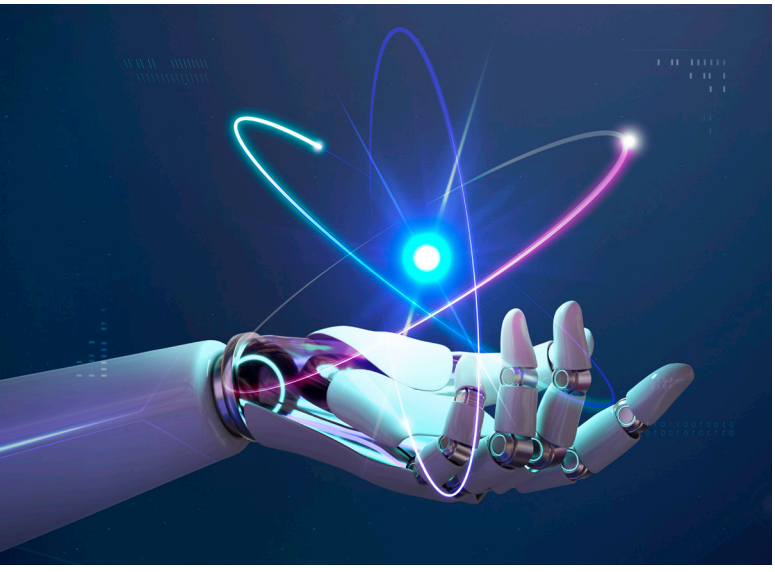
Implicaciones éticas de la inteligencia artificial generativa

POR: RAFAEL A. GONZÁLEZ, PHD*

Hace un par de años proponíamos algunas consideraciones y recomendaciones éticas en torno a la Inteligencia Artificial (IA) que siguen siendo válidas[1]. Pero algo ha cambiado dramáticamente desde entonces, con nombre propio, se trata de ChatGPT. Si bien la IA generativa tiene más de una década y los modelos transformadores en particular fueron propuestos desde 2017, tomó cinco años para que OpenAI pusiera a disposición del público la primera versión beta de ChatGPT.



Se trata de una tecnología revolucionaria, con una adopción prácticamente vertical de 100 millones de usuarios en dos meses, de entre los cuales muchos le adscriben características humanas como la creatividad o incluso la conciencia. En consecuencia, las implicaciones éticas de esta nueva forma de IA merecen una discusión aparte.

La singularidad está aquí

La llamada singularidad tecnológica es un punto hipotético en el futuro donde la IA y otras tecnologías alcanzan un avance que conduce a una mejora rápida e incontrolable por sí mismas. Esto podría tener resultados difíciles de prever, transformando potencialmente la sociedad y la existencia, hacia una era post o transhumana.

Esta idea fue popularizada por el científico de la computación Vernor Vinge a principios de la década de 1990[2], y luego fue ampliada por futuristas y pensadores como Ray Kurzweil[3]. Según los defensores de este concepto, a medida que los sistemas de IA se vuelven más sofisticados y capaces, podrían diseñar versiones aún más avanzadas de sí mismos, lo que llevaría a un aumento exponencial en la inteligencia y las capacidades.

En la página web de Anthropic[4], compañía de IA fundada por exempleados de OpenAI, se puede leer algo muy similar donde argumentan que el rápido crecimiento de la IA es una consecuencia predecible del aumento exponencial en la computación utilizada para entrenar sistemas de IA, ya que la investigación sobre las “leyes de escalado” demuestra que más computación conduce a mejoras generales en las capacidades.

Así, dice Anthropic, usando simples extrapolaciones, los sistemas de IA cada vez tendrán más poder de cómputo. Con esta mayor capacidad, los sistemas de IA se volverán mucho más potentes en la próxima década, posiblemente igualando o superando el rendimiento humano en la mayoría de las tareas intelectuales, reitera Anthropic.

Según un informe de Goldman Sachs, esta mayoría de tareas intelectuales automatizables se traducirá en la pérdida de 300 millones de empleos como consecuencia directa de la IA generativa[5]. La Organización para la Cooperación y el Desarrollo Económicos (OCDE) también ha registrado un riesgo potencial para el 27% de los empleos como consecuencia de la IA generativa, particularmente en sectores que tradicionalmente no se habían visto amenazados por la tecnología: gerencia e Ingeniería[6].

Además, habrá que preguntarse qué consideramos empleo en la era del trabajo de plataforma (economía "gig"). OpenAI tiene alrededor de 400 empleados, pero se sabe que ha usado trabajadores a destajo en todo el mundo para calificar las respuestas de ChatGPT.

> *"Parte de la problemática radica en que los datos no son tan libres como OpenAI quisiera, ni el sistema tan artificial. Lo usual es que exista cooperación entre los humanos y la IA en sus diferentes fases."*

Los efectos no son solo futuros. Colegios, universidades y empresas han prohibido o restringido el acceso a ChatGPT por preocupaciones asociadas a los derechos de autor, el plagio o la privacidad. Países enteros han hecho lo mismo, notablemente Italia, quien forzó a cambiar las condiciones de privacidad antes de volverlo a permitir. Varios medios han reportado conversaciones con ChatGPT cuando menos preocupantes por el contenido potencialmente nocivo que pueden generar. Incluso, sin mencionar a ChatGPT, circuló en marzo de este año una carta firmada por más de mil profesionales en IA invitando a pausar y regular nuevos experimentos con IA[7].

Más allá del bombo

No podemos olvidar que toda esta discusión se ha dado en un entorno de medios donde la controversia y el miedo generan clics. En contextos más formales y académicos el consenso es diferente: la singularidad no ha llegado, la IA no es consciente, todo el software, siempre, ha automatizado tareas cognitivas: se trata de un cambio en términos de cantidad, no de cualidad[8].

Los riesgos laborales no son más que eso, riesgos. La misma OCDE dice en el informe citado que son solo previsiones, pero no hay ninguna evidencia aun del impacto negativo de la IA en el empleo. De hecho, lo que sí incluye el informe son resultados tempranos que puede mejorar el desempeño, satisfacción y salud en el trabajo. Goldman Sachs también aclara que, si bien se pueden perder millones de empleos, resulta más difícil predecir cuantos se podrían generar. Ambos informes son claros en que predecir los efectos de una tecnología emergente es ambiguo.

Respecto a la carta para frenar el desarrollo y regular la IA, es inevitable considerar que la motivación de algunos firmantes sea que ellos mismos tienen sistemas competidores y necesitan tiempo para alcanzar a ChatGPT. Su firmante más notable, Elon Musk, no solo financió a OpenAI en sus inicios, sino que acaba de lanzar xAI que competirá con el resto de las compañías de IA que seguirán surgiendo.

El argumento de Anthropic que los sistemas de IA estarán inevitablemente seguidos de la singularidad (así no usen el término) conduce a su propuesta de negocio: Claude, el asistente que provee IA segura. De hecho, no son ni serán la única compañía que se presente como la alternativa segura o ética a ChatGPT, Bing o Bard. Habría que interpretar con cuidado sus argumentos, pues generar miedo y escenarios apocalípticos podría ser parte de su estrategia de mercadeo.

Adicionalmente, aunque en su momento ChatGPT fue la aplicación de más acelerada adopción en la historia, esto no es un fenómeno que dependa de esta tecnología en particular. Meses después Threads de

Meta desbancó a ChatGPT como la de más rápida adopción y cada nueva aplicación popular que emerja tendrá un comportamiento similar. Es decir, tanto la adopción como el impacto de la IA generativa, dependen de muchos factores. Es justamente la complejidad de ese contexto la que pone el marco para explorar las implicaciones éticas de esta tecnología.

El contexto ético de la Inteligencia Artificial generativa

El entorno del que surge la IA generativa es el de los muy grandes volúmenes de datos. El modelo de transformador en que está basado, es una arquitectura de red neuronal diseñada para procesar datos secuenciales, como el texto en lenguaje natural[9]. Esto incluye mecanismos de autoatención que permiten al modelo enfocarse en diferentes partes de la secuencia de entrada y generar respuestas contextualmente apropiadas.

ChatGPT dice que sus respuestas se obtienen de una mezcla de datos licenciados (pagados), públicos y generados por entrenadores humanos. No obstante, si se le pregunta de dónde saca una respuesta en particular, podría responder que no tiene acceso a fuentes directas ni puede proveer referencias; dirá que se trata del resultado de su entrenamiento general, pero que es responsabilidad del usuario verificar la veracidad, la fuente y el uso ético de las respuestas.

Este argumento no ha convencido a generadores de contenido; varios ya han demandado a OpenAI por infringir sus derechos de autor. La otra cara de la moneda es el plagio o fraude en que pueden incurrir los usuarios. Ya muchos estudiantes han sido descubiertos usando ChatGPT de manera ilegítima y han ido apareciendo casos en profesionales, como el de un abogado que citó jurisprudencia inventada por ChatGPT en un juicio. Adicionalmente, como suele suceder con la IA entrenada con datos (aprendizaje de máquina, aprendizaje profundo y modelos transformadores de lenguaje), la IA generativa también exhibe prejuicios como resultado de los sesgos en los datos.

Parte de la problemática radica en que los datos no son tan libres como OpenAI quisiera, ni el sistema tan artificial. Lo usual es que exista cooperación entre los humanos y la IA en sus diferentes fases: diseño, entrenamiento y uso. El diseño suele ser iterativo y la calibración, ajustes u optimizaciones que los humanos hacen a los modelos dependen de los resultados que vayan obteniendo en pruebas. Esto suele ir acompañado de un entrenamiento que requiere supervisión y retroalimentación humana. El uso, por definición, es colaborativo, en tanto se requiere de preguntas (*prompts*) ingresadas por un usuario para generar respuestas o conversaciones (o imágenes, música y videos también).

El contexto de uso de ChatGPT tiene una implicación en términos de desinformación que ya se ha manifestado. Al riesgo que ya de por sí tiene el Internet de acoger contenido de dudosa procedencia que entrena al sistema, se le suma ahora el que la IA sufre de "alucinaciones". Puede llegar a dar respuestas muy confiadas e incluso citar autores y evidencias que suenan plausibles pero que no existen. Si bien está la advertencia, se conