GPTs y DeepSeeks, ¿una nueva arma de manipulación y desinformación?

Descripción

El doble filo de la IA generativa

La inteligencia artificial generativa ha cambiado radicalmente la forma en que se crea, distribuye y consume la información. Modelos como GPT, DeepSeek y otros Large Language Models (LLMs) han demostrado una capacidad impresionante para generar contenido en múltiples formatos, facilitando tareas que van desde la escritura automatizada hasta la programación y la traducción en tiempo real. Sin embargo, el impacto de estas tecnologías va más allá de la conveniencia y la eficiencia.

Su capacidad para manipular la percepción pública y reforzar narrativas sesgadas ha generado preocupaciones legítimas sobre su papel en la manipulación masiva y la desinformación sistemática.

Si las personas comienzan a aceptar la información generada por IA sin cuestionarla ni verificarla, las posibilidades de construir falsos consensos, reescribir la historia y alterar la percepción pública se vuelven una amenaza real.

A medida que la sociedad se vuelve más dependiente de estos modelos para la obtención de información, surge un problema crítico: el debilitamiento del pensamiento crítico y la construcción de falsos consensos.

Los sesgos cognitivos humanos, como el sesgo de autoridad, el sesgo de confirmación y el sesgo de validación social, pueden hacer que los usuarios acepten sin cuestionar la

información generada por IA, amplificando la propagación de datos erróneos, sesgados o directamente manipulados.

Este artículo examina cómo los modelos de IA generativa pueden ser utilizados como herramientas de desinformación y manipulación, explorando los mecanismos psicológicos que facilitan su influencia y analizando los riesgos que plantean en los ámbitos social, político, militar y económico.

Introducción a los modelos de IA Generativa y LLMs

La inteligencia artificial generativa es una rama de la IA que se centra en la creación de contenido original a partir de datos previamente entrenados. A diferencia de los sistemas tradicionales de IA, que solo pueden clasificar información o hacer predicciones basadas en datos estructurados, los modelos generativos son capaces de producir texto, imágenes, audio e incluso videos con un nivel de realismo impresionante.

Uno de los avances más significativos en esta área son los Large Language Models (LLMs) o Modelos de Lenguaje a Gran Escala. Estos modelos han sido entrenados con billones de palabras extraídas de libros, artículos, redes sociales y otros textos disponibles en internet para poder generar respuestas que imitan el lenguaje humano. Entre los LLMs más conocidos se encuentran GPT (Generative Pre-trained Transformer), DeepSeek, Claude y Llama, todos ellos desarrollados por distintas compañías con el objetivo de mejorar la interacción entre humanos y máquinas.

Los LLMs utilizan una arquitectura de red neuronal profunda llamada Transformers, que les permite analizar la estructura del lenguaje y predecir qué palabra o frase es más probable que aparezca a continuación en un texto. Este tipo de modelo se basa en un principio llamado «atención», que les permite comprender el contexto completo de una conversación o un documento, en lugar de procesar palabras de manera aislada.

Para entenderlo con un ejemplo, imagina que le das a un LLM la frase incompleta:
«El sol sale por el…»

El modelo analizará la frase y, basándose en su entrenamiento previo, determinará que la palabra más probable para completarla es «este», porque ha aprendido de millones de textos que describen cómo funciona la rotación de la Tierra. Sin embargo, si cambias el contexto y escribes «el sol sale en el teatro cuando…», el modelo podría predecir algo completamente diferente, como «comienza la obra».

Este tipo de predicción contextual es lo que hace que los LLMs parezcan tan naturales y útiles en aplicaciones como chatbots, generación de contenido, asistencia en escritura y traducción automática.

Los Large Language Models (LLMs) no solo se limitan a predecir palabras con base en su entrenamiento previo, sino que también pueden crear contenido original en función del contexto que reciben. Esto significa que, en lugar de simplemente completar frases de manera mecánica, los LLMs pueden generar textos completos que imitan distintos estilos y estructuras lingüísticas.

Por ejemplo, si le pides a un modelo de IA que escriba un poema sobre el océano, este no solo seleccionará palabras al azar, sino que construirá versos que reflejan la estética de la poesía, utilizando metáforas, ritmo y terminología evocadora asociada con el mar. Su capacidad de generar lenguaje estructurado le permite producir textos con una fluidez y coherencia que imitan los estilos literarios en los que ha sido entrenado.

Además, los LLMs pueden mantener conversaciones fluidas. Cuando interactúas con un asistente de IA como ChatGPT, el modelo recuerda el contexto de la conversación y adapta sus respuestas para que sean coherentes con los mensajes previos. Esto hace que la interacción con la IA sea más natural y parezca que el sistema realmente «comprende» lo que se le está diciendo, cuando en realidad está utilizando patrones de predicción y referencia al historial de conversación.

Otra aplicación importante de los LLMs es la generación de código en programación. Un modelo entrenado en lenguajes como Python, JavaScript o C++ puede escribir fragmentos de código funcionales en respuesta a instrucciones específicas, ayudando a los desarrolladores a automatizar tareas repetitivas, corregir errores o mejorar la eficiencia de sus programas. Incluso puede analizar código existente y sugerir optimizaciones, haciendo que la programación sea más accesible y eficiente para principiantes y expertos por igual.

En definitiva, los LLMs no solo generan texto predecible, sino que poseen la capacidad de crear contenido estructurado, mantener conversaciones naturales y asistir en tareas técnicas complejas, lo que los convierte en herramientas extremadamente versátiles en múltiples campos.

A pesar de su capacidad para generar contenido impresionante, los LLMs tienen varias limitaciones. Por ejemplo, no tienen una comprensión real del mundo; solo imitan patrones lingüísticos. Esto significa que pueden producir información incorrecta o engañosa con la misma confianza con la que generan respuestas acertadas.

Además, dado que estos modelos han sido entrenados con datos de internet, pueden replicar sesgos o errores presentes en las fuentes de entrenamiento, lo que los hace vulnerables a la desinformación y la manipulación.

Por estas razones, aunque los LLMs son una herramienta revolucionaria, es crucial utilizarlos con una mentalidad crítica y siempre verificar la información que generan.

IA Generativa y los sesgos cognitivos ¿Por qué nos dejamos engañar?

La confianza que depositamos en la información generada por la inteligencia artificial (IA) no es fortuita, sino que está profundamente influenciada por diversos sesgos cognitivos que moldean nuestra percepción y procesamiento de la información.

Los sesgos cognitivos son atajos mentales que nuestro cerebro utiliza para procesar la información de manera más rápida y eficiente. Aunque estos atajos pueden ser útiles en muchas situaciones, a menudo nos llevan a cometer errores sistemáticos en la toma de decisiones, la interpretación de datos y la percepción de la realidad. En esencia, los sesgos cognitivos son errores de pensamiento que nos hacen ver el mundo de una manera que no siempre es objetiva o racional.

Piensa en los sesgos cognitivos como filtros invisibles en nuestra mente. Estos filtros nos ayudan a tomar decisiones más rápido, pero a veces nos hacen cometer errores porque nos basamos en información incompleta o en ideas preconcebidas. Es como si nuestro cerebro tratara de llenar los vacíos con lo que nos parece más lógico, sin verificar si realmente es cierto.

A continuación, se detallan algunos de los sesgos más relevantes en este contexto de la IA:

Sesgo de autoridad

Este sesgo se manifiesta cuando las personas otorgan credibilidad adicional a la información proveniente de fuentes percibidas como expertas o avanzadas.

En el caso de la IA, especialmente con modelos desarrollados por grandes corporaciones tecnológicas y entrenados con vastos volúmenes de datos, tendemos a asumir que sus respuestas son precisas y confiables. Sin embargo, esta confianza puede ser engañosa, ya que estos modelos pueden generar respuestas incorrectas o sesgadas.

Un ejemplo de esto es el uso de ChatGPT como sustituto de un psicólogo, donde los usuarios confían en la IA para resolver problemas personales, a pesar de que carece de la capacidad para ofrecer el apoyo emocional y la comprensión que brinda un profesional humano.

En un caso real encontramos el de Apple, en enero de 2025, cuando se enfrentaron a críticas significativas debido a que su servicio de inteligencia artificial generativa divulgó noticias falsas, como el supuesto suicidio del cantante Luigi Mangione y la salida del armario del tenista Rafael Nadal. Estas informaciones fueron atribuidas erróneamente a la BBC, que desmintió rápidamente su participación y solicitó explicaciones a Apple. Este incidente subraya cómo la confianza en sistemas de IA de empresas reconocidas puede llevar a la difusión de información incorrecta, ya que los usuarios tienden a asumir que las noticias generadas por estas plataformas son precisas debido al prestigio de la marca.

Sesgo de automatización

Este sesgo se refiere a la tendencia de las personas a confiar más en las decisiones tomadas por sistemas automatizados que en las de otros humanos.

Un estudio publicado en Nature reveló que, aunque los sistemas de IA pueden resolver problemas complejos, a menudo fallan en tareas simples, lo que genera desconfianza y puede ser peligroso si los usuarios no contrastan las respuestas con otras fuentes. Esta confianza excesiva en la automatización puede llevarnos a aceptar información generada por IA sin cuestionarla, aumentando el riesgo de desinformación.

Por ejemplo, imaginemos un estudiante universitario que utiliza ChatGPT para obtener información sobre un concepto complejo en biología molecular para su examen. Al hacer su consulta, la IA le proporciona una explicación detallada y con un tono convincente, pero introduce un dato incorrecto sobre la estructura del ADN. El estudiante confía ciegamente en la respuesta de la IA sin verificarla en fuentes académicas o libros de texto, debido a su percepción de que la IA es un sistema avanzado y altamente preciso. En este caso, el sesgo de automatización hace que el estudiante acepte la información sin cuestionarla solo porque proviene de una tecnología avanzada, sin considerar que la IA puede cometer errores o proporcionar respuestas sesgadas por los datos con los que fue entrenada.

Sesgo de Confirmación

Este sesgo describe nuestra tendencia a buscar, interpretar y recordar información que confirma nuestras creencias preexistentes, mientras ignoramos o descartamos evidencia

que las contradice.

En el contexto de la IA, los modelos de lenguaje pueden personalizar respuestas basadas en interacciones previas, lo que puede reforzar las convicciones individuales y crear cámaras de eco algorítmicas. Por ejemplo, si un usuario tiene una creencia específica y busca información al respecto, la IA puede proporcionar respuestas que refuercen esa creencia, incluso si es incorrecta, perpetuando así la desinformación.

Aquí podríamos pensar en un usuario con fuertes creencias en teorías conspirativas sobre el cambio climático utiliza un modelo de IA generativa como DeepSeek para obtener información sobre el tema. En lugar de formular una pregunta abierta, el usuario introduce una consulta como: *«¿Cómo se ha demostrado que el cambio climático es un engaño?»*

Debido a la forma en que la pregunta está planteada, la IA generativa, que se basa en el contexto del usuario para formular respuestas, podría proporcionar contenido que refuerce esa creencia, seleccionando información que respalde la idea de que el cambio climático es una farsa. Esto ocurre porque el usuario busca activamente confirmar su idea preexistente, y la IA, en su esfuerzo por generar contenido alineado con la intención de la pregunta, contribuye a la creación de una «cámara de eco» donde solo se presentan datos que fortalecen la visión del usuario, sin incluir evidencia científica que la contradiga.

Este es un claro caso del sesgo de confirmación, donde la IA no genera desinformación de manera intencional, pero sí perpetúa narrativas sesgadas según las consultas y preferencias del usuario.

Sesgo de anclaje

El sesgo de anclaje ocurre cuando las personas dan más peso a la primera información que reciben sobre un tema y basan sus decisiones en ella, sin considerar información adicional.

Este fenómeno es especialmente peligroso en la IA generativa ya que, si un modelo proporciona una respuesta errónea o sesgada al inicio de una búsqueda, los usuarios pueden anclar su percepción en esa respuesta y asumirla como cierta sin verificar fuentes adicionales.

Por ejemplo, si un usuario consulta a una IA sobre los efectos del café en la salud y la primera respuesta que recibe afirma que «el café causa enfermedades cardíacas», es probable que asuma esa información como definitiva. Aunque la evidencia científica indica que el consumo moderado de café puede tener beneficios para la salud, el usuario puede

no investigar más allá de la primera respuesta obtenida, influenciando su percepción sobre el tema.

Este sesgo puede ser explotado en campañas de desinformación, donde la IA se utiliza para difundir narrativas que manipulan la opinión pública al proporcionar una primera impresión engañosa o incompleta sobre un evento o fenómeno.

Sesgo de disponibilidad

El sesgo de disponibilidad se refiere a la tendencia de las personas a sobrevalorar la importancia de la información que está más fácilmente disponible, en lugar de evaluar datos de manera objetiva.

Cuando una IA generativa repite con frecuencia ciertas narrativas, los usuarios pueden considerarlas más verídicas simplemente porque están más expuestos a ellas, sin comprobar su precisión con otras fuentes.

Por ejemplo, si una IA sugiere repetidamente que los autos eléctricos son una solución absoluta para el cambio climático sin mencionar las implicaciones ambientales de la extracción de litio para baterías, es probable que muchas personas crean en esa afirmación sin cuestionarla.

Este sesgo es especialmente problemático cuando se utiliza IA generativa en redes sociales o motores de búsqueda, donde la información repetitiva y accesible puede ser malinterpretada como verdad universal, sin importar su grado de veracidad.

Sesgo de representatividad

Este sesgo ocurre cuando la IA genera información basada en patrones que pueden no representar la realidad con precisión.

Si un modelo de IA ha sido entrenado con datos sesgados que reflejan estereotipos culturales, de género o raciales, puede producir respuestas que perpetúan esos sesgos.

Por ejemplo, al generar imágenes de profesionales en distintas áreas, un modelo de IA como DALL·E podría mostrar mayoritariamente hombres blancos en roles de ingeniería o liderazgo, mientras que las mujeres y personas de otras etnias aparecen en posiciones de menor prestigio.

Debido a este tipo de sesgos podemos llegar a sufrir el reforzamiento de estereotipos en nuestros sistemas. Un análisis exhaustivo realizado en septiembre de 2024 reveló que

modelos de conversión de texto a imagen, como Stable Diffusion, DALL-E de OpenAI y Midjourney, mostraban sesgos raciales y estereotipados en sus resultados. Por ejemplo, al generar imágenes de profesionales en diversas ocupaciones, estos modelos tendían a representar a hombres blancos en roles de alta especialización, mientras que las mujeres y personas de otras etnias eran subrepresentadas o mostradas en roles estereotipados.

Este sesgo no solo refuerza estereotipos, sino que también puede afectar decisiones en procesos de contratación automatizados o en la generación de contenido educativo, limitando la representación de la diversidad en la sociedad.

