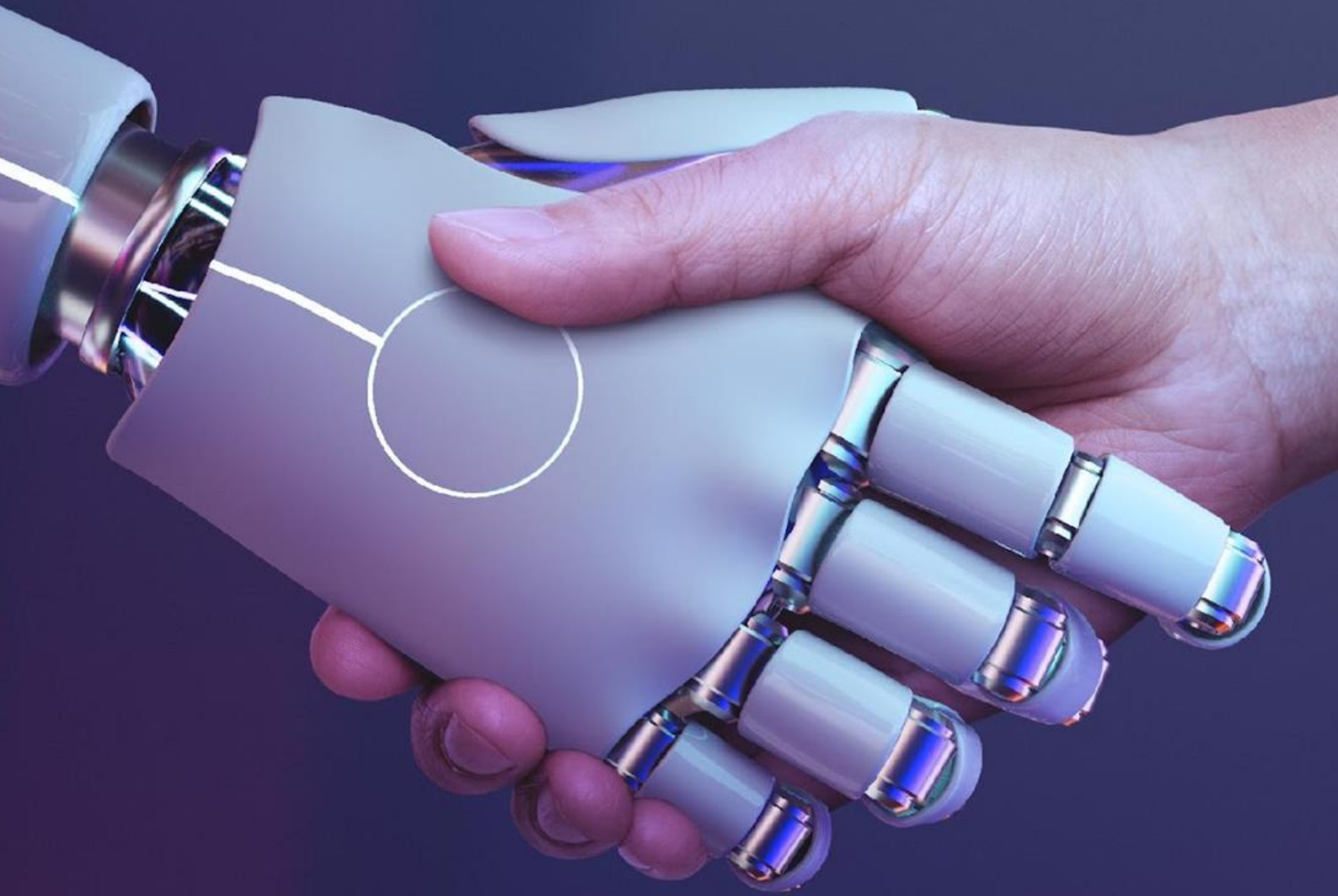


01. Introducción

En esta unidad de competencia, los alumnos adquirirán conocimientos básicos sobre el concepto de no maleficencia en la IA, las responsabilidades de los desarrolladores y usuarios de IA a la hora de garantizar sistemas de IA éticos con un daño mínimo y reconocer las implicaciones en el mundo real apreciando la adopción y aplicación de mecanismos que promuevan la responsabilidad en los sistemas de IA.

