2.2 Veracidad: sesgos, alucinaciones y deepfakes

El uso de sistemas de inteligencia artificial plantea importantes cuestiones relacionadas con la **fiabilidad de la información que producen**: pueden incorporar **sesgos**, producir **alucinaciones** y también facilitar la creación de contenidos manipulados, como los **deepfakes**, que imitan de forma convincente imágenes, voces o vídeos de personas reales. Todo ello hace necesario analizar con atención la **veracidad de los contenidos generados por la IA** y fomentar un uso crítico de estas herramientas, especialmente en el ámbito educativo.

Sesgos y discriminación

> Un **sesgo** puede entenderse como una **tendencia a favorecer o perjudicar a alguien o algo de manera sistemática.**

En la vida cotidiana, los sesgos aparecen cuando nuestras decisiones o juicios no son completamente neutrales, sino que están influidos por experiencias previas, estereotipos o información incompleta. Por ejemplo, si una persona cree que los estudiantes que hablan más en clase son siempre los que mejor aprenden, podría valorar más sus intervenciones y prestar menos atención a quienes participan menos, aunque estos también comprendan bien el contenido.

> Cuando hablamos de **sesgos en la inteligencia artificial**, nos referimos a algo similar: los sistemas de IA pueden producir resultados que favorecen o perjudican a determinados grupos de personas, categorías, resultados o situaciones de forma sistemática, generando errores recurrentes en las predicciones o decisiones automatizadas.

Esto ocurre porque los algoritmos aprenden a partir de datos generados por personas y por la sociedad, que ya pueden contener desigualdades o representaciones incompletas de la realidad. Como resultado, **la IA puede reproducir o amplificar** esas tendencias, perpetuando esas disparidades o estereotipos sociales existentes.

El origen de los sesgos en la inteligencia artificial es **multifacético** y puede surgir en diferentes etapas del proceso de creación y uso de un sistema. Ferrara (2023) distingue principalmente tres orígenes:

1. **Sesgo de datos:** ocurre cuando los datos utilizados para entrenar los modelos **no son representativos o están incompletos**. Esto sucede si los datos provienen de fuentes ya sesgadas, contienen errores o carecen de información importante sobre ciertos grupos. Los modelos de aprendizaje automático aprenden y replican estos patrones de sesgo presentes en los datos de entrenamiento. Por ejemplo, según el informe de la OCDE los modelos de IA suelen basarse de forma abrumadora **en culturas occidentales y de habla inglesa, perjudicando especialmente a hablantes de otras lenguas y dialectos específicos** y silenciando realidades o valores que no resultan convenientes para los intereses de quienes programan el algoritmo.

   Dentro de este tipo de sesgo, podemos integrar el **sesgo histórico**, que aparece cuando los datos reflejan desigualdades del pasado y estas se trasladan a las decisiones automatizadas.

2. **Sesgo algorítmico:** este sesgo es **inherente al diseño e implementación del algoritmo**. Surge cuando los algoritmos se basan en suposiciones sesgadas o utilizan criterios que priorizan ciertos atributos de manera que generan resultados injustos. Es decir, cuando el propio diseño del sistema favorece ciertos resultados.

3. **Sesgo del usuario:** se produce cuando las **personas que utilizan los sistemas introducen sus propios prejuicios** de forma consciente o inconsciente. Esto puede ocurrir al proporcionar datos de entrenamiento sesgados por parte de la persona que desarrolla el sistema o por el propio usuario en sus interacciones con la IA, de manera que reflejen sus prejuicios personales.

Además, en el artículo de Jeff Shuford, encontramos una tabla donde se describen los siguientes tipos de sesgos:

Sesgo de Muestreo	Se da cuando los datos de entrenamiento no representan a la población a la que sirven, como un algoritmo de reconocimiento facial entrenado mayoritariamente con personas blancas.
Sesgo de Representación	Sucede cuando el conjunto de datos no modela con precisión a la población, como bases de datos médicas que subrepresentan a las mujeres.
Sesgo de Confirmación	Ocurre cuando el sistema de IA se utiliza para confirmar prejuicios o creencias preexistentes de sus creadores o usuarios.
Sesgo de Medición	Emerge cuando el sistema de recolección de datos sobrerrepresenta o subrepresenta sistemáticamente a ciertos grupos.

