La Inteligencia Artificial Generativa en la educación y la investigación

UNESCO – 2023

Apartados 1 y 2

Traducción de la Cátedra Comunicación 3 – FCS – UNLZ

1. ¿Qué es la IA generativa y cómo funciona?

1.1 ¿Qué es la IA generativa?

La IA generativa (GenAI) es una tecnología de inteligencia artificial (IA) que crea contenido automático en respuesta a indicaciones escritas en lenguaje natural e interfaces conversacionales. En lugar de simplemente seleccionar las páginas web existentes, GenAI genera contenido nuevo a partir de datos existentes. El contenido puede aparecer en formatos que abarcan todas las representaciones simbólicas del pensamiento humano: textos escritos en lenguaje natural, imágenes (incluyendo fotografías, pinturas digitales y dibujos animados), vídeos, música y código de software.

La GenAI se entrena utilizando datos recopilados de páginas web, conversaciones en redes sociales y otros medios online. Genera su contenido mediante un análisis estadístico de las distribuciones de palabras, píxeles u otros elementos en los datos que ha absorbido, identificando y repitiendo patrones comunes (por ejemplo, qué palabras suelen seguir a otras).

Aunque la GenAI puede producir contenido nuevo, no puede generar ideas nuevas ni soluciones a problemas del mundo real, ya que no comprende los objetos del mundo real ni las relaciones sociales que sustentan el lenguaje. Además, a pesar de su producción fluida e impresionante, no se puede confiar en que la GenAI sea precisa. De hecho, incluso el proveedor de ChatGPT reconoce que "aunque herramientas como ChatGPT a menudo pueden generar respuestas que parecen razonables, no se puede confiar en su precisión" (OpenAI, 2023). En la mayoría de los casos, los errores pasarán desapercibidos a menos que el usuario tenga un conocimiento sólido del tema en cuestión.

1.2 ¿Cómo funciona la IA generativa?

Las tecnologías específicas detrás de la IA generativa pertenecen a la familia de tecnologías de Inteligencia Artificial (IA) llamadas aprendizaje de máquina (ML), que utilizan algoritmos para mejorar continuamente y de forma automática su rendimiento a partir de los datos. El tipo de ML que ha impulsado muchos de los avances que hemos visto en la IA en los últimos años, como el uso de la IA para el reconocimiento facial, se conoce como **redes neuronales artificiales (ANNs),** inspiradas en el funcionamiento del cerebro humano y sus conexiones sinápticas entre neuronas. Existen muchos tipos de ANNs.

Tanto las tecnologías de IA generativa de texto como de imagen se basan en un conjunto de tecnologías de IA que han estado disponibles para los investigadores durante varios años. Por ejemplo, ChatGPT utiliza un transformador pre-entrenado generativo (GPT), mientras que la IA generativa de imágenes suele utilizar lo que se conoce como redes adversarias generativas (GANs) (véase la Tabla 1).

<table>
<tr><td colspan="2">Aprendizaje de máquina (ML)</td><td>Un tipo de IA que utiliza datos para mejorar automáticamente su rendimiento.</td></tr>
<tr><td colspan="2">Red Neuronal Artificial (ANN)</td><td>Un tipo de ML inspirado en la estructura y el funcionamiento del cerebro humano (por ejemplo, las conexiones sinápticas entre las neuronas).</td></tr>
<tr><td rowspan="3">IA Generativa de Texto</td><td>Transformador de propósito general</td><td>Un tipo de ANN que es capaz de centrarse en diferentes partes de los datos para determinar cómo se relacionan entre sí.</td></tr>
<tr><td>Modelos de lenguaje grande (LLM)</td><td>Un tipo de transformador de propósito general que se entrena con grandes cantidades de datos de texto.</td></tr>
<tr><td>Transformador pre-entrenado generativo (GPT)</td><td>Un tipo de LLM que se pre-entrena con cantidades aún mayores de datos, lo que permite al modelo capturar los matices del lenguaje y generar un texto coherente y consciente del contexto.</td></tr>
<tr><td rowspan="2">IA Generativa de imágenes</td><td>Redes Adversarias Generativas (GANs)</td><td rowspan="2">Un tipo de ANN utilizado para la generación de imágenes.</td></tr>
<tr><td>Autoencoders Variacionales (VAEs)</td></tr>
</table>