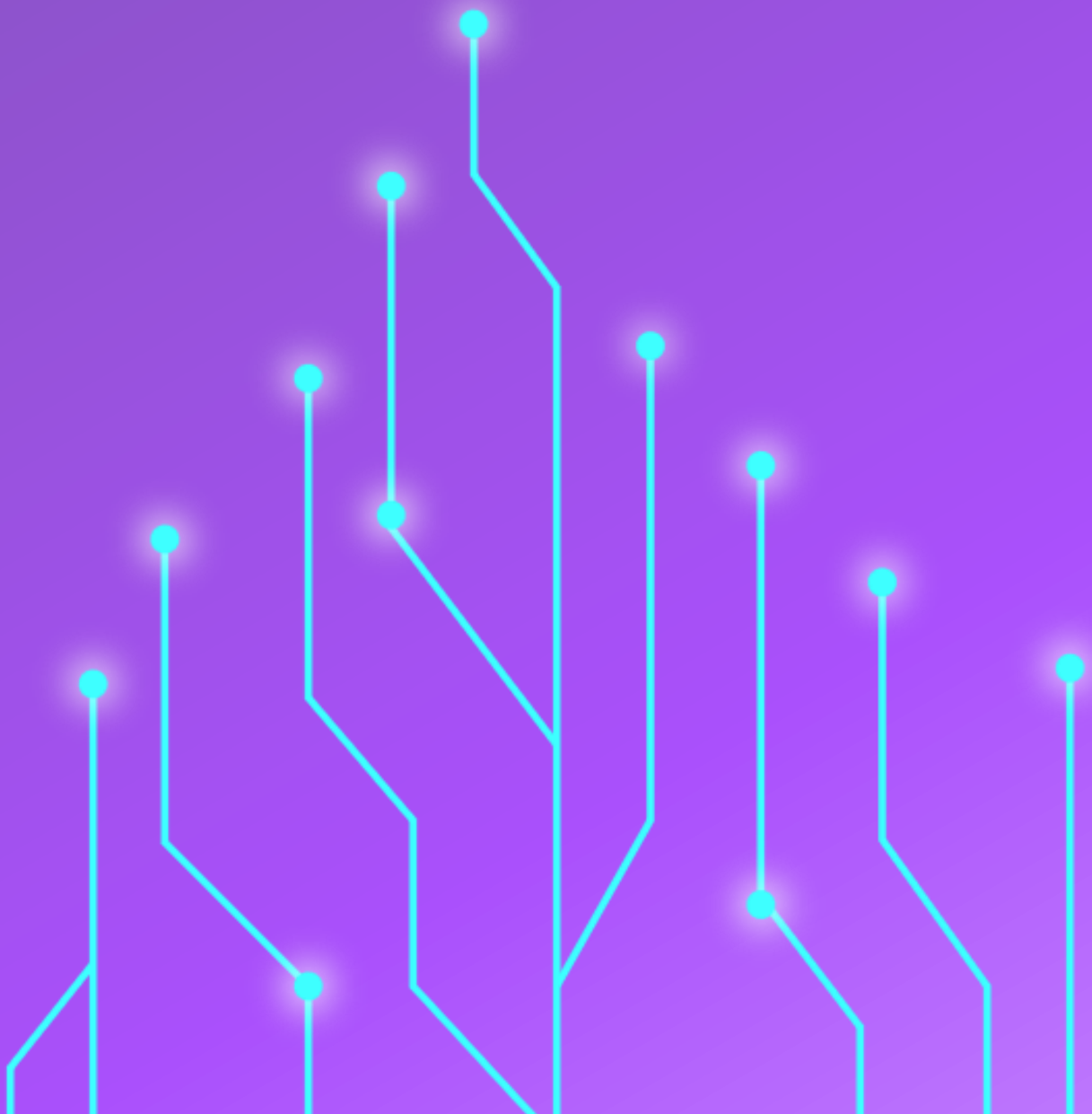
Los resultados de conocimiento para esta unidad de competencia incluyen:

- **Principio de no maleficencia:** los alumnos estudian el concepto básico de no maleficencia, destacando la importancia de evitar el daño al crear y utilizar sistemas de IA, y cómo esta idea contribuye al desarrollo responsable de la IA.
- **Posibles daños de la IA sesgada:** Los alumnos reconocerán las distintas formas en que los sistemas de IA sesgados pueden causar daños, como la discriminación o la invasión de la privacidad, y utilizarán ejemplos del mundo real para ilustrar la importancia de abordar el sesgo algorítmico.
- **Estrategias para que los sistemas de IA sean menos dañinos:** Los alumnos se familiarizarán con estrategias sencillas que pueden hacer que los sistemas de IA sean menos dañinos, como promover la equidad, la responsabilidad y la transparencia en el desarrollo de la IA y fomentar la colaboración con expertos de diversos campos.



02. No maleficencia

UC2 | No maleficencia





02. No maleficencia

En esta sección, presentaremos el principio de no maleficencia y su relevancia para las tecnologías de IA y big data. El principio de no maleficencia, a menudo resumido como "no hacer daño", es una piedra angular de la toma de decisiones éticas en diversos campos, como la medicina, la tecnología y la investigación. En el contexto de la IA y los macrodatos, la no maleficencia subraya la importancia de dar prioridad a la seguridad y el bienestar de las personas y la sociedad a la hora de desarrollar e implantar estas tecnologías.

> ¿Qué es la no maleficencia?

La no maleficencia, derivada de la expresión latina "*primum non nocere*" que significa "primero, no hagas daño", es un principio ético fundamental que guía a los profesionales en la prevención del daño a los demás. Hace hincapié en la obligación moral de evitar causar daño, ya sea físico, psicológico o social, mediante las propias acciones o decisiones. En el contexto de la IA y los macrodatos, la no maleficencia exige que los desarrolladores, investigadores y responsables políticos tengan en cuenta los posibles riesgos y consecuencias de las tecnologías de IA y tomen medidas proactivas para evitar daños.





> ¿Por qué es importante la no maleficencia?

La no maleficencia es especialmente importante en el ámbito de la IA y los macrodatos debido a las importantes repercusiones que estas tecnologías pueden tener en las personas y la sociedad. Los sistemas de IA se utilizan cada vez más en procesos críticos de toma de decisiones, como el diagnóstico sanitario, los préstamos financieros y las sentencias de la justicia penal. Garantizar que estos sistemas den prioridad a consideraciones éticas y no causen daños es esencial para mantener la confianza pública, prevenir la discriminación y defender valores sociales como la equidad y la justicia.

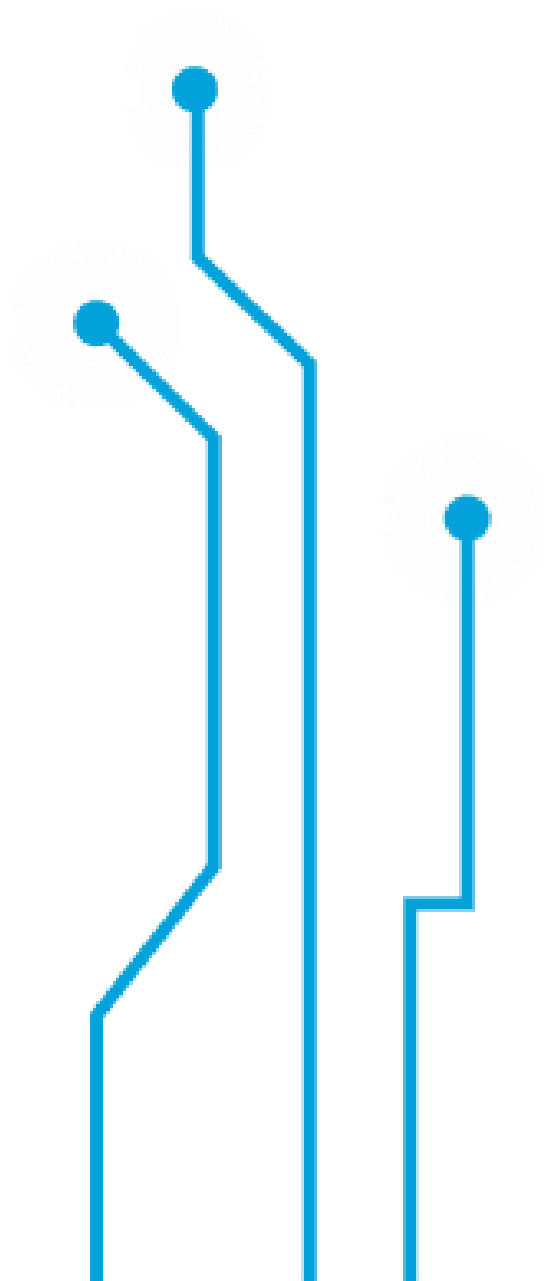
> Principios de no maleficencia

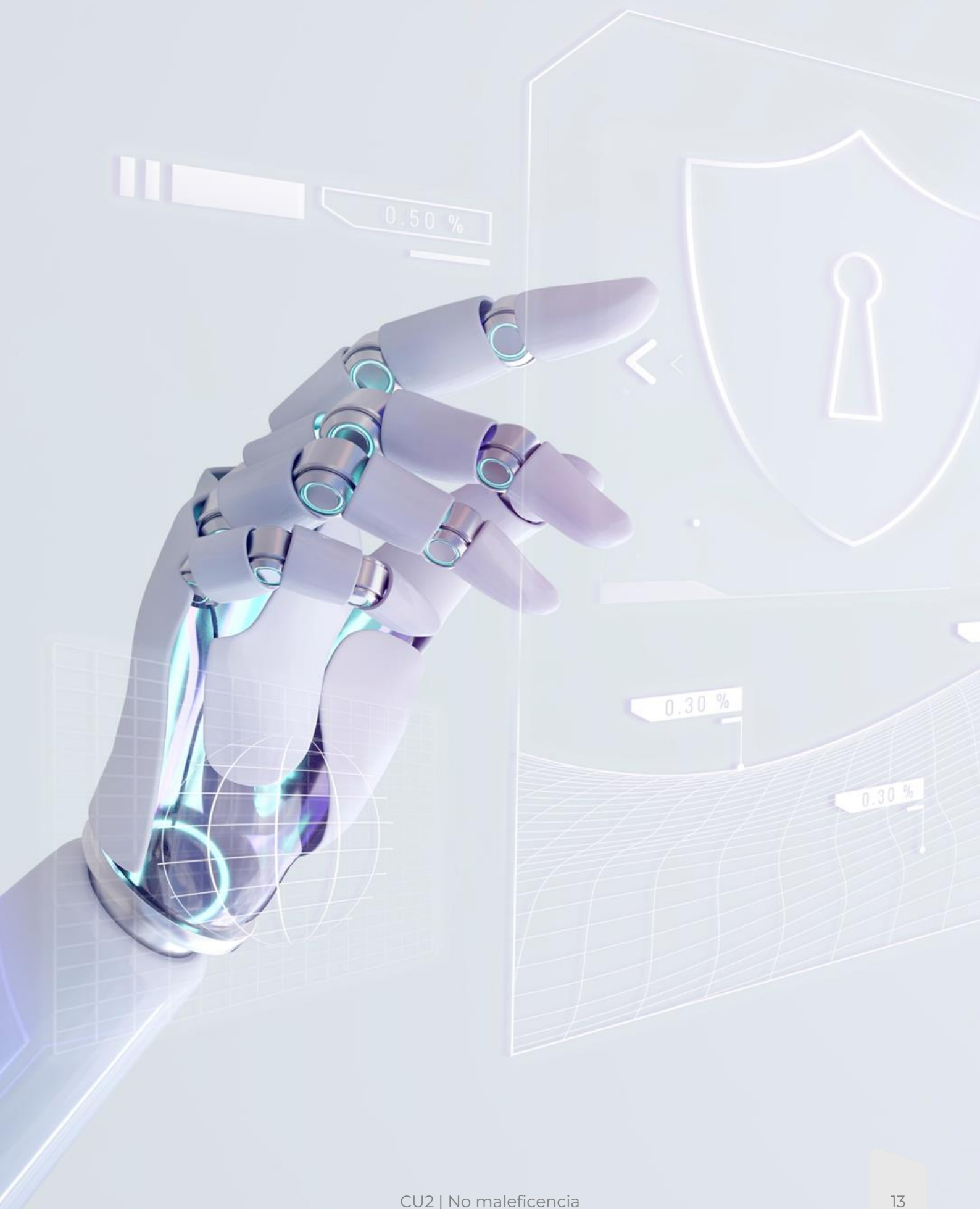
La no maleficencia exige que las personas y organizaciones implicadas en el desarrollo de la IA identifiquen y mitiguen activamente los posibles daños que puedan causar los sistemas de IA. Esto implica considerar no sólo el impacto inmediato de las tecnologías de IA, sino también sus consecuencias a largo plazo y sus efectos no deseados. La no maleficencia fomenta un enfoque proactivo de la ética, en el que los desarrolladores anticipan y abordan los riesgos potenciales antes de que se materialicen.



