La búsqueda de información con Inteligencia Artificial: desafíos éticos

2ª Jornada de Búsqueda de Información

Biblioteca del Instituto Universitario de la Policía Federal Argentina

Nicolás Petrosini

Universidad de Palermo

npetro@palermo.edu

Resumen:

En esta ponencia se abordará la importancia de considerar aspectos éticos al utilizar grandes modelos de lenguaje (LLMs) en la búsqueda de información dentro de entornos académicos y de investigación. La presentación tiene como objetivo principal que los asistentes comprendan la importancia de validar la veracidad de la información recuperada, identifiquen el sesgo inherente a estos modelos, reconozcan adecuadamente el uso de un LLM en sus investigaciones y se concienticen sobre la importancia de resguardar los datos privados.

Palabras clave:

Búsqueda de información, Inteligencia Artificial Generativa, Ética

Introducción

En la mañana del 13 de junio de 2018, un millón y medio de personas en Hawái recibió un mensaje de alerta en sus celulares. Un misil norcoreano iba a impactar en ese territorio en 15 minutos. A partir de ese momento, el caos se apoderó de la población. Autos abandonados en autopistas, residentes y turistas buscando refugio con desesperación. Llamadas al 911 colapsadas. Incluso, algunas personas grabaron mensajes de despedida para sus seres queridos.

Sin embargo, 40 minutos más tarde se informó que la alerta era falsa. Todo había sido causado por un empleado del gobierno que, durante un simulacro, creyó que la amenaza era real y presionó el botón equivocado. Y, como el sistema no tenía los controles necesarios, propagó la alerta por todo Hawái. Además, este error se amplificó cuando las personas compartieron el mensaje masivamente en sus redes sociales.

Esta historia nos demuestra lo que puede suceder cuando no verificamos la información adecuadamente. Lo mismo ocurre con el uso de herramientas de Inteligencia Artificial generativa, como ChatGPT, donde la veracidad de la información no siempre está garantizada. Además, al usar estos modelos, nos enfrentamos a otros desafíos éticos: los sesgos, la integridad académica y la privacidad de los datos. En esta presentación vamos a explorar estrategias para mitigarlos al utilizar IA en la búsqueda de información.

Veracidad de la información

Hablemos de la veracidad. Los motores de búsqueda tradicionales como Google extraen información de millones de sitios web y muestran resultados concretos. En cambio, los modelos de lenguaje como ChatGPT predicen las palabras más coherentes y convincentes en un contexto. Esto significa que no siempre proveen información verdadera, sino verosímil.

Por ejemplo, si realizamos la misma búsqueda en Google varias veces, es muy probable que obtengamos los mismos resultados. Pero, si usamos un modelo de IA, podríamos recibir respuestas distintas cada vez, incluso si la pregunta es la misma.

Otro problema es que los modelos de IA, al solicitar que den las referencias bibliográficas en las que se basaron para elaborar una respuesta, pueden inventarlas. Porque sus respuestas no provienen de una base de datos.

Entonces, ¿de qué forma podemos validar la información que obtenemos de la IA? Una estrategia consiste en aplicar nuestra experiencia del tema sobre el cual interrogamos al modelo. Es decir,

si somos expertos en la materia. Este diagrama de flujo, adaptado de un informe de la UNESCO, define cuándo es seguro utilizar la IA Generativa y cuándo no.

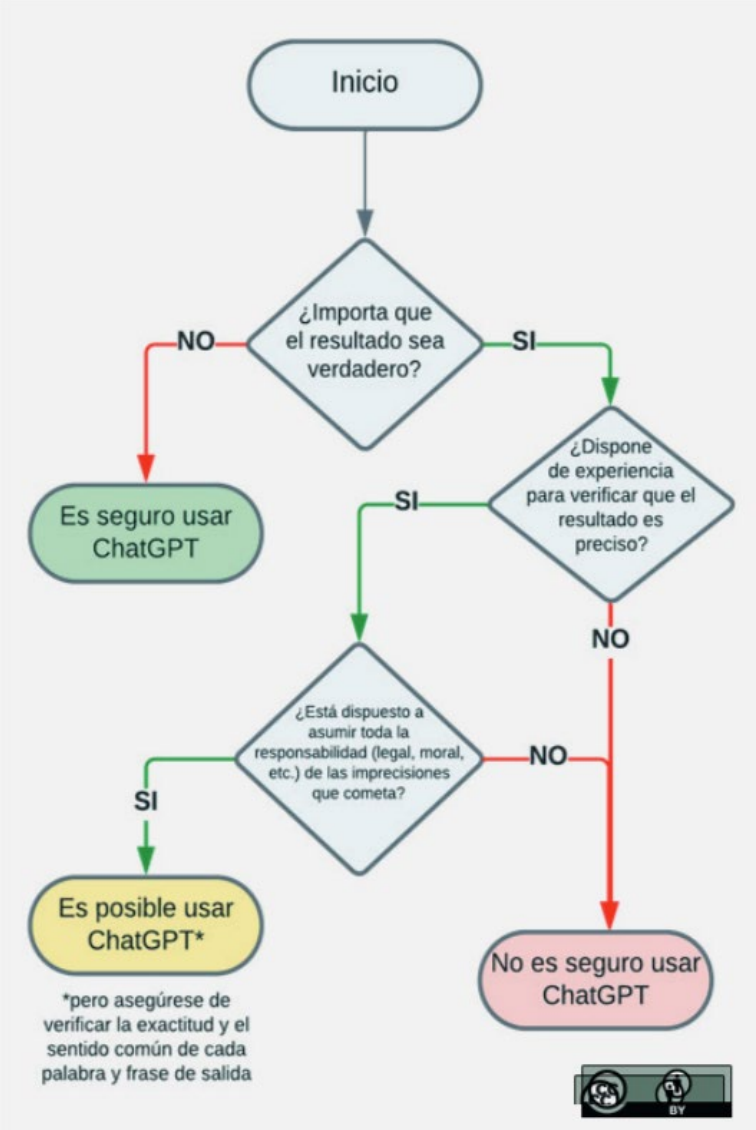

Ilustración 1: Diagrama de flujo elaborado por Aleksandr Tiulkanov, AI and Data Policy Lawyer, enero de 2023. UNESCO (2023) CC BY 3.0 IGO

Lo primero que tenemos que preguntarnos es si el resultado que esperamos obtener debe ser verdadero. Si no lo es, es seguro usar la IA. En cambio, si tiene que ser verdadero, nos podemos preguntar si, desde nuestra experiencia sobre la materia, podemos verificar la precisión del resultado. Si es así, tenemos que asumir la responsabilidad de la respuesta, legal y moral de las imprecisiones que el modelo pueda cometer.

Otra opción, para los entornos académicos o de investigación, es explorar diferentes herramientas de IA. Existen algunas que fueron entrenadas solamente con literatura científica.

Por ejemplo, Consensus. Es un motor de búsqueda potenciado con IA, que permite conversar sobre los artículos, hacer preguntas, elaborar resúmenes, etc.

Sesgos

El segundo desafío en el uso ético de la Inteligencia Artificial Generativa son los sesgos. ¿Qué es un sesgo? La tendencia o inclinación hacia algo o alguien. Algunos sesgos pueden ser positivos, por ejemplo, cuando elegimos comer solo alimentos saludables. Pero otros pueden estar basados en estereotipos, en vez de en un conocimiento real. En definitiva, los sesgos son atajos de la mente que pueden devenir en prejuicios.

Y las máquinas, la IA, no es neutral. Reproduce valores e ideologías. Porque aprende de nosotros. Para poder responder, primero tienen que entrenarse con datos, de libros, artículos, foros, etc., los cuales suelen provenir de una de las más importantes creaciones de la humanidad: internet. Y como sabemos, en internet hay información correcta e incorrecta.