1.2.1. Cómo funcionan los modelos de IA generativa de texto

La IA generativa de texto utiliza un tipo de red neuronal artificial (ANN) conocida como transformador de propósito general, específicamente un subtipo llamado modelo de lenguaje grande (LLM). Por ello, a los sistemas de IA generativa de texto a menudo se les llama LLMs. El tipo de LLM utilizado por la IA generativa de texto se conoce como transformador pre-entrenado generativo (GPT), de ahí el "GPT" en "ChatGPT".

ChatGPT está construido sobre GPT-3, desarrollado por OpenAI. Esta fue la tercera iteración de su GPT, la primera lanzada en 2018 y la más reciente, GPT-4,

en marzo de 2023 (ver Tabla 2). Cada OpenAI GPT mejoró iterativamente respecto al anterior gracias a avances en arquitecturas de IA, métodos de entrenamiento y técnicas de optimización. Un aspecto conocido de su progreso continuo es el uso de cantidades crecientes de datos para entrenar su número exponencialmente creciente de "parámetros". Los parámetros pueden considerarse como botones metafóricos que se pueden ajustar para afinar el rendimiento del GPT. Incluyen los "pesos" del modelo, parámetros numéricos que determinan cómo procesa la entrada y produce la salida.

Además de los avances en la optimización de arquitecturas de IA y métodos de entrenamiento, esta rápida iteración también ha sido posible gracias a las cantidades masivas de datos y las mejoras en las capacidades de computación disponibles para las grandes empresas. Desde 2012, la capacidad de cómputo utilizada para entrenar modelos de IA generativa se ha duplicado cada 3-4 meses. En comparación, la Ley de Moore tenía un período de duplicación de dos años (OpenAI, 2018; Universidad de Stanford, 2019).

Modelo	Lanzamiento	Datos de entrenamiento	Número de parámetros	Características
GPT-1	2018	40 GB	117 millones	Capaz de tareas de procesamiento del lenguaje natural como completar textos y responder preguntas.
GPT-2	2019	40 GB	1.500 millones	Capaz de tareas de procesamiento del lenguaje natural más complejas como traducción automática y resumen.
GPT-3	2020	17.000 GB	175.000 millones	Capaz de tareas avanzadas de procesamiento del lenguaje natural como escribir párrafos coherentes y generar artículos completos. También es capaz de adaptarse a nuevas tareas con solo unos pocos ejemplos.
GPT-4	2023	1.000.000 GB	170 billones (reportado pero no confirmado)	Mayor confiabilidad y capacidad para procesar instrucciones más complejas.

Una vez que el GPT ha sido entrenado, generar una respuesta de texto a una indicación involucra los siguientes pasos:

1. **La indicación se divide en unidades más pequeñas (llamadas tokens) que se introducen en el GPT.**
2. **El GPT usa patrones estadísticos para predecir palabras o frases probables que podrían formar una respuesta coherente a la indicación.**
    - El GPT identifica patrones de palabras y frases que comúnmente coexisten en su modelo de datos grande preconstruido (que comprende texto extraído de internet y de otros lugares).
    - Utilizando estos patrones, el GPT estima la probabilidad de que palabras o frases específicas aparezcan en un contexto determinado.

    - **Comenzando con una predicción aleatoria, el GPT usa estas probabilidades estimadas para predecir la próxima palabra o frase probable en su respuesta.**
3. Las palabras o frases predichas se convierten en texto legible.
4. El texto legible se filtra a través de lo que se conoce como "barreras de seguridad" para eliminar cualquier contenido ofensivo.
5. Los pasos 2 a 4 se repiten hasta que se termina una respuesta. La respuesta se considera finalizada cuando alcanza un límite máximo de tokens o cumple con criterios de detención predefinidos.
6. **La respuesta se procesa posteriormente para mejorar la legibilidad mediante la aplicación de formato, puntuación y otras mejoras (como comenzar la respuesta con palabras que un humano podría usar, como "Claro", "Por supuesto" o "Lo siento").**