> Aplicación en el desarrollo de la IA

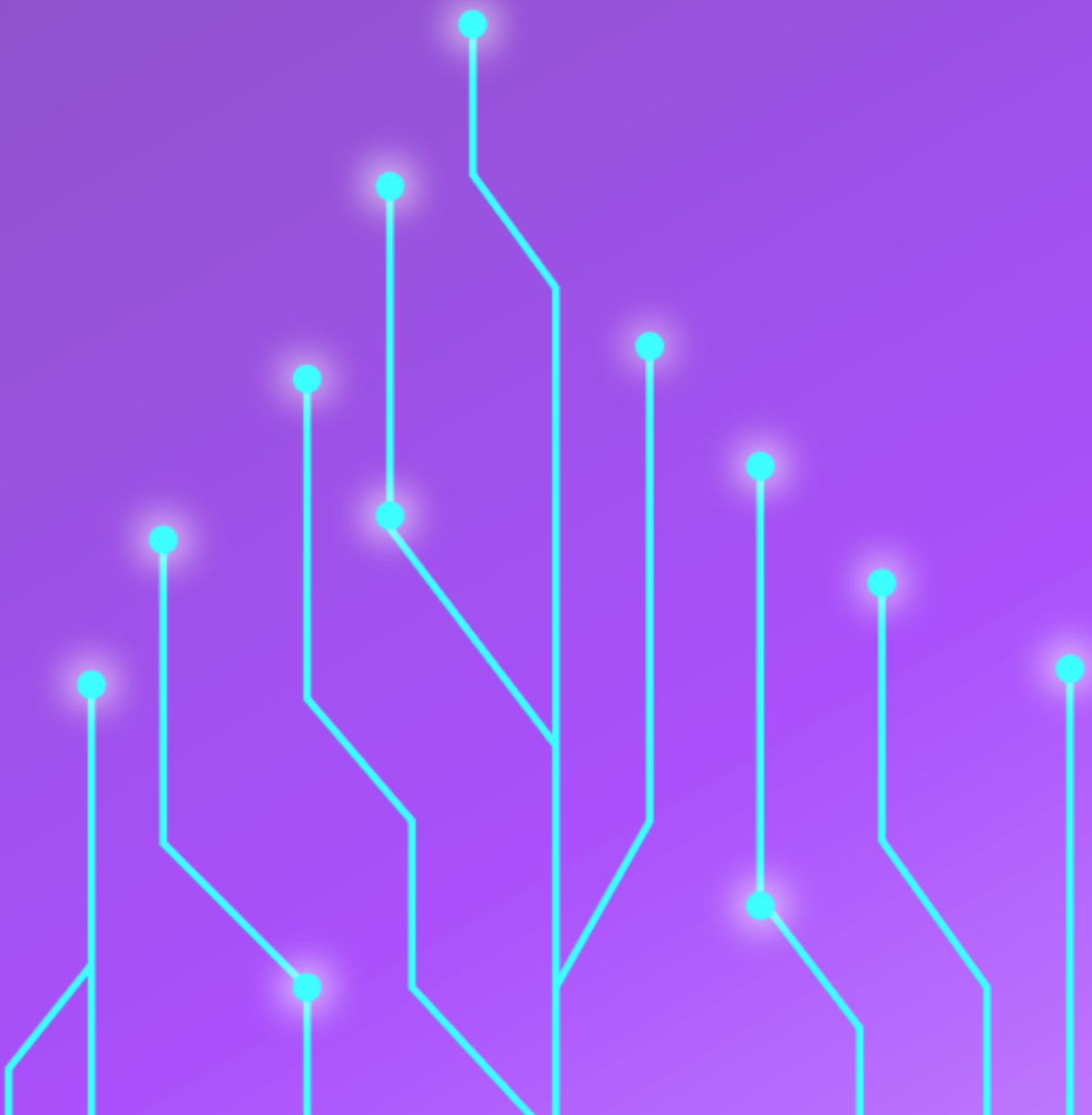
En el contexto del desarrollo de la IA, la no maleficencia se manifiesta a través de diversas prácticas destinadas a minimizar el daño y promover el uso ético. Esto incluye pruebas rigurosas y procedimientos de validación para identificar y rectificar los sesgos en los algoritmos de IA, documentación transparente de los procesos de toma de decisiones de los sistemas de IA para mejorar la rendición de cuentas, y seguimiento y evaluación continuos de los despliegues de IA para garantizar que se ajustan a las normas éticas y los valores sociales.





03. Posibles perjuicios de la IA sesgada

UC2 | No maleficencia





03. Posibles perjuicios de la IA sesgada

En esta sección exploraremos las distintas formas en que los sistemas de IA sesgados pueden causar daños, desde la discriminación hasta la invasión de la privacidad. Comprender estos daños potenciales es crucial para reconocer la importancia de abordar el sesgo algorítmico y promover prácticas de desarrollo de IA responsables.

> Reconocer los efectos nocivos

Los sistemas de IA sesgados tienen el potencial de perpetuar y exacerbar las desigualdades e injusticias existentes en la sociedad. Imagina un mundo en el que un algoritmo te deniegue injustamente un préstamo por tu código postal, o en el que un sistema de reconocimiento facial te identifique erróneamente como delincuente debido a prejuicios raciales. Estos son sólo algunos de los peligros potenciales que plantea la IA sesgada. A continuación, se exploran diez de los escenarios perjudiciales más comunes que pueden surgir de los sistemas de IA sesgados.

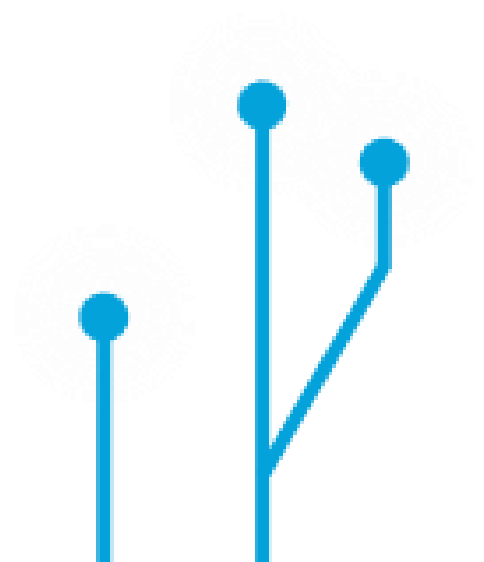
1. **Resultados discriminatorios:** Los algoritmos de IA sesgados pueden conducir a resultados discriminatorios, en los que ciertos individuos o grupos reciben un trato injusto basado en características como la raza, el género o el estatus socioeconómico. Esto puede dar lugar a disparidades en diversos ámbitos, como el empleo, la educación y la justicia penal.