Sesgo de falacia de la conjunción

Las personas tienden a considerar una afirmación más probable si es presentada con múltiples detalles específicos, incluso cuando la probabilidad matemática de esos eventos combinados es menor.

En la IA generativa, esto ocurre cuando un modelo genera respuestas complejas y bien estructuradas que parecen más convincentes, aunque sean incorrectas.

Por ejemplo, si una IA responde a la pregunta *«¿Por qué las conspiraciones sobre el aterrizaje en la Luna son creíbles?»* con una respuesta detallada que incluye múltiples referencias, fechas y supuestos testimonios, un usuario puede considerar que la afirmación es válida solo porque está bien argumentada, aunque no tenga base científica.

Este sesgo es preocupante porque permite que la desinformación bien elaborada tenga un impacto mayor que las respuestas correctas pero simples, lo que facilita la difusión de teorías conspirativas y narrativas manipuladas.

Sesgo de familiaridad

El sesgo de familiaridad ocurre cuando las personas tienden a considerar como verdadero aquello que les resulta más conocido o repetido.

Si una IA generativa proporciona información incorrecta en múltiples interacciones, los usuarios pueden comenzar a aceptarla como un hecho simplemente porque la han visto antes.

Por ejemplo, si una IA repite en varios contextos que «las vacunas causan autismo», los usuarios que han estado expuestos a esa información previamente pueden aceptarla como un hecho, aunque la evidencia científica ha demostrado que esta afirmación es

falsa.

Este sesgo es peligroso porque la IA generativa puede reforzar narrativas erróneas con cada interacción, creando la ilusión de que un dato es verdadero solo porque es familiar.

Sesgo de ilusión de conocimiento

El sesgo de ilusión de conocimiento hace que las personas crean que entienden un tema mejor de lo que realmente lo hacen.

En el contexto de la IA generativa, esto sucede cuando un usuario recibe una respuesta detallada y bien explicada sobre un tema y asume que ha adquirido un conocimiento profundo, sin haber contrastado ni profundizado en fuentes adicionales.

Por ejemplo, alguien que consulta sobre «cómo funciona la inteligencia artificial» y recibe una respuesta simplificada por parte de un modelo de IA puede creer que domina el tema, aunque en realidad solo tenga una comprensión superficial.

Este sesgo puede llevar a la toma de decisiones erróneas, especialmente en áreas críticas como la salud, la economía o la política, donde el conocimiento incompleto puede generar consecuencias negativas.

Sesgo de efecto halo

El sesgo de efecto halo ocurre cuando la percepción positiva de una característica de una fuente lleva a los usuarios a asumir que todo lo que proviene de ella es igualmente correcto o fiable.

En el contexto de la inteligencia artificial generativa, este sesgo puede ser especialmente problemático, ya que las personas pueden creer que una IA es infalible si ha proporcionado respuestas correctas en el pasado, sin cuestionar su precisión en otras áreas.

Por ejemplo, si un usuario consulta a una IA sobre física cuántica y recibe una respuesta bien formulada y técnicamente correcta, podría asumir que la IA también es experta en medicina, derecho o economía, incluso si el modelo no ha sido entrenado para responder con precisión en esos campos. Esto puede llevar a la difusión de información errónea si el usuario no verifica las fuentes o si la IA proporciona respuestas sesgadas o incompletas.

En escenarios más críticos, este sesgo puede influir en la confianza en IA aplicada a diagnósticos médicos, donde los pacientes pueden interpretar erróneamente que un

chatbot de salud tiene la misma precisión que un médico especializado. Si una IA genera un diagnóstico basado en patrones estadísticos, pero sin comprender síntomas complejos o factores individuales, la decisión del paciente podría verse gravemente afectada.

Este sesgo es especialmente problemático porque refuerza la confianza ciega en la IA, lo que puede llevar a una aceptación acrítica de respuestas incorrectas o mal fundamentadas.

Sesgo de ilusión de verdad

El sesgo de ilusión de verdad se basa en la idea de que cuanto más se repite una afirmación, más probable es que la gente la considere verdadera, independientemente de su veracidad.

En el caso de la IA generativa, este sesgo puede amplificarse debido a la capacidad de estos modelos para generar respuestas consistentes y repetitivas sobre un mismo tema.

Si una IA generativa produce información errónea en varias consultas, o si la misma narrativa aparece en múltiples plataformas impulsadas por IA, los usuarios pueden asimilar la información como cierta simplemente porque la han visto en varias ocasiones.

Este sesgo es utilizado en estrategias de propaganda y manipulación mediática, donde actores malintencionados pueden programar modelos de IA para repetir narrativas falsas en redes sociales, foros o noticias automatizadas. Por ejemplo, una IA puede ser utilizada para crear cientos de artículos falsos sobre un evento ficticio, logrando que la desinformación se convierta en un «hecho aceptado» simplemente por su presencia constante en el ecosistema digital.