Y no solo las fuentes de aprendizaje son las causantes de los sesgos en los modelos de IA. Una vez que terminan su entrenamiento, empiezan a tomar sus propias decisiones y lo que aprenden de estas se añade a sus datos. Es decir, la IA se realimenta de lo aprendido. Y si ha sido entrenada con datos sesgados, puede amplificarlos, resultando en un mensaje distorsionado para el usuario.

Les voy a contar un hecho muy llamativo que ocurrió hace unos años. La empresa Amazon desarrolló un algoritmo para seleccionar el personal de manera autónoma, con el fin de optimizar recursos humanos y ahorrar tiempo. El programa tenía que filtrar los cinco mejores currículums entre 100 candidatos. Para eso revisaba las solicitudes y asignaba un puntaje de 1 a 5 estrellas a cada uno. Pero el programa resultó ser machista: prefería la contratación de hombres por sobre mujeres, especialmente para puestos técnicos. Esto ocurrió porque fue entrenado con las solicitudes acumuladas de diez años, en las cuales prevalecía la elección de hombres. Por lo cual, como dice un artículo de la BBC, “se enseñó a sí mismo que los candidatos masculinos eran una preferencia”. Es decir, aprendió y replicó un sesgo presente en los datos. Cuando los reclutadores de Amazon se dieron cuenta de lo que estaba ocurriendo, ¡el algoritmo fue despedido! Entonces, volviendo a la búsqueda de información con IA; ¿cómo podemos mitigar los sesgos?

Una estrategia que podemos aplicar es comparar la información generada por el modelo con fuentes confiables y verificables. Es decir, tomar la respuesta como un punto de partida y no como un producto final. Para eso tenemos que acostumbrarnos a reflexionar y evaluar las respuestas al usar la IA. Los investigadores y bibliotecarios debemos entrenarnos para evaluar la información generada por estos sistemas.

Integridad académica

El tercer tema que vamos a tratar en esta charla es la integridad académica en el uso de la Inteligencia Artificial Generativa. ¿Qué es la integridad? El compromiso de actuar con honestidad y responsabilidad en una comunidad académica. Entre otras conductas, incluye la presentación de trabajos genuinos y el reconocimiento de las ideas de otros mediante las citas.

De acuerdo a la UNESCO, la posibilidad de copiar las respuestas de los modelos de lenguaje y no atribuirlas es la principal preocupación que se ha despertado en el ámbito académico, en particular desde la irrupción de ChatGPT en noviembre de 2022.

Por otra parte, en la investigación científica, la irrupción de la IA Generativa ha causado una verdadera revolución. Por ejemplo, hubo casos en que se incluyó a ChatGPT como autor de artículos. Pero revistas como Nature han modificado sus políticas editoriales, prohibiendo dicha atribución. Por otro lado, publicaciones como The Lancet permiten el uso del chat, aunque establecen en sus políticas que "debe limitarse a mejorar la legibilidad y el lenguaje del trabajo, lo cual debe declararse en el manuscrito." O sea, hay posturas que prohíben ciertos usos de la IA Generativa y otras que lo incorporan, pero con límites muy claros. En definitiva, debemos recordar que la autoría implica responsabilidad por nuestro trabajo, lo cual no puede ser asumido por una IA.

La propuesta que quiero compartir con ustedes hoy es integrar el uso de la Inteligencia Artificial siendo transparentes. Lo primero que debemos hacer es informarnos sobre las políticas, ya sean de una publicación o institucionales. Luego, si necesitamos reconocer el uso de la IA, podríamos aplicar una norma.