Sesgo de Interacción	Aparece cuando la IA interactúa con los humanos de forma sesgada, como un chatbot que responde de manera distinta a hombres y mujeres.
Sesgo Generativo	Específico de modelos de IA generativa (como DALL-E o GPT), donde los resultados reflejan de manera desproporcionada patrones o perspectivas específicas de los datos de entrenamiento. Por ejemplo, al solicitar imágenes de "CEOs", los modelos suelen producir mayoritariamente imágenes de hombres, y al solicitar imágenes de "criminales", tienden a mostrar de forma abrumadora a personas de color.

En la siguiente imagen no sólo percibimos el sesgo generativo, sino también un sesgo de idioma ya que probablemente el modelo se entrenó en lengua inglesa donde "teacher" no tiene género.

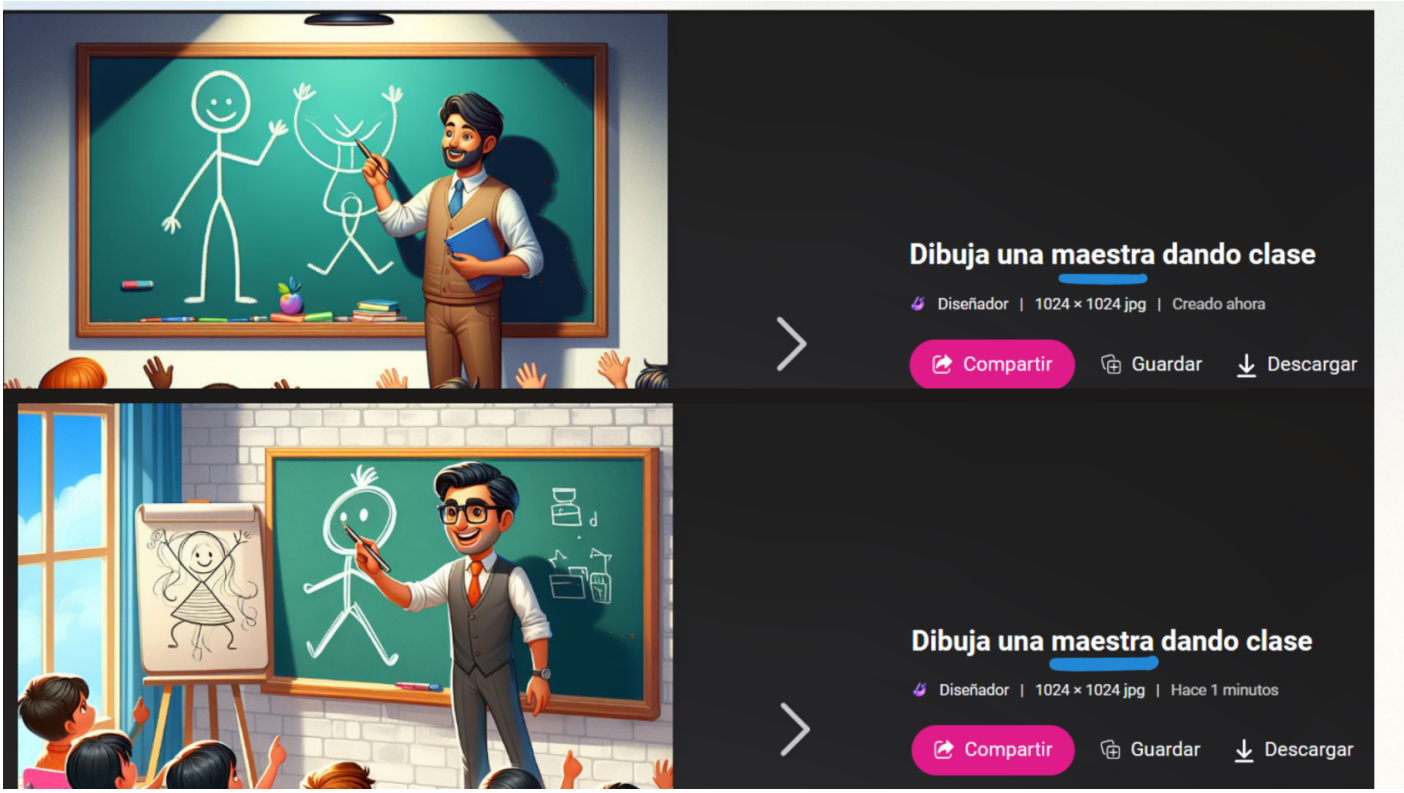


Imagen generada con Bing Image Creator (2023)

En relación a este último sesgo y los modelos de IA generativa (IAGen), es menester mencionar de forma explícita las siguientes problemáticas derivadas del mismo:

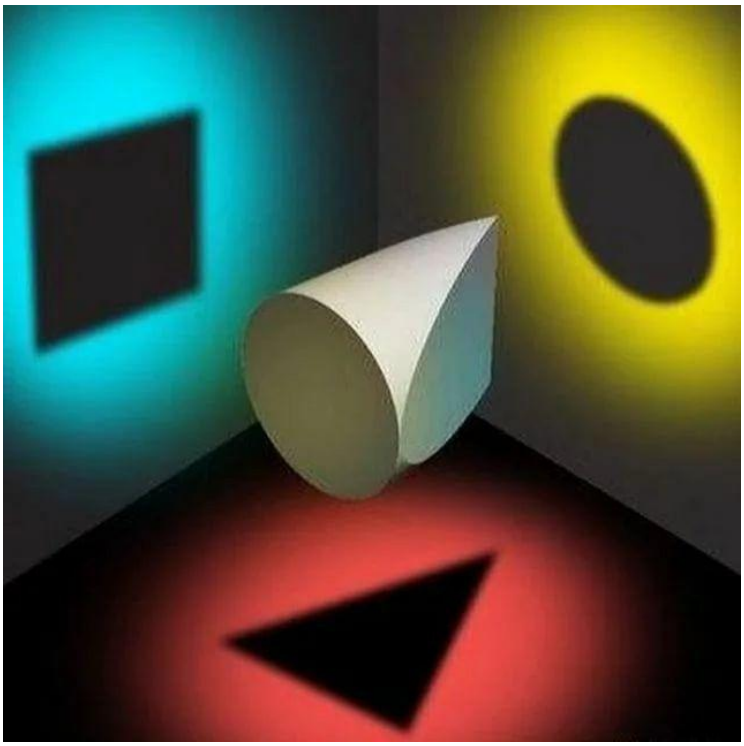
- **Amplificación de Estereotipos:** Los modelos de IAGen pueden **reproducir y amplificar estereotipos sociales** al generar contenido. Por ejemplo cuando la IA representa sistemáticamente a las enfermeras como mujeres y a los doctores como hombres.
- **Riesgo de las Realidades Sintéticas:** Se advierte que, a medida que avanzamos en la creación de **realidades sintéticas** cada vez más sofisticadas, existe el peligro de que sesgos sutiles se infiltren y moldeen la sociedad de formas no deseadas y potencialmente dañinas.
- **Impacto de los Datos de Internet:** Los modelos de IAGen entrenados con imágenes o textos extraídos de internet suelen heredar las **disparidades existentes** en el mundo real, lo que contamina sus resultados generados.

> Como docentes, es importante añadir el **sesgo de automatización**: tendencia humana a favorecer las sugerencias de los sistemas IA e ignorar el resto.

> El impacto de estos sesgos es profundo, ya que pueden **perpetuar desigualdades sociales**, reforzar estereotipos dañinos y limitar el acceso a servicios esenciales como la salud o el empleo.

Según la OCDE, aproximadamente cuatro de cada diez docentes temen que la IA pueda amplificar sesgos que refuercen conceptos erróneos en los estudiantes.

El algoritmo tiene el poder de iluminar partes de la realidad y dejar a oscuras, silenciadas, realidades no convenientes a los intereses que lo programan.



Tres proyecciones de una misma realidad

Como vemos, esto no es solo debido a unas malas intenciones del diseñador sino también a la cantidad histórica de datos estructurados generados por cada uno de los sectores de la sociedad en función de su riqueza, país de origen o de razones históricas. El riesgo socioeducativo es la desmaterialización de la diversidad.

Si bien es cierto, como docentes debemos tener en cuenta que toda selección de contenidos, parte de un sesgo, o, al menos, de una perspectiva y todos los docentes seleccionamos en un océano infinito de contenidos aquellos que trabajamos con nuestro alumnado. Para ello disponemos del currículo pero también de nuestra perspectiva personal y humana en ese tercer nivel de concreción que es nuestra programación de aula. Si bien parece inevitable partir de un cierto sesgo, si se pueden valorar los objetivos y motivaciones de cada selección de información.

> El abordaje del análisis crítico de los sistemas inteligentes implica partir del hecho de que los datos y los algoritmos no vienen dados, responden a los contextos históricos, políticos, sociales, culturales de su producción y existe una dimensión subjetiva tanto en la producción como en la mediación algorítmica (Martins 2024).

Alucinaciones

Además de los sesgos, otro aspecto importante que afecta a la fiabilidad de los sistemas de inteligencia artificial es su **capacidad de generar información incorrecta que parece plausible**. Mientras que los sesgos se refieren a tendencias sistemáticas que pueden favorecer o perjudicar determinados resultados o grupos de personas, los sistemas de IA generativa también pueden producir **afirmaciones, datos o referencias que no son verdaderos**. Este fenómeno se conoce como **alucinaciones de la inteligencia artificial.**

> Podemos definir las **alucinaciones** como los **contenidos generados por la IA** que parecen coherentes y convincentes, pero que en realidad son **inventados** o no están respaldados por información verificable.

Este fenómeno ocurre porque los sistemas de IA generativa no "entienden" realmente la información que procesan, sino que generan contenido basándose en patrones estadísticos.

Un **ejemplo que se hizo viral** fue el caso Mata y la aerolínea Avianca en 2023:

> “El abogado de un hombre que demandó a una aerolínea por daños personales utilizó ChatGPT para preparar una presentación, pero el bot de inteligencia artificial entregó casos falsos que el abogado presentó después ante el tribunal, lo que llevó a un juez a considerar sanciones mientras la comunidad jurídica lidia con uno de los primeros casos de **"alucinaciones" de IA** que hacen acto de presencia en los tribunales.
>
> Revista Forbes Argentina

> Este tipo de errores pone de relieve la importancia de **verificar siempre la información producida por sistemas de IA** antes de utilizarla como fuente.

El fenómeno de las **alucinaciones** en la IA representa un desafío profundo para la verdad educativa, ya que estos sistemas están diseñados para priorizar la verosimilitud sobre la veracidad. Como advierte la OCDE (2026), estos modelos pueden generar información que resulta totalmente plausible y bien estructurada pero que es fundamentalmente errónea, llegando incluso a fabricar citas bibliográficas inexistentes. Esto crea una suerte de "caverna digital" donde alumnado y docentes pueden terminar interactuando con "sombras lingüísticas": proyecciones estadísticas de textos que no tienen nada que ver con la realidad del mundo.

Imagen de @philosophymeme0

> En contextos educativos es imprescindible tener este fenómeno en cuenta, ya que **premisa falsa introducida por la IA puede descarrilar todo el proceso de aprendizaje de un estudiante.**

Aceptar estos resultados sin un filtro riguroso transforma el aula en un espacio de "infodemia", donde el tsunami de datos desaloja la acción racional y la comprensión profunda.