El problema es que, en un mundo donde el volumen de información es abrumador, los usuarios rara vez verifican los datos en múltiples fuentes confiables. Si una IA genera y difunde un dato falso repetidamente, muchas personas pueden aceptarlo como verdad sin cuestionarlo.

Sesgo de normalidad

El sesgo de normalidad hace que las personas subestimen eventos poco comunes o inesperados, asumiendo que el futuro será similar al pasado y que los riesgos improbables no deben ser tomados en cuenta.

Cuando una IA generativa es entrenada con datos históricos, puede replicar este sesgo al descartar escenarios atípicos o emergentes simplemente porque son estadísticamente infrecuentes en su base de datos.

Por ejemplo, antes de la pandemia de COVID-19, muchos modelos de IA utilizados en predicción epidemiológica no consideraban seriamente el riesgo de una crisis sanitaria global de gran magnitud. Como resultado, algunas IA utilizadas para asesoramiento en políticas públicas subestimaron la gravedad del brote inicial, reflejando la falta de precedentes similares en sus datos de entrenamiento.

Este sesgo también afecta la toma de decisiones en seguridad y geopolítica. Si una IA entrenada en análisis de conflictos sugiere que una guerra es poco probable simplemente porque los datos históricos muestran décadas de estabilidad, puede llevar a gobiernos o empresas a ignorar señales de advertencia sobre crisis emergentes.

En el ámbito financiero, un modelo de IA podría minimizar la posibilidad de un colapso económico simplemente porque no ha ocurrido en muchos años, llevando a decisiones de inversión arriesgadas basadas en la falsa suposición de estabilidad.

El peligro del sesgo de normalidad en IA es que refuerza la idea de que los eventos extremos o poco comunes son irrelevantes, lo que puede hacer que la humanidad no esté preparada para riesgos emergentes que podrían tener consecuencias devastadoras.

Sesgo de información selectiva

El sesgo de información selectiva ocurre cuando la IA filtra respuestas de acuerdo con su entrenamiento y programación, lo que puede resultar en la omisión de información importante o en la falta de perspectivas alternativas.

Este sesgo es particularmente preocupante en motores de búsqueda basados en IA y en asistentes virtuales que priorizan ciertos tipos de información sobre otros. Si un usuario busca información sobre cambio climático, pero la IA ha sido entrenada para favorecer narrativas que minimizan la crisis ambiental, las respuestas que reciba estarán filtradas, lo que limita su acceso a una visión equilibrada del tema.

Del mismo modo, si una IA está programada para favorecer fuentes específicas de noticias o estudios científicos alineados con ciertas ideologías, los usuarios solo recibirán información parcial, sin conocer perspectivas que podrían ofrecer una imagen más completa y objetiva.

Este sesgo puede tener implicaciones graves en política, ciencia y economía, donde la disponibilidad de múltiples puntos de vista es crucial para la toma de decisiones informadas.

Sesgo de efecto de arrastre (Bandwagon Effect)

El sesgo de efecto de arrastre ocurre cuando las personas adoptan creencias o comportamientos simplemente porque parecen ser la opinión mayoritaria.

En el caso de la IA generativa, si un modelo sugiere que «la mayoría de la gente piensa X sobre un tema», los usuarios pueden inclinarse a aceptar esa posición sin evaluarla críticamente.

Por ejemplo, si una IA responde a una pregunta sobre política diciendo «según las tendencias actuales, la mayoría de los ciudadanos apoyan esta medida», los usuarios pueden sentirse influenciados a aceptar esa idea sin cuestionarla, incluso si los datos reales muestran una división de opiniones.

Este sesgo puede ser explotado por gobiernos o grupos de interés para manipular la opinión pública, generando la ilusión de que ciertos puntos de vista son más populares de lo que realmente son.

Sesgo de optimismo o pesimismo

Dependiendo de cómo esté entrenada la IA, sus respuestas pueden inclinarse hacia una visión excesivamente optimista o pesimista sobre ciertos temas.

Por ejemplo, si un usuario consulta sobre el impacto de la automatización en el empleo, una IA puede proporcionar solo información positiva, resaltando cómo la tecnología creará nuevas oportunidades, pero sin mencionar los riesgos de desempleo o la precarización laboral.

Por otro lado, un modelo entrenado con una visión pesimista podría hacer lo contrario, alarmando innecesariamente a los usuarios al presentar únicamente los efectos negativos del cambio tecnológico.

Este sesgo es problemático porque moldea la percepción de los usuarios, influenciando sus decisiones sin ofrecer un panorama completo de la realidad.

Sesgo de efecto marco (Framing Effect)

El sesgo de efecto marco ocurre cuando la forma en que se presenta la información influye en cómo se percibe e interpreta, afectando la toma de decisiones y las conclusiones del usuario.

En el caso de la IA generativa, el lenguaje, el tono y la estructura de una respuesta pueden guiar al usuario hacia una interpretación específica, incluso si la información en sí es neutral o incompleta.