   Si bien los GPT y su capacidad para generar texto automáticamente han estado disponibles para los investigadores desde 2018, lo que hizo que el lanzamiento de ChatGPT fuera tan novedoso fue su acceso gratuito a través de una interfaz fácil de usar, lo que significa que cualquier persona con acceso a internet podía explorar la herramienta. El lanzamiento de ChatGPT causó gran conmoción en todo el mundo y rápidamente llevó a otras compañías tecnológicas globales a ponerse al día, junto con numerosas empresas emergentes, ya sea lanzando sus propios sistemas similares o creando nuevas herramientas a partir de ellos.

1.3 Ingeniería de prompts para generar resultados deseados

Si bien usar IA generativa puede ser tan simple como escribir una pregunta u otro indicador, la realidad es que aún no es sencillo para el usuario obtener exactamente el resultado que desea. Por ejemplo, la innovadora imagen de IA "Théâtre D'opéra Spatial", que ganó un premio en la Feria Estatal de Colorado en los Estados Unidos de América, requirió semanas de escribir indicadores y ajustar cientos de imágenes para generar la presentación final (Roose, 2022). El desafío similar de escribir indicadores efectivos para IA generativa de texto ha llevado a un aumento de puestos de trabajo de "ingeniería de prompts" en sitios web de reclutamiento (Popli, 2023).

"Ingeniería de prompts" se refiere a los procesos y técnicas para componer entradas que produzcan salidas de IA generativa que se asemejen más a la intención deseada del usuario.

La ingeniería de prompts es más exitosa cuando el prompt articula una cadena coherente de razonamiento centrada en un problema particular o una cadena de pensamiento en un orden lógico. Las recomendaciones específicas incluyen:

- Usar lenguaje simple, claro y directo que sea fácil de entender, evitando palabras complejas o ambiguas.

- Incluir ejemplos para ilustrar la respuesta deseada o el formato de las terminaciones generadas.
- Incluir contexto, que es crucial para generar terminaciones relevantes y significativas.
- Refinar e iterar según sea necesario, experimentando con diferentes variaciones.
- Ser ético, evitando prompts que puedan generar contenido inapropiado, sesgado o dañino.

También es importante reconocer de inmediato que no se puede confiar en los resultados de la IA generativa sin una evaluación crítica. Como OpenAI escribe sobre su GPT más sofisticado:

En vista de la calidad de los resultados de la IA generativa, se deben realizar pruebas rigurosas de usuario y evaluaciones de rendimiento antes de validar las herramientas para una adopción a gran escala o de alto riesgo. Estos ejercicios deben diseñarse con una métrica de rendimiento que sea más relevante para el tipo de tarea para la cual los usuarios piden resultados de la IA generativa. Por ejemplo, para resolver problemas matemáticos, "precisión" podría usarse como la métrica principal para cuantificar con qué frecuencia una herramienta de IA generativa produce la respuesta correcta; para responder a preguntas sensibles, la métrica principal para medir el rendimiento podría ser "tasa de respuesta" (la frecuencia con la que la IA generativa responde directamente a una pregunta); para la generación de código, la métrica puede ser "la fracción de los códigos generados que son directamente ejecutables" (si el código generado se puede ejecutar directamente en un entorno de programación y pasar las pruebas unitarias); y para el razonamiento visual, la métrica podría ser "coincidencia exacta" (si los objetos visuales generados coinciden exactamente con la verdad real) (Chen et al., 2023).

En resumen, a un nivel superficial, la IA generativa es fácil de usar; sin embargo, los resultados más sofisticados necesitan de intervención humana especializada y deben evaluarse críticamente antes de ser utilizados.

Implicaciones para la educación y la investigación

Si bien la IA generativa podría ayudar a maestros e investigadores a generar textos y otros resultados útiles para su trabajo, no es necesariamente un proceso sencillo. Pueden ser necesarias múltiples iteraciones de un prompt antes de lograr el resultado deseado. Una preocupación es que los estudiantes jóvenes, al ser por definición menos expertos que los maestros, podrían aceptar inconscientemente y sin un análisis crítico de los resultados de IA generativa que sean superficiales, inexactos o incluso dañinos.