Para contrarrestar este riesgo, los docentes debemos tratar de ser el último "guía de la razón humana", ejerciendo su juicio profesional para validar y respaldar cada resultado antes de que sea integrado en el proceso de aprendizaje. El objetivo es que nuestro alumnado no caiga en una confianza ciega en la "razón" algorítmica (sesgo de automatización), a través de una alfabetización crítica que permita usar la IA como un amplificador del saber humano.

Pieter Brueghel el Viejo (1568). *La parábola de los ciegos* [Óleo sobre tabla]. Museo di Capodimonte de Nápoles https://es.wikipedia.org/wiki/La_par%C3%A1bola_de_los_ciegos

> Dejadlos: son ciegos que guían a ciegos. Y si un ciego guía a otro ciego, los dos caerán en el hoyo
>
> Mateo 15, 14.

Además, para mitigar las alucinaciones, la OCDE alude al uso de **técnicas** como la Generación Aumentada por Recuperación **(RAG)**, que ancla las respuestas en bases de datos confiables como libros de texto, y de la que hablaremos en el curso 2 de este itinerario "IA y diseño curricular". Por otro lado, algunos **enfoques pedagógicos** proponen permitir que los docentes ajusten el "**porcentaje de alucinación**" de las herramientas para fomentar el pensamiento crítico de los alumnos al obligarlos a verificar la información.

Comprender los tipos de sesgo y las alucinaciones es fundamental para analizar críticamente el funcionamiento y uso de la IA especialmente en contextos como la educación, donde la veracidad

de la información es esencial.

Deepfakes

Las tecnologías de la IA desempeñan una función cada vez más importante en el procesamiento, la estructuración y el suministro de información; las cuestiones del periodismo automatizado y del suministro algorítmico de noticias y la moderación y la conservación de contenidos en los medios sociales y los buscadores son solo algunos ejemplos que plantean cuestiones relacionadas con el acceso a la información, la desinformación, la información errónea, el discurso de odio, la aparición de nuevas formas de narrativa social, la discriminación, la libertad de expresión, la privacidad y la alfabetización mediática e informacional, entre otras (UNESCO 2022).

Aunque la **desinformación** no es un fenómeno nuevo, la IA permite producirla a gran escala y con menor esfuerzo, lo que incrementa el riesgo de que los **usuarios compartan contenidos sin comprobar su autenticidad.** Los sistemas de IA pueden generar textos, imágenes, audios o vídeos con gran apariencia de realismo, lo que facilita la producción y difusión de contenidos engañosos. En el contexto digital actual, donde gran parte de la información circula a través de redes sociales y plataformas en línea, esta capacidad puede contribuir a la propagación de **noticias falsas (fake news)** o **contenidos manipulados** que resultan **difíciles de distinguir de la información verificada**.

Un caso especialmente relevante es el de los **deepfakes**, es decir, vídeos, imágenes o audios generados o manipulados mediante IA que imitan de manera muy convincente la apariencia o la voz de una persona real. Es decir, la IAGen permite crear imágenes, vídeos o audios falsos que pueden representar a una persona diciendo o haciendo algo que nunca ocurrió. Estos contenidos pueden utilizarse con fines humorísticos o creativos, pero también para difundir desinformación, suplantar identidades o manipular la opinión pública.

Aquí puedes consultar imágenes deepfakes que se hicieron virales en 2023 como este vídeo de un supuesto Morgan Freeman:

https://www.youtube.com/embed/oxXpB9pSETo?si=lLJWFYArxk21_K5R

Además, la creciente sofisticación y el avance de estas tecnologías hace cada vez más complicado detectar las manipulaciones únicamente mediante la observación directa, por lo que se vuelve necesario desarrollar herramientas técnicas y competencias críticas para evaluar la credibilidad de la información.

> Comparison AI of Will Smith eating spaghetti from 2023 vs 2026 is going viral
> pic.twitter.com/nS1Dl49irC
>
> — kira □□ (@kirawontmiss) February 12, 2026

Conviene destacar a este respecto, que la **Ley de IA** introduce obligaciones de informar que un contenido está hecho con IA cuando pueda surgir un riesgo por falta de transparencia en torno a su uso:

> “ En algunos casos, el resultado de la IA generativa debe estar **visiblemente etiquetado**, como en el caso de los «deepfakes» y los textos destinados a informar al público sobre asuntos de interés público.