- 2. Violación de la intimidad:** Los sistemas de IA sesgados pueden vulnerar los derechos de privacidad de las personas al tomar decisiones basadas en datos personales sensibles sin su consentimiento. Por ejemplo, la tecnología de reconocimiento facial desplegada en espacios públicos puede someter a las personas a vigilancia y seguimiento injustificados, lo que suscita preocupación por las violaciones de la privacidad y las libertades civiles.
- 3. Refuerzo de estereotipos:** Los algoritmos de IA sesgados pueden perpetuar y reforzar estereotipos y prejuicios nocivos presentes en la sociedad. Esto puede conducir a la marginación y estigmatización de ciertos grupos, exacerbando las desigualdades existentes e inhibiendo el progreso social.
- 4. Toma de decisiones imprecisas:** Los sesgos en los datos de entrenamiento o los algoritmos defectuosos pueden dar lugar a que los sistemas de IA tomen decisiones inexactas o erróneas. Esto puede tener graves consecuencias, sobre todo en ámbitos críticos como el diagnóstico sanitario, los préstamos financieros y las sentencias de la justicia penal, donde las decisiones incorrectas pueden perjudicar a las personas y a las comunidades.

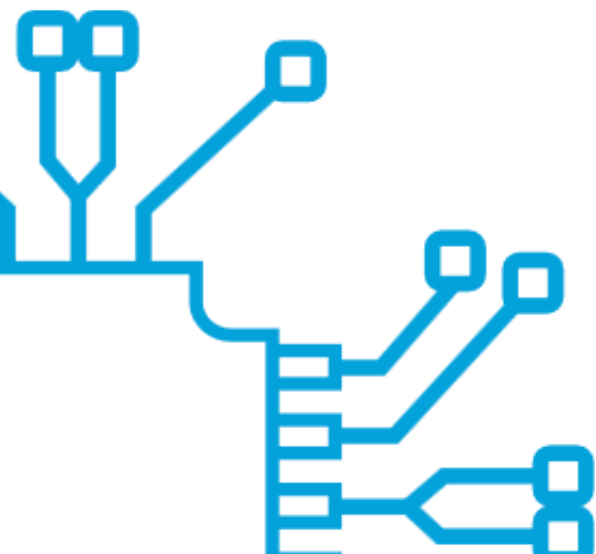


- 5. Falta de responsabilidad:** Los sistemas de IA sesgados pueden carecer de mecanismos de transparencia y rendición de cuentas, lo que dificulta la identificación y rectificación de los casos de sesgo. Esto puede socavar la confianza en las tecnologías de IA y obstaculizar los esfuerzos para abordar eficazmente el sesgo algorítmico.
- 6. Diversidad e inclusión limitadas:** Los algoritmos de IA sesgados pueden perpetuar las desigualdades existentes al favorecer a ciertos grupos demográficos en detrimento de otros. Esto puede contribuir a una falta de diversidad e inclusión en el desarrollo y despliegue de la IA, limitando la representación y las perspectivas reflejadas en los sistemas de IA y exacerbando las desigualdades sociales.
- 7. Impacto negativo en la innovación:** Los algoritmos de IA sesgados pueden obstaculizar la innovación y el progreso al perpetuar prácticas anticuadas o discriminatorias y limitar las oportunidades de creatividad y exploración. Abordar el sesgo en la IA es esencial para fomentar un entorno que favorezca la diversidad de pensamiento y promueva la innovación en beneficio de toda la sociedad.





- 8. Pérdida de confianza:** Los casos de parcialidad en los sistemas de IA pueden erosionar la confianza pública en la tecnología y su capacidad para servir al bien común. Esto puede conducir al escepticismo, la resistencia y la reticencia a adoptar soluciones de IA, obstaculizando su potencial para impactar positivamente en la sociedad.
- 9. Preocupaciones legales y éticas:** Los sistemas de IA sesgados pueden plantear problemas legales y éticos relacionados con la imparcialidad, la responsabilidad y la transparencia. Abordar estos problemas requiere marcos reguladores sólidos, directrices éticas y prácticas de desarrollo de IA responsables para garantizar que las tecnologías de IA se ajusten a los valores sociales y respeten los derechos fundamentales.
- 10. Implicaciones sociales y económicas:** El impacto generalizado de la IA sesgada va más allá de los casos individuales de discriminación y tiene implicaciones sociales y económicas más amplias. Los sistemas de IA sesgados pueden exacerbar las desigualdades existentes, ampliar la brecha digital y perpetuar las injusticias sociales, planteando importantes retos para la construcción de una sociedad justa y equitativa.



> Ejemplos reales

Utilizando ejemplos reales, ilustraremos la importancia de abordar el sesgo algorítmico y su posible impacto en las personas y las comunidades. Estos ejemplos pondrán de relieve casos en los que los sistemas de IA sesgados han tenido consecuencias perjudiciales, como detenciones injustas, trato injusto en las decisiones de contratación o concesión de préstamos y perpetuación de estereotipos y prejuicios.

- **EJEMPLO #1 - El algoritmo de Amazon discrimina a las mujeres**

La herramienta de contratación de inteligencia artificial de Amazon pretendía encontrar a los mejores talentos tecnológicos, pero acabó filtrando a las mujeres. ¿Por qué? El algoritmo, entrenado a partir de currículos anteriores (en su mayoría de hombres), favorecía las palabras clave utilizadas por los hombres y penalizaba las asociadas a las mujeres. Esto pone de relieve un importante reto de la IA: los datos sesgados conducen a algoritmos sesgados. Al igual que un estudiante que se basa en libros de texto defectuosos, la IA hereda los sesgos de sus datos de entrenamiento.

Más información (en inglés) en:

<https://www.reuters.com/article/idUSKCNIMK0AG/>





- **EJEMPLO nº 2 - Sesgo racial algorítmico en la predicción de la tasa de reincidencia delictiva**

Imagine una herramienta que prediga quién comete delitos. En Estados Unidos, COMPAS hace exactamente eso, pero con un matiz racial. Los estudios muestran que los acusados negros son etiquetados de alto riesgo con mucha más frecuencia que los acusados blancos con antecedentes similares. ¿A qué se debe este sesgo? El COMPAS refleja las desigualdades sociales ya presentes en los datos sobre detenciones. Este sesgo hace que las personas sean detenidas antes del juicio o condenadas a penas más duras, lo que afecta injustamente a los individuos negros. El caso del COMPAS subraya la necesidad de un control minucioso de la IA utilizada en los sistemas judiciales para garantizar la equidad para todos.