1.4 EdGPT emergente y sus implicaciones

Dado que los modelos de IA generativa pueden servir como base o punto de partida para desarrollar modelos más especializados o específicos de un dominio, algunos investigadores han sugerido que los GPT deberían renombrarse como "modelos base" (Bommasani et al., 2021). En educación, los desarrolladores e investigadores han comenzado a ajustar un modelo base para desarrollar "EdGPT". Los modelos EdGPT se entrenan con datos específicos para servir a propósitos educativos. En otras palabras, EdGPT tiene como objetivo refinar el modelo que se ha derivado de cantidades masivas de datos de entrenamiento generales con cantidades más pequeñas de datos educativos de alta calidad y específicos del dominio.

Esto potencialmente le da a EdGPT un mayor alcance para apoyar el logro de las transformaciones enumeradas en la Sección 4.3. Por ejemplo, los modelos EdGPT dirigidos al codiseño del currículo pueden permitir a los educadores y estudiantes generar materiales educativos adecuados, como planes de lecciones, cuestionarios y actividades interactivas que se alineen estrechamente con un enfoque pedagógico eficaz y objetivos curriculares específicos y niveles de desafío para estudiantes particulares.

De manera similar, en el contexto de un entrenador de habilidades lingüísticas 1:1, un modelo base refinado con textos apropiados para un idioma en particular podría usarse para generar oraciones, párrafos o conversaciones de ejemplo para la práctica. Cuando los estudiantes interactúan con el modelo, puede responder con texto relevante y gramaticalmente correcto al nivel adecuado para ellos.

Teóricamente, las salidas de los modelos EdGPT también podrían contener menos sesgos generales o contenido objetable que el GPT estándar, pero aún podrían generar errores. Es fundamental tener en cuenta que, a menos que los modelos y enfoques subyacentes de IA generativa cambien significativamente, EdGPT aún puede generar errores y demostrar otras limitaciones. En consecuencia, sigue siendo importante que los principales usuarios de EdGPT, especialmente los profesores y los alumnos, adopten una perspectiva crítica ante cualquier resultado.

Actualmente, el refinamiento de modelos base para un uso más específico de GPT en la educación se encuentra en una etapa temprana. Los ejemplos existentes incluyen EduChat, un modelo base desarrollado por la Universidad Normal del Este de China para proporcionar servicios de enseñanza y aprendizaje, y cuyos códigos, datos y parámetros se comparten como código abierto. Otro ejemplo es MathGPT, desarrollado por el Grupo de Educación TAL: un LLM que se centra en la resolución de problemas relacionados con las matemáticas y las conferencias para usuarios de todo el mundo.

Sin embargo, antes de que sea posible un progreso significativo, es esencial que los esfuerzos se dediquen a refinar los modelos base no solo agregando

conocimiento del tema y eliminando sesgos, sino también agregando conocimiento sobre métodos de aprendizaje relevantes y cómo se puede reflejar esto en el diseño de algoritmos y modelos.

El desafío consiste en determinar hasta qué punto los modelos EdGPT pueden ir más allá del conocimiento del tema para también dirigirse a la pedagogía centrada en el estudiante y las interacciones positivas entre profesor y estudiante. El desafío adicional es determinar hasta qué punto los datos de alumnos y profesores pueden recopilarse y utilizarse éticamente para informar a un EdGPT. Finalmente, también es necesario que haya una investigación sólida para garantizar que EdGPT no vulnere los derechos humanos de los estudiantes ni desempodere a los profesores.

2. Controversias en torno a la IA generativa y sus implicaciones para la educación

Habiendo discutido previamente qué es GenAI y cómo funciona, esta sección examina las controversias y los riesgos éticos que plantean todos los sistemas GenAI y considera algunas de las implicaciones para la educación.

2.1 Empeoramiento de la pobreza digital

Como se señaló anteriormente, GenAI depende de enormes cantidades de datos y de una potencia informática masiva, además de sus innovaciones iterativas en arquitecturas de IA y métodos de capacitación, que en su mayoría solo están disponibles para las mayores empresas tecnológicas internacionales y unas pocas economías (principalmente las Estados Unidos, República Popular China y, en menor medida, Europa).