En el ámbito educativo, estas cuestiones tienen implicaciones importantes: puede derivar en situaciones graves como la creación y difusión de **contenidos manipulados de alumnado o profesorado**, incluidos montajes de **carácter sexual o desnudos falsos**, que se **comparten rápidamente** a través de redes sociales o por WhatsApp. Este tipo de prácticas puede convertirse en una forma de **ciberacoso**, con **consecuencias psicológicas, sociales y reputacionales** muy serias para las personas afectadas.

En la siguiente gráfica del estudio **“Hand in Hand: Schools’ Embrace of AI Connected to Increased Risks to Students”,** vemos cómo los deepfakes son un tema destacado en los centros educativos con una creciente conciencia entre las familias:

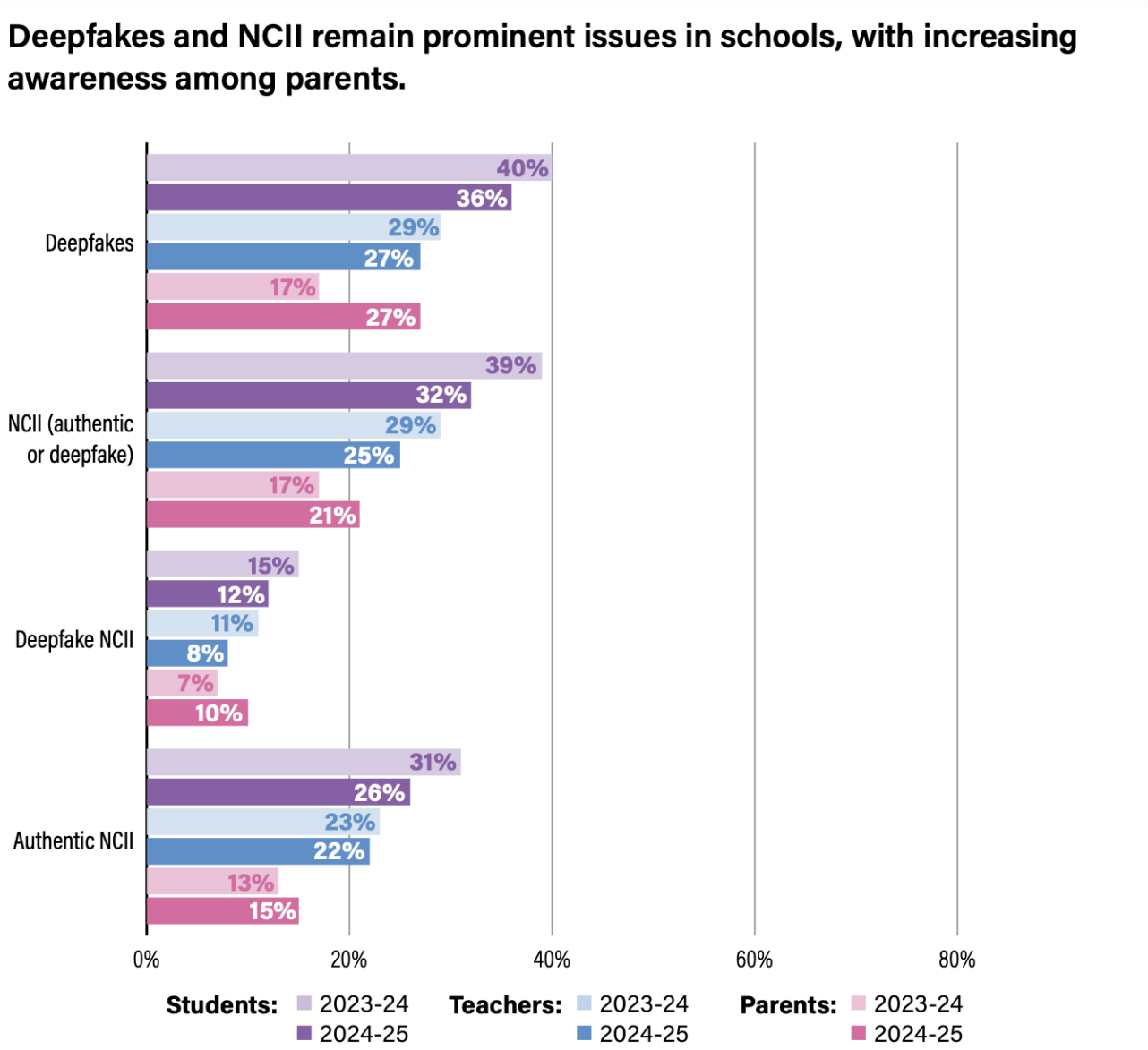


Figure 32. Percentage (%) of students, teachers, and parents who have heard of deepfakes, NCII, deepfake NCII, or authentic NCII being shared that depicts someone associated with their/their child's school

Hand in Hand: Schools' Embrace of AI Connected to Increased Risks to Students

**NCII: Non Consensual Intimate Image. Difusión no consentida de imágenes íntimas.*

Por ello, los centros educativos se enfrentan al reto de **prevenir, detectar y abordar estas situaciones**, promoviendo una educación digital responsable, el respeto a la privacidad y la conciencia sobre las implicaciones éticas y legales del uso de estas tecnologías.

En este sentido, conviene hacer consciente tanto al profesorado como al alumnado del **derecho de imagen**, (lo veremos en el capítulo cuatro de este curso), ya que:

La**ley prohíbe** captar, difundir o **utilizar la imagen de alguien sin su consentimiento expreso**, y esto incluye **modificar una fotografía mediante herramientas de inteligencia artificial**, crear montajes o aplicar filtros sobre la imagen de alguien sin su autorización, aunque sea sin mala intención.

Para completar esta información puedes visualizar el episodio **Porno, IA y menores** de rtve.