En esta presentación, lo voy a ejemplificar con las Normas APA, a partir de las cuales se han recomendado diferentes alternativas. En primer lugar, en un trabajo de investigación se puede describir cómo se utilizó un modelo de IA en el apartado de metodología. Si es un ensayo u otro texto más breve se puede aclarar en la introducción. Se puede incluir el prompt, es decir, la instrucción que damos al modelo, en el texto del trabajo; y cualquier porción significativa de la respuesta que nos haya brindado el chat, reconociéndolo como una cita textual. Esto último es muy importante, porque los modelos son entrenados con nuestra producción, que está en internet, pero las empresas que los desarrollan no suelen revelar las fuentes. Entonces, ¿cómo sabemos que no estamos plagiando a un autor?

También, se puede aplicar el formato de citas y referencias: para más detalles, les recomiendo consultar el blog oficial de APA. Para finalizar con este punto, les voy a hacer una confesión

personal. El texto que escribí en esta charla es de mi autoría, pero pasó por la revisión de ChatGPT. Y gracias al chat, eliminé párrafos enteros que seguramente los hubiera aburrido. Otras partes estaban bien, a pesar de la opinión del chat: por lo cual los dejé.

Privacidad de datos

Y el último desafío de la ética en la búsqueda de información con IA, que les voy a explicar hoy, es la importancia de resguardar los datos. Datos que hacen, por ejemplo, a nuestra identidad, pero también, a la de otras personas, de las cuales, por ejemplo, podemos recolectar información para una investigación. Incluso, puede ser también información confidencial de nuestro empleador. Por ejemplo, en el contexto empresarial, un informe de nuevos productos que no salieron al mercado y que subimos a una herramienta de IA para que lo analice. Si estos datos pudieran ser conocidos de antemano por la competencia, se perdería la ventaja competitiva.

En el caso de ChatGPT, nuestras conversaciones y los archivos que subimos pueden ser visualizados por empleados de OpenAI, la empresa que desarrolla el chat; y también por proveedores externos "de confianza". Entre otras razones, porque una forma de mejorar a los modelos es suministrándoles datos de nuestras conversaciones para aumentar su precisión al dar una respuesta. Es decir, que cuando utilizamos la IA generativa ayudamos a que el modelo se desarrolle más rápido.

Esta cuestión de lo que hacen las compañías que elaboran los modelos con nuestros datos ha provocado grandes controversias. Una de las más conocidas ocurrió en abril del año pasado. Italia fue el primer país occidental en prohibir el uso de ChatGPT en su territorio, porque se reveló que hubo una filtración de las conversaciones de los usuarios. Es importante mencionar que nuestras conversaciones quedan guardadas en un servidor; por lo tanto, pueden ser hackeadas. O tal vez no haya que llegar a ese extremo. Nuestros datos pueden ser utilizados, por ejemplo, para influenciar en nuestras decisiones, desde qué productos compramos hasta qué candidatos votamos en una elección presidencial. Sí, en el siglo XXI, los datos cotizan más que el petróleo. En el caso italiano, ¿saben cómo terminó esta historia? OpenAI tuvo que agregar las sesiones de incógnito, similares a las que podemos utilizar en los navegadores web, y creó suscripciones empresariales que no utilizan los datos para entrenamiento del modelo.

Y si tenemos una cuenta gratuita, ¿qué medidas podemos tomar para minimizar el riesgo de perder nuestros datos? En el caso de ChatGPT, en la configuración es posible deshabilitar la opción para aportar datos para entrenamiento. Otra sugerencia que les puedo hacer es suponer que los prompts que ingresemos al modelo pueden ser, en algún momento, públicos, y si hay

datos que no nos gustaría revelar, eliminarlos, de forma que nuestro prompt sea más impersonal. Tal como podemos hacer con las redes sociales. Hay cosas que publicamos abiertamente, pero otras no.

Conclusión

Como aprendimos en esta presentación, el uso de los grandes modelos de lenguaje para propósitos académicos o de investigación implica ser conscientes de los aspectos éticos. Entre ellos, mencionamos comprobar la veracidad de la información, analizar la presencia de los sesgos, respetar la integridad académica y resguardar los datos. Para mitigarlos, propusimos diferentes estrategias.