Esto significa que la posibilidad de crear y controlar GenAI está fuera del alcance de la mayoría de las empresas y de la mayoría de los países, especialmente aquellos del Sur Global. A medida que el acceso a los datos se vuelve cada vez más esencial para el desarrollo económico de los países y para las oportunidades digitales de las personas, aquellos países y personas que no tienen acceso o no pueden permitirse suficientes datos quedan en una situación de "pobreza de datos" (Marwala , 2023).).

La situación es similar en el caso del acceso a la potencia informática. La rápida penetración de la GenAI en países y regiones tecnológicamente avanzados ha acelerado exponencialmente la generación y el procesamiento de datos y, al mismo tiempo, ha intensificado la concentración de la riqueza de la IA en el Norte Global. Como consecuencia inmediata, las regiones con escasez de datos han sido aún más excluidas y expuestas a un riesgo a largo plazo de ser colonizadas por los estándares incorporados en los modelos GPT. Los modelos ChatGPT actuales se basan en datos de usuarios en línea que reflejan los valores y normas del Norte Global, lo que los hace inapropiados para algoritmos de IA localmente relevantes en comunidades con escasez de datos en muchas partes del Sur Global o en comunidades más desfavorecidas en el Sur Global. Norte.

2.2 Superar la adaptación regulatoria nacional

Los proveedores dominantes de GenAI también han sido criticados por no permitir que sus sistemas estén sujetos a una rigurosa revisión académica independiente (Dwivedi et al., 2023).44 Las tecnologías fundamentales de GenAI de una empresa tienden a protegerse como propiedad intelectual corporativa . Mientras tanto, a muchas de las empresas que están empezando a utilizar GenAI les resulta cada vez más difícil mantener la seguridad de sus sistemas (Lin, 2023). Además, a pesar de los llamados a la regulación por parte de la propia industria de la IA,45 la redacción de legislación sobre la creación y el uso de toda la IA, incluida la GenAI, a menudo va a la zaga del rápido ritmo de desarrollo.

Esto explica en parte los desafíos que enfrentan las agencias nacionales o locales para comprender y gobernar las cuestiones legales y éticas.46 Si bien la GenAI puede aumentar las capacidades humanas para completar ciertas tareas, existe un control democrático limitado de las empresas que promueven la GenAI. Esto plantea la cuestión de las regulaciones, en particular con respecto al acceso y uso de datos nacionales, incluidos datos sobre instituciones e individuos locales, así como datos generados en el territorio de los países. Se necesita una legislación adecuada para que las agencias gubernamentales locales puedan obtener cierto control sobre las crecientes olas de GenAI para garantizar su gobernanza como un bien público.

2.3 Uso de contenido sin consentimiento

Como se señaló anteriormente, los modelos GenAI se construyen a partir de grandes cantidades de datos (por ejemplo, texto, sonidos, códigos e imágenes) a menudo extraídos de Internet y generalmente sin el permiso de ningún propietario. En consecuencia, muchos sistemas GenAI de imágenes y algunos sistemas GenAI de código han sido acusados de violar los derechos de propiedad intelectual. Al momento de escribir este artículo, hay varios casos legales internacionales en curso que se relacionan con este tema. Además, algunos han señalado que los GPT pueden contravenir leyes como el Reglamento General de Protección de Datos de la Unión Europea (2016) o GDPR, especialmente el derecho de las personas al olvido, ya que actualmente es imposible eliminar los datos de alguien (o los resultados de esos datos). de un modelo GPT una vez que ha sido entrenado

2.4 Modelos inexplicables utilizados para generar resultados

Hace tiempo que se reconoce que las redes neuronales artificiales (RNA) suelen ser "cajas negras"; es decir, que su funcionamiento interno no está abierto a inspección. Como resultado, las RNA no son "transparentes" ni "explicables" y no es posible determinar cómo se determinaron sus resultados. Si bien el enfoque general, incluidos los algoritmos utilizados, es generalmente explicable, los modelos particulares y sus parámetros, incluidos los pesos del modelo, no son inspeccionables, razón por la cual no se puede explicar un resultado específico que se genera. Hay miles de millones de parámetros/pesos en un modelo como GPT-4 (ver Tabla 2) y son los pesos en conjunto los que contienen los patrones aprendidos que el modelo utiliza para generar sus resultados.

Como los parámetros/ponderaciones no son transparentes en las RNA (Tabla 1), no se puede explicar la forma precisa en que estos modelos crean una salida específica. La

falta de transparencia y explicabilidad de GenAI es cada vez más problemática a medida que GenAI se vuelve cada vez más compleja (ver Tabla 2), produciendo a menudo resultados inesperados o no deseados. Además, los modelos GenAI heredan y perpetúan sesgos presentes en sus datos de entrenamiento que, dada la naturaleza poco transparente de los modelos, son difíciles de detectar y abordar. Finalmente, esta opacidad también es una causa clave de los problemas de confianza en torno a GenAI (Nazaretsky et al., 2022a). Si los usuarios no entienden cómo un sistema GenAI llegó a un resultado específico, es menos probable que estén dispuestos a adoptarlo o utilizarlo (Nazaretsky et al., 2022b).

2.5 Contenido generado por IA que contamina Internet

Debido a que los datos de entrenamiento de GPT generalmente se extraen de Internet, que con demasiada frecuencia incluye lenguaje discriminatorio y otros lenguajes inaceptables, los desarrolladores han tenido que implementar lo que llaman "barandillas" para evitar que la salida de GPT sea ofensiva y/o o poco ético. Sin embargo, debido a la ausencia de regulaciones estrictas y mecanismos de monitoreo efectivos, los materiales sesgados generados por GenAI se están extendiendo cada vez más por Internet, contaminando una de las principales fuentes de contenido o conocimiento para la mayoría de los estudiantes en todo el mundo. Esto es especialmente importante porque el material generado por GenAI puede parecer bastante preciso y convincente, cuando a menudo contiene errores e ideas sesgadas. Esto plantea un alto riesgo para los estudiantes jóvenes que no tienen conocimientos previos sólidos sobre el tema en cuestión. También plantea un riesgo recursivo para futuros modelos GPT que se entrenarán con texto extraído de Internet que los propios modelos GPT han creado y que también incluye sus sesgos y errores.

2.6 Falta de comprensión del mundo real

Los GPT de texto a veces se denominan peyorativamente "loros estocásticos" porque, como se señaló anteriormente, si bien pueden producir texto que parece convincente, ese texto a menudo contiene errores y puede incluir declaraciones dañinas (Bender et al. al., 2021). Todo esto ocurre porque los GPT solo repiten patrones de lenguaje encontrados en sus datos de entrenamiento (generalmente texto extraído de Internet), comenzando con patrones aleatorios (o "estocásticos") y sin comprender su significado, del mismo modo que un loro puede imitar sonidos sin comprenderlos realmente. lo que está diciendo. La desconexión entre los modelos GenAI que "parecen" comprender el texto que utilizan y generan, y la "realidad" de que no comprenden el lenguaje y el mundo real puede llevar a profesores y estudiantes a depositar un cierto nivel de confianza en el resultado que generan. no garantiza. Esto plantea serios riesgos para la educación futura.

De hecho, GenAI no se basa en observaciones del mundo real u otros aspectos clave del método científico, ni está alineado con valores humanos o sociales. Por estas razones, no puede generar contenido genuinamente novedoso sobre el mundo real, los objetos y sus relaciones, las personas y las relaciones sociales, las relaciones entre humanos y objetos o las relaciones tecnológicas humanas. Se cuestiona si el contenido aparentemente novedoso generado por los modelos GenAI puede reconocerse como conocimiento científico. Como ya se señaló, los GPT frecuentemente pueden producir texto inexacto o poco confiable. De hecho, es bien sabido que los GPT constituyen algunas cosas que no existen en la vida real. Algunos llaman a esto "alucinación", aunque otros critican el uso

de un término tan antropomórfico y, por tanto, engañoso. Así lo reconocen las empresas que producen GenAI.