Además, la educación debe reforzar el desarrollo de **competencias de alfabetización mediática y digital**, que incluyan la capacidad de contrastar fuentes, identificar señales de desinformación y analizar críticamente los contenidos generados por IA..

En **commonsense.org** hay una página destinada a **IA y centros educativos con diversos juegos;** concretamente **aquí** puedes seleccionar edades y nivel de juego para adivinar qué cartel de película ha sido creado o generado por IA.

En esta página puedes jugar a adivinar **qué persona es real** y cuál ha sido generada por IA

Puntos clave

1. La **supervisión humana** se considera el salvaguarda ético fundamental para corregir la deriva lógica y las inexactitudes de la IA.
2. Siempre se ha de **evaluar** cualquier **resultado generado por la IA,** tanto para respaldarlo como para rechazarlo o modificarlo, asegurándonos en el ámbito educativo de la calidad pedagógica que aporta.
3. Aprender a **evaluar la veracidad de la información** se convierte en una **habilidad clave** para participar de manera informada y responsable en la **sociedad digital.**

Así pues, es vital desarrollar la "**alfabetización en IA**" tanto en alumnado como en profesorado para que puedan evaluar críticamente la credibilidad de la información y reconocer sesgos potenciales. Y por supuesto, para ello es necesario que haya un desarrollo del pensamiento crítico a través del fomento de lo que Maryanne Wolf denomina como lectura profunda (deep Reading).

Lectura profunda es el estado en el que usamos la corteza cerebral para realizar analogías e inferencias. Este proceso es fundamental para desarrollar un **pensamiento crítico y analítico**; sin él, solo se obtiene información superficial. Cuando se logra fluidez, el cerebro

> utiliza rutas más rápidas y eficientes, lo que libera tiempo para generar **pensamientos más profundos** e integrar sentimientos con la experiencia personal. Si dejamos de practicarla, podríamos **perder la capacidad de comprender contenidos complejos** y de involucrar nuestra imaginación.

Revision #50
Created 2026-01-01 19:36:18 CET by Maria
Updated 2026-05-26 11:14:33 CEST by Maria