Y me gustaría compartir una idea de dos autores, Sigman y Bilinkis. Si la Inteligencia Artificial está inspirada en nuestro cerebro y es entrenada con nuestros datos, ¿por qué tenemos la expectativa de que no cometa errores? ¿Por qué les exigimos un grado de perfección que nosotros no alcanzamos? "En el proceso de incluir más y más IA en la vida cotidiana, es importante saber que, al igual que nosotros, se va a equivocar."

Lo importante es reconocer nuestro rol al usar la IA, de manera que no toquemos el botón equivocado.

Referencias

2018 Hawaii false missile alert. (2024). Wikipedia. Recuperado 4 de octubre de 2024, https://en.wikipedia.org/w/index.php?title=2018_Hawaii_false_missile_alert&oldid=1236563628

Arévalo, J. A. (2024, julio 31). *La IA complica el plagio. ¿Cómo deben responder los científicos?* Universo Abierto. https://universoabierto.org/2024/07/31/la-ia-complica-el-plagio-como-deben-responder-los-cientificos/

ChatGPT: Italia se convierte en el primer país occidental en bloquear el acceso al programa de inteligencia artificial. (2023, 31 marzo). BBC News Mundo. Recuperado 4 de octubre de 2024, de https://www.bbc.com/mundo/noticias-65142505

El algoritmo de Amazon al que no le gustan las mujeres. (2018, octubre 11). BBC News Mundo. https://www.bbc.com/mundo/noticias-45823470

Garay, J. (2023, marzo 31). *Italia se convierte en el primer país en prohibir la IA de ChatGPT*. Wired. https://es.wired.com/articulos/italia-bloquea-chatgpt-por-riesgos-a-la-proteccion-de-datos-personales

Hervieux, S., & Wheatley, A. (Eds.). (2022). Ethical Implications of Implicit Bias in AI: Impact for Academic Libraries. En *The Rise of AI: Implications and Applications of Artificial Intelligence in Academic Libraries* (pp. 165-174). Association of College and Research Libraries.

International Center for Academic Integrity. (2014). *The Fundamental Values of Academic*. Edición del Autor. https://academicintegrity.org/images/pdfs/20019_ICAI-Fundamental-Values_R12.pdf

Lopezosa, C. (2023). La Inteligencia artificial generativa en la comunicación científica: Retos y oportunidades. *Revista de Investigación e Innovación en Ciencias de la Salud, 5*(1), 1-5. https://doi.org/10.46634/riics.211

Marquez, J. (2024, junio 13). *Cuando usamos ChatGPT en el trabajo estamos exponiendo la información de nuestra empresa: Así podemos evitarlo*. Xataka. https://www.xataka.com/robotica-e-ia/cuando-usamos-chatgpt-trabajo-podemos-estar-exponiendo-informacion-confidencial-esto-podemos-hacer-para-evitarlo

McAdoo, T. (2023, abril 7). *How to cite ChatGPT*. APA Style Blog. https://apastyle.apa.org/blog/how-to-cite-chatgpt

Sesgo. (s. f.). Psychology Today en español. Recuperado 22 de septiembre de 2024, de https://www.psychologytoday.com/es/fundamentos/sesgo

Sigman, M., & Bilinkis, S. (2023). La moral de un algoritmo. En *Artificial: La nueva inteligencia y le contorno de lo humano* (pp. 169-185). Debate.

Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura. (2023). *ChatGPT e inteligencia artificial en la educación superior: Guía de inicio rápido*. Edición del Autor. https://unesdoc.unesco.org/ark:/48223/pf0000385146_spa

Valenzuela, G. U. (2023). El desafío del uso de inteligencia artificial para la elaboración de la literatura científica: El caso de ChatGPT, un debate abierto. *Cuadernos médico sociales*, *63*(1), 27-31. https://dialnet.unirioja.es/servlet/articulo?codigo=9510768