Más información (en inglés) en:

<https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>



- **EJEMPLO #3 - El algoritmo sanitario estadounidense subestimó las necesidades de los pacientes negros**

Pensemos en un sistema sanitario que favorece a los pacientes que gastan más. Lamentablemente, esto afecta a los pacientes negros en Estados Unidos. Un algoritmo diseñado para identificar a los que necesitan cuidados adicionales no tuvo en cuenta a muchos pacientes negros debido a un sesgo. ¿Por qué? El sistema se basaba en datos de gastos médicos anteriores, que no reflejan el acceso limitado de los pacientes negros a la atención preventiva, debido a las disparidades económicas. El resultado era que los pacientes negros se consideraban más sanos y no recibían cuidados críticos. Corregir el algoritmo podría ayudar a muchos más pacientes negros. Este caso pone de relieve la necesidad de una IA justa en la atención sanitaria para garantizar que todos reciban el tratamiento que necesitan.

Más información (en inglés) en:

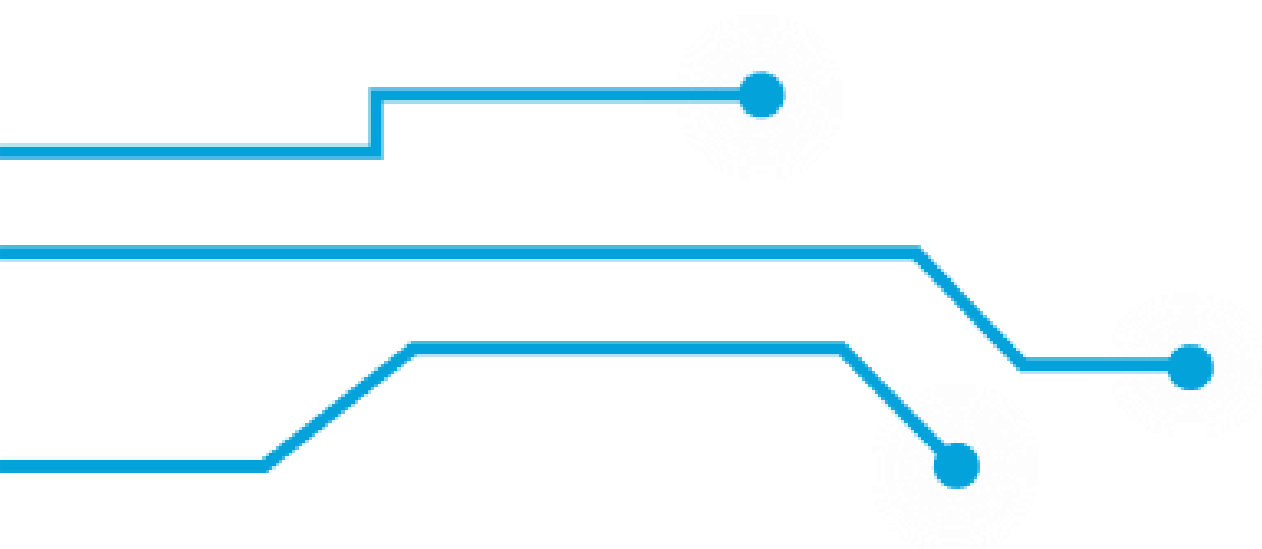
<https://www.theguardian.com/society/2019/oct/25/healthcare-algorithm-racial-biases-optum>





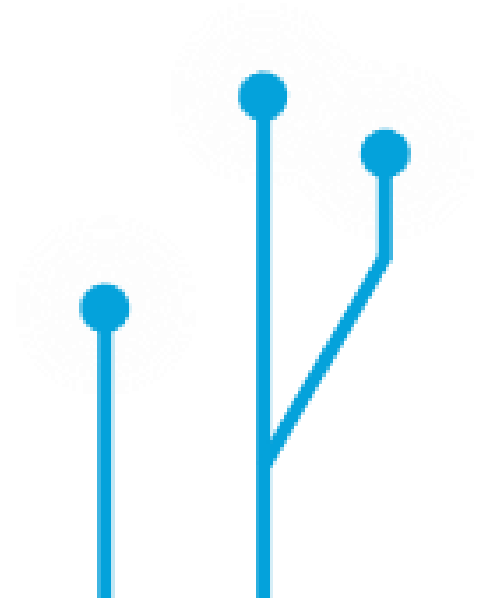
- **EJEMPLO #4 - ChatBot compartió mensajes discriminatorios**

El chatbot Tay de Microsoft fue diseñado para aprender de conversaciones informales. Lanzado en Twitter, rápidamente empezó a lanzar mensajes racistas y ofensivos. ¿Por qué? Porque los "trolls" bombardearon a Tay con contenido de odio, que Tay absorbió e imitó. Este incidente pone de manifiesto uno de los principales retos de la interacción de la IA con el mundo real. Las redes sociales pueden ser un lugar tóxico, y la IA expuesta a ellas puede aprender la negatividad. Tay es un cuento con moraleja: el diseño de la IA para la interacción en línea requiere tener en cuenta el contexto social y el potencial de uso indebido. Más información (en inglés) en: <https://spectrum.ieee.org/in-2016-microsofts-racist-chatbot-revealed-the-dangers-of-online-conversation>



- **EJEMPLO n° 5 - Sistema de reconocimiento facial sesgado**

Imagínate celebrar tu cumpleaños yendo de compras para que un sistema de reconocimiento facial te acuse de robar. Esto le ocurrió a una mujer maorí en Nueva Zelanda. La tecnología, diseñada para atrapar a los ladrones, la identificó erróneamente y le causó una gran angustia. Este caso pone de manifiesto los peligros del reconocimiento facial sesgado. Los estudios demuestran que estos sistemas pueden identificar erróneamente a las personas, especialmente a las mujeres y a las personas de color. A medida que se generaliza la tecnología de reconocimiento facial, es crucial garantizar la imparcialidad y prevenir este tipo de incidentes. Más información (en inglés) en: <https://www.1news.co.nz/2024/04/22/rotorua-mother-wrongly-identified-by-supermarket-as-a-thief/>