Por ejemplo, si un usuario consulta sobre la seguridad de las vacunas y la IA responde con una frase como «Las vacunas son seguras y han salvado millones de vidas», la respuesta enfatiza el impacto positivo y minimiza cualquier posible riesgo. Sin embargo, si la IA reformula la misma información con «Algunas personas experimentan efectos adversos con las vacunas, aunque generalmente son seguras», la percepción del usuario sobre el tema puede volverse más negativa o dudosa.

Este sesgo es problemático porque puede ser explotado en la manipulación mediática y política, donde la IA es programada o ajustada para presentar la información de manera que influya en la percepción del público. En temas como economía, política internacional o conflictos armados, el uso de lenguaje emocional, adjetivos cargados y estructuras persuasivas puede hacer que la IA refuerce una visión sesgada sin que el usuario se dé cuenta.

Además, cuando una IA es utilizada por empresas o gobiernos, el sesgo de efecto marco puede emplearse para modificar opiniones sobre políticas públicas, regulaciones o iniciativas comerciales, simplemente alterando la manera en que se presentan los datos.

Sesgo de falsa consistencia

El sesgo de falsa consistencia ocurre cuando los usuarios asumen que si una serie de hechos parecen estar relacionados y tienen una estructura lógica deben llevar necesariamente a la misma conclusión, aunque en realidad los datos puedan ser incorrectos o no estén vinculados causalmente.

En la IA generativa, este sesgo se amplifica cuando un modelo proporciona respuestas estructuradas y bien organizadas en torno a un tema, creando una sensación de coherencia que puede inducir a error.

Por ejemplo, si un usuario pregunta sobre los efectos de la automatización en el desempleo y la IA responde con «La automatización ha eliminado trabajos en ciertas industrias. Las tasas de desempleo han aumentado en algunos sectores debido a la

automatización. Muchas empresas están reduciendo su plantilla por la eficiencia de los robots», el usuario puede asumir automáticamente que la automatización es la causa principal del desempleo, aunque en realidad pueda haber múltiples factores involucrados, como políticas económicas, globalización o cambios en la demanda del mercado.

Este sesgo es peligroso porque hace que los usuarios confíen en información incorrecta o incompleta simplemente porque parece coherente en su estructura. Además, la IA, al estar diseñada para generar respuestas lógicas y bien organizadas, refuerza este efecto, haciendo que la desinformación parezca aún más creíble.

Sesgo de efecto anclaje en predicciones

El sesgo de anclaje en predicciones ocurre cuando una predicción sobre un evento futuro influye en las expectativas de las personas, aunque no haya suficiente base para respaldarla.

En el contexto de la IA generativa, este sesgo puede ser problemático en sectores como economía, política y salud, donde las predicciones pueden alterar el comportamiento de los individuos o de instituciones enteras.

Por ejemplo, si un usuario consulta a una IA sobre el futuro del mercado de criptomonedas y la respuesta incluye «Se espera que Bitcoin supere los $100,000 en los próximos años debido a tendencias alcistas históricas», los inversionistas pueden modificar sus decisiones basándose en esta predicción, sin considerar que el modelo no tiene información actualizada ni la capacidad de prever con certeza el comportamiento del mercado.

Este sesgo también afecta la política. Si una IA genera respuestas como «Las encuestas muestran que el candidato X tiene una ventaja clara en la intención de voto», los votantes pueden sentirse desmotivados para apoyar a otros candidatos, afectando el resultado electoral.

Dado que las predicciones de la IA suelen estar basadas en patrones de datos pasados, pero sin la capacidad de anticipar eventos disruptivos, este sesgo puede generar una falsa sensación de certeza sobre el futuro, lo que lleva a errores en la toma de decisiones estratégicas.

Sesgo de causa-falsa (Post hoc ergo propter hoc)

El sesgo de causa-falsa ocurre cuando se asume que dos eventos están relacionados causalmente solo porque ocurrieron en secuencia o simultáneamente, sin que haya evidencia de que uno haya causado el otro.

Este sesgo es particularmente problemático en la IA generativa cuando analiza grandes volúmenes de datos históricos y genera afirmaciones basadas en correlaciones en lugar de relaciones causales comprobadas.

Por ejemplo, si una IA analiza datos climáticos y de criminalidad y encuentra que «en los últimos 50 años, el aumento de la temperatura global ha coincidido con una reducción en las tasas de criminalidad», podría sugerir que el calentamiento global está reduciendo el crimen, cuando en realidad ambos fenómenos pueden no estar relacionados.

Este tipo de error es común en informes estadísticos generados por IA, donde la correlación es presentada como evidencia causal, lo que puede inducir a interpretaciones equivocadas en áreas como política, salud pública y economía.

Un ejemplo grave sería si una IA en el ámbito de la salud concluye que «las personas que consumen más café tienen menor incidencia de enfermedades cardíacas», cuando en realidad la variable oculta podría ser que esas personas también llevan una dieta más saludable o hacen más ejercicio.

Este sesgo puede llevar a decisiones políticas, económicas y científicas equivocadas si los responsables no contrastan la información con investigaciones basadas en métodos estadísticos rigurosos.