La parte inferior de la interfaz pública de ChatGPT, por ejemplo, dice: "ChatGPT puede producir información inexacta sobre personas, lugares o hechos".2 Algunos defensores también han sugerido que GenAI representa un paso significativo en el camino hacia la Inteligencia Artificial General (AGI), un término que sugiere una clase de IA que es más inteligente que los humanos. Sin embargo, esto ha sido criticado desde hace tiempo, con el argumento de que la IA nunca progresará hacia la IAG al menos hasta que de alguna manera reúna, en simbiosis, tanto la IA basada en el conocimiento (también conocida como IA simbólica o basada en reglas) como la IA basada en datos (también conocida como aprendizaje automático) (Marcus, 2022). Las afirmaciones de IAG o de conciencia también nos distraen de una consideración más cuidadosa de los daños actuales que se perpetran con la IA, como la discriminación oculta contra grupos ya discriminados (Metz, 2021).

2.7 Reducir la diversidad de opiniones y marginar aún más las voces ya marginadas

ChatGPT y herramientas similares tienden a generar solo respuestas estándar que asumen los valores de los propietarios/creadores de los datos utilizados para entrenar los modelos. De hecho, si una secuencia de palabras aparece con frecuencia en los datos de capacitación –como es el caso de temas comunes y no controvertidos y creencias dominantes o dominantes– es probable que el GPT la repita en su producción. Esto corre el riesgo de limitar y socavar el desarrollo de opiniones plurales y expresiones plurales de ideas. Las poblaciones con escasez de datos, incluidas las comunidades marginadas del Norte Global, tienen una presencia digital en línea mínima o limitada. En consecuencia, sus voces no se escuchan y sus preocupaciones no están representadas en los datos que se utilizan para capacitar a los GPT, por lo que rara vez aparecen en los resultados. Por estas razones, dada la metodología de capacitación previa basada en datos de páginas web de Internet y conversaciones en redes sociales, los modelos GPT pueden marginar aún más a las personas que ya están desfavorecidas.

2.8 Generación de deepfakes más profundos

Además de las controversias comunes a todas las GenAI, GAN GenAI se puede utilizar para alterar o manipular imágenes o vídeos existentes para generar imágenes falsas que son difíciles de distinguir de las reales. GenAI está haciendo que sea cada vez más fácil crear estos 'deepfakes' y las llamadas 'noticias falsas'. En otras palabras, GenAI está facilitando que ciertos actores cometan actos poco éticos, inmorales y criminales, como difundir desinformación, promover discursos de odio e incorporar rostros de personas, sin su conocimiento o consentimiento, en películas completamente falsas y a veces comprometedoras.

Observaciones finales

Desde la perspectiva de un enfoque centrado en el ser humano, las herramientas de IA deben diseñarse para ampliar o aumentar las capacidades intelectuales y sociales humanas, y no socavarlas, entrar en conflicto con ellas o usurparlas. Durante mucho tiempo se ha esperado que las herramientas de IA puedan integrarse aún más como parte integrante de las herramientas disponibles para los humanos para respaldar el análisis y la acción para futuros más inclusivos y sostenibles.

Para que la IA se convierta en un elemento confiable y esencial de la colaboración humano-máquina, a nivel individual, institucional y sistémico, el enfoque centrado en el ser humano, basado en la Recomendación de la UNESCO sobre la Ética de la IA de 2021, debe ser especificado e implementado de acuerdo con las características específicas de las tecnologías emergentes como la IA Generativa (GenAI). Solo así podremos garantizar que la GenAI se convierta en una herramienta confiable para investigadores, docentes y estudiantes.

Si bien la GenAI debería utilizarse al servicio de la educación y la investigación, todos debemos ser conscientes de que la GenAI también podría cambiar los sistemas establecidos y sus fundamentos en estos ámbitos. La transformación de la educación y la investigación que impulsará la GenAI, si la hay, debe revisarse rigurosamente y dirigirse mediante un enfoque centrado en el ser humano . Sólo así podremos garantizar que el potencial de la IA en particular, y de todas las demás categorías de tecnologías utilizadas en la educación en general, mejoren las capacidades humanas para construir futuros digitales inclusivos para todos.