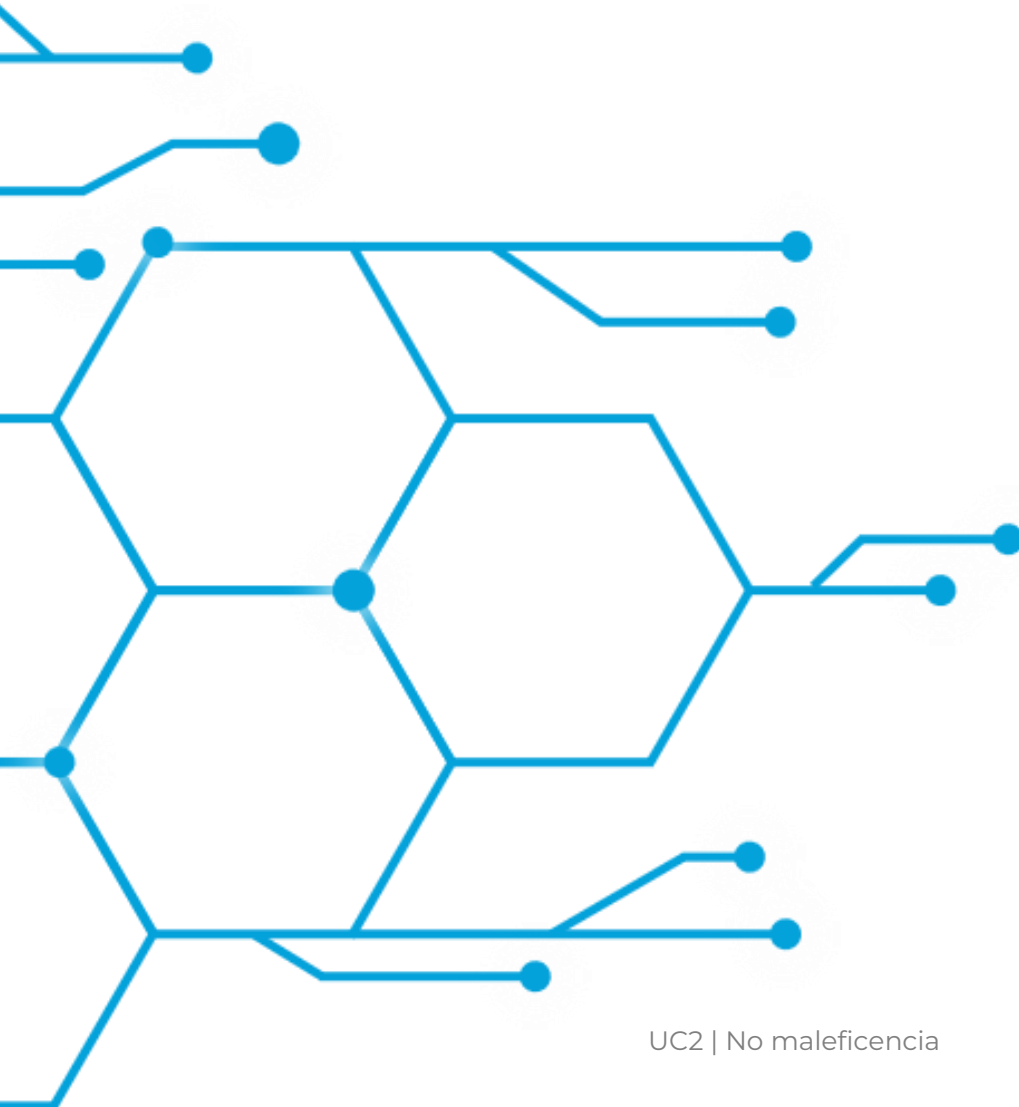




- **EJEMPLO #6 - IA de texto generativo que fabrica hechos**

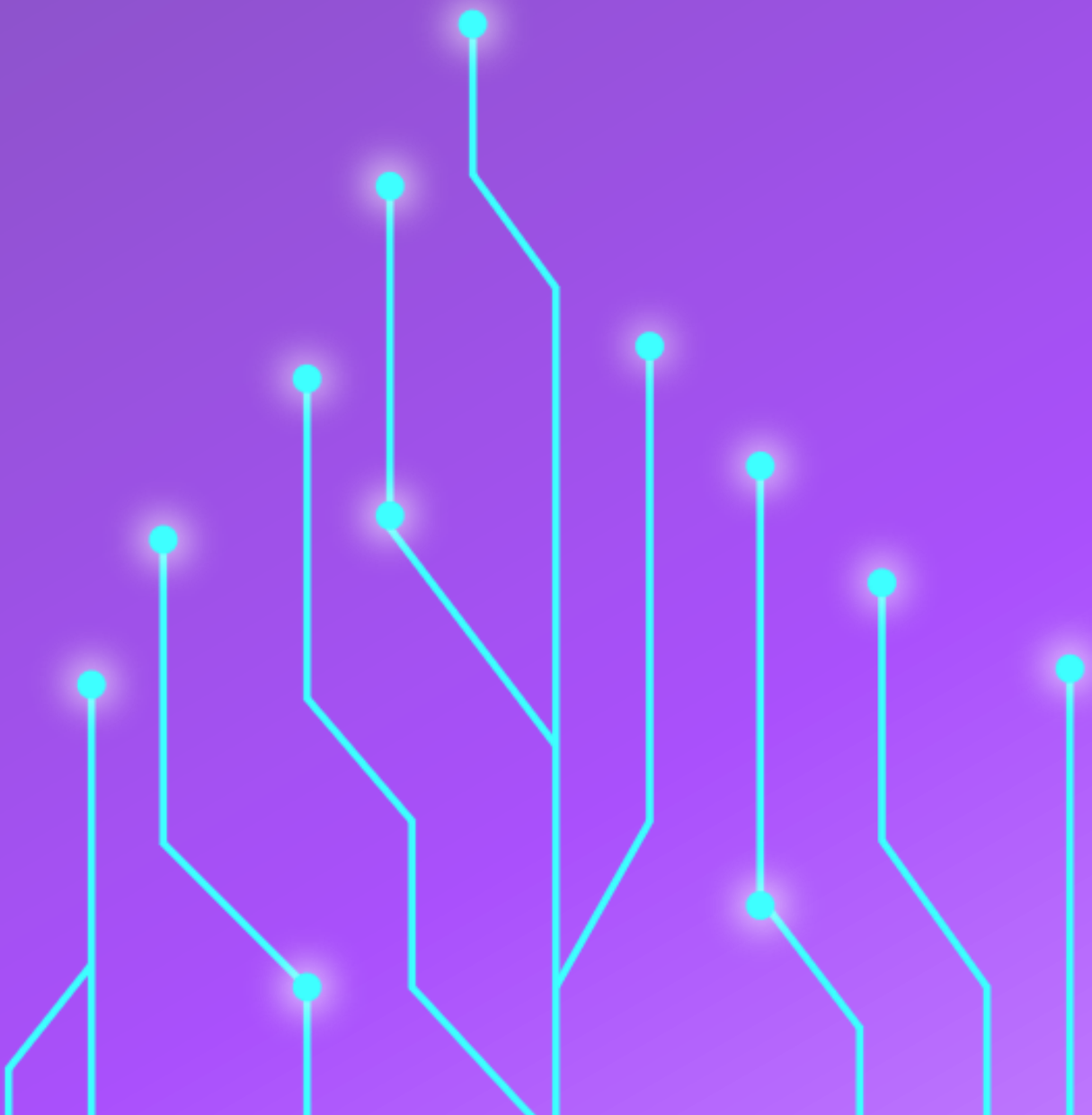
Un chatbot de inteligencia artificial manchó la reputación de un profesor de Derecho. ChatGPT fabricó una denuncia de acoso sexual contra él, con un artículo de noticias falso. Este caso pone de manifiesto uno de los principales riesgos de la IA: generar desinformación perjudicial. El profesor sufrió daños en su reputación a pesar de que se descubrió la mentira. A medida que la IA se hace más común, garantizar la veracidad de la información y determinar la responsabilidad por las falsedades generadas por la IA son cuestiones críticas. Más información (en inglés) en:

<https://www.washingtonpost.com/technology/2023/04/05/chatgpt-lies/>



04. Estrategias para que los sistemas de IA sean menos dañinos

UC2 | No maleficencia





04. Estrategias para que los sistemas de IA sean menos dañinos

En esta sección, presentaremos estrategias destinadas a hacer que los sistemas de IA sean menos dañinos mediante la promoción de la equidad, la responsabilidad y la transparencia en su desarrollo y despliegue. Estas estrategias permiten a los desarrolladores, los responsables políticos y las partes interesadas abordar de forma proactiva el sesgo algorítmico y mitigar sus posibles consecuencias negativas.

> Promover la equidad

Una estrategia clave para mitigar los perjuicios de los sistemas de IA sesgados es promover la equidad en los procesos algorítmicos de toma de decisiones. Esto implica garantizar que los modelos de IA se entrenen con conjuntos de datos diversos y representativos, libres de sesgos discriminatorios. Además, pueden emplearse técnicas de aprendizaje automático que tengan en cuenta la equidad para identificar y mitigar los sesgos en las predicciones algorítmicas, promoviendo así resultados equitativos para todas las personas.



➤ **Aumentar la responsabilidad**

Otro aspecto importante para reducir los daños de los sistemas de IA sesgados es aumentar la responsabilidad de los desarrolladores, las organizaciones y los responsables políticos. Esto incluye la aplicación de directrices éticas y mejores prácticas para el desarrollo de la IA, como la realización de evaluaciones de impacto exhaustivas para