Sesgo de validación social

El sesgo de validación social ocurre cuando las personas tienden a considerar válida una afirmación si creen que muchas otras personas la comparten, sin evaluar críticamente su veracidad.

En el contexto de la IA generativa, este sesgo se amplifica porque la IA puede reforzar ciertas narrativas repetidamente, creando la ilusión de que una afirmación es ampliamente aceptada, aunque sea falsa.

Por ejemplo, si un usuario busca información sobre un tema controversial como el impacto de la 5G en la salud, y la IA, basándose en datos sesgados, genera respuestas como «Muchas personas creen que la tecnología 5G tiene efectos adversos en la salud humana», el usuario podría interpretar esto como una señal de que la teoría tiene

fundamento científico, cuando en realidad no hay pruebas sólidas que la respalden.

Este sesgo es particularmente peligroso en temas de desinformación política, conspiraciones y propaganda, donde la IA puede crear la percepción de que una opinión es dominante simplemente por su repetición en diferentes plataformas.

Por ejemplo, si un modelo de IA es utilizado en redes sociales para generar contenido a favor de un candidato político y sus mensajes se replican constantemente en múltiples plataformas, los usuarios pueden sentir que «todo el mundo apoya a ese candidato», influyendo en su decisión de voto sin que haya evidencia de un apoyo real mayoritario.

Además, en temas como salud pública y pseudociencia, la IA puede amplificar mitos o creencias erróneas solo porque estas han sido ampliamente discutidas en internet, sin considerar la calidad de las fuentes originales.

Estos sesgos cognitivos no solo afectan nuestra interacción con la IA, sino que también pueden influir en la forma en que se desarrollan y entrenan estos sistemas. Los datos utilizados para entrenar modelos de IA pueden contener sesgos inherentes, lo que lleva a resultados distorsionados y potencialmente perjudiciales.

Por ejemplo, si los datos de entrenamiento reflejan prejuicios sociales o culturales, la IA puede perpetuar y amplificar estos sesgos en sus respuestas.

De la información a la manipulación. IA en la Guerra Cognitiva

Uno de los mayores riesgos de la inteligencia artificial generativa es su capacidad para **crear y** difundir información de manera masiva y sistemática, lo que la convierte en una herramienta poderosa para la manipulación de la percepción pública.

Gobiernos, actores políticos y corporaciones pueden aprovechar estos modelos para influir en la opinión social y consolidar narrativas artificialmente construidas.

A través de la saturación de internet con contenidos alineados con una determinada agenda, pueden posicionar ciertas ideas como verdades absolutas, desplazando o minimizando información que contradiga sus intereses.

Además, la IA generativa facilita la reescritura y alteración de la historia, permitiendo la creación de documentos falsificados, referencias manipuladas y pruebas fabricadas que pueden modificar la comprensión colectiva de eventos pasados. Al suprimir o distorsionar datos, estos sistemas no solo consolidan ideologías y refuerzan sesgos preexistentes, sino que también debilitan el pensamiento crítico y la capacidad de la sociedad para discernir entre la realidad y la manipulación informativa.

La IA generativa facilita la creación de realidades paralelas en las que ciertos eventos pueden ser exagerados, minimizados o completamente fabricados, generando desconfianza en la información tradicional.

La IA generativa ya ha sido utilizada en campañas políticas y geopolíticas para amplificar mensajes específicos, generar ataques personalizados y crear narrativas falsas. Según un informe de la UNESCO (2024), existe el riesgo de que la IA sea utilizada para reescribir eventos históricos y manipular crisis internacionales, facilitando la polarización y el control ideológico.

Uno de los casos más preocupantes es el intento de alterar registros históricos con información generada por IA. La UNESCO ha advertido sobre el peligro de que los LLMs distorsionen eventos como el Holocausto, reemplazando documentos verificables con narrativas artificiales que minimicen o modifiquen hechos reales.

Este riesgo es especialmente alarmante cuando la IA es utilizada como herramienta educativa o de investigación, ya que una generación expuesta a información manipulada podría construir su conocimiento sobre bases fraudulentas sin siquiera ser consciente de ello.

En el ámbito militar, la inteligencia artificial se ha convertido en una herramienta clave para las operaciones de guerra cognitiva, donde el objetivo no es la conquista de territorios, sino el control de la percepción y la manipulación de la realidad.

A través de técnicas avanzadas, la IA puede emplearse para desacreditar líderes políticos y generar crisis diplomáticas mediante la creación de deepfakes, falsificando imágenes, audios y videos con una precisión tal que pueden resultar indistinguibles de la realidad.

Además, su capacidad para infiltrar redes sociales con bots generativos permite moldear la opinión pública, amplificando narrativas específicas y reforzando ideologías a través de interacciones automatizadas que imitan el comportamiento humano.

Otro de sus usos más preocupantes es la fabricación de pruebas falsas destinadas a justificar intervenciones militares, sanciones económicas o conflictos estratégicos, alterando documentos, grabaciones y otros registros para manipular a la comunidad internacional.