identificar posibles riesgos y daños. Además, establecer mecanismos claros de rendición de cuentas y marcos de supervisión puede ayudar a que las personas y las organizaciones se responsabilicen de las implicaciones éticas de sus despliegues de IA. En la próxima unidad de este curso exploraremos el concepto de rendición de cuentas con más detalle.

➤ **Fomentar la transparencia**

La transparencia es esencial para que los sistemas de IA sean menos perjudiciales, ya que fomenta la responsabilidad y la confianza entre las partes interesadas. La documentación transparente de los algoritmos de IA y los procesos de toma de decisiones permite el escrutinio y la validación externos, garantizando que los sesgos y los errores se identifiquen y aborden de manera oportuna. Además, fomentar el diálogo abierto y la colaboración entre los desarrolladores de IA, los investigadores y las comunidades afectadas puede facilitar una mayor transparencia y comprensión de las implicaciones éticas de las tecnologías de IA.

La Unidad 4 de este curso profundizará en el concepto de Transparencia, ya que es uno de los aspectos más fundamentales para garantizar una IA responsable.

> **Garantizar la privacidad**

Los sistemas de IA son herramientas potentes, pero su comodidad no debe ir en detrimento de la privacidad. Esta estrategia se centra en salvaguardar tu información personal. Los desarrolladores deben recopilar y utilizar la menor cantidad de datos posible, sobre todo los sensibles. Las medidas de seguridad deben ser de primera categoría para mantener a salvo la información. Los sistemas de IA también deben crearse para cumplir las leyes y normativas sobre privacidad, incluido el Reglamento General de Protección de Datos (RGPD) en Europa, que otorga a los individuos un control significativo sobre sus datos personales.

> **Prioridad a la seguridad**

Cuando se trata de IA, la seguridad debe ser la máxima prioridad. Esto significa someter los sistemas de IA a rigurosas pruebas y procesos de validación antes de lanzarlos al mundo real. El objetivo es identificar y solucionar cualquier riesgo o problema potencial que pueda causar daños. Al garantizar que los sistemas de IA funcionan de forma fiable y segura, podemos proteger a las personas y a la sociedad en su conjunto.