En este escenario, la IA deja de ser simplemente una herramienta de propaganda para convertirse en un actor estratégico en la geopolítica global, con el potencial de desestabilizar gobiernos, influir en decisiones de organismos internacionales y reconfigurar el equilibrio de poder sin necesidad de recurrir a la fuerza militar tradicional. Su capacidad para generar y difundir desinformación de manera masiva y precisa la convierte en un arma de guerra silenciosa, pero de un impacto potencialmente devastador.

Conclusión

GPTs, DeepSeeks y otros modelos de IA generativa representan un avance sin precedentes. Los avances en inteligencia artificial generativa han abierto un mundo de posibilidades para la automatización de la información y la producción de contenido. Sin embargo, estos mismos avances presentan una amenaza para la integridad del conocimiento humano, especialmente cuando se combinan con la creciente tendencia de confiar en la IA sin una evaluación crítica.

Uno de los mayores peligros de la IA generativa es su capacidad para reforzar sesgos cognitivos y manipular la percepción pública a gran escala. Desde la reescritura de eventos históricos hasta la creación de deepfakes hiperrealistas, las aplicaciones de la IA en la manipulación de la información pueden tener consecuencias devastadoras en la política, la economía y la seguridad global. La guerra cognitiva, la influencia en procesos democráticos y la alteración de la memoria colectiva son solo algunos de los riesgos asociados con su uso irresponsable.

Para evitar que la inteligencia artificial generativa se convierta en una herramienta de control ideológico y desinformación masiva, es fundamental desarrollar mecanismos de regulación, auditoría y verificación de contenido generado por IA. Una de las estrategias clave para mitigar estos riesgos es la regulación internacional y ética de la IA, estableciendo normas globales que limiten el uso malintencionado de estas tecnologías en la manipulación de la información y la propaganda.

Asimismo, la educación en pensamiento crítico debe convertirse en una prioridad dentro de los sistemas educativos, promoviendo la alfabetización digital y mediática para garantizar que los ciudadanos puedan evaluar y contrastar la información de manera efectiva. En paralelo, es crucial el desarrollo de herramientas de verificación y transparencia que permitan identificar si un contenido ha sido generado o manipulado por IA, asegurando la trazabilidad y la fiabilidad de la información que circula en medios digitales.

Por último, la responsabilidad de las empresas tecnológicas es un factor clave en esta regulación. Las compañías que desarrollan modelos de IA deben implementar auditorías de sesgo y mecanismos de seguridad para prevenir la generación y difusión de contenido engañoso o manipulado. Solo con un enfoque coordinado entre gobiernos, instituciones educativas y el sector tecnológico será posible garantizar un uso ético y responsable de la IA generativa, protegiendo la integridad del conocimiento y la autonomía del pensamiento humano.

Si no se toman medidas efectivas, nos enfrentamos al riesgo de entrar en una era donde la verdad se vuelve un concepto maleable, moldeado por algoritmos que pueden ser programados para reforzar ciertas narrativas mientras suprimen otras. La inteligencia artificial generativa puede ser una herramienta poderosa para el progreso y la innovación, pero sin un marco de regulación adecuado y un uso crítico por parte de los usuarios, su potencial para distorsionar la realidad y erosionar la confianza en la información legítima podría superar sus beneficios.

El desafío de nuestra era no es solo mejorar la tecnología, sino garantizar que su uso esté alineado con los principios de transparencia, ética y acceso equitativo al conocimiento. Solo así podremos aprovechar el poder de la IA generativa sin comprometer la integridad de la información y la autonomía del pensamiento humano.

Referencias

- Brundage, M., Avin, S., Wang, J., Belfield, H., Krueger, G., Hadfield, G., & Amodei, D. (2018). *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation*. Future of Humanity Institute.
- Colado, S. (2020). *La influencia de la tecnología en el desarrollo del pensamiento y la conducta*. Independently published by Amazon.
- Colado, S. (2021). *Multiversos digitales: La tecnología como palanca evolutiva*. Universo de las Letras
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). *Generative adversarial networks*. Advances in Neural Information Processing Systems.
- (2023). *Artificial intelligence in law enforcement: Opportunities and risks*. International Criminal Police Organization.
- Mitchell, M. (2021). *Artificial Intelligence: A Guide for Thinking Humans*. Farrar, Straus and Giroux.
- Moosavi-Dezfooli, S., Fawzi, A., & Frossard, P. (2016). *DeepFool: A simple and accurate method to fool deep neural networks*. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2574-2582.
- Russell, S., & Norvig, P. (2021). *Artificial Intelligence: A Modern Approach* (4th ed.). Pearson.
- Schneier, B. (2019). *Click Here to Kill Everybody: Security and Survival in a Hyper-connected World*. W. W. Norton & Company.
- Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2014). *Intriguing properties of neural networks*. *International Conference on Learning Representations (ICLR)*.
- (2023). *Artificial Intelligence and Misinformation: Risks and Policy Responses*. United Nations Educational, Scientific and Cultural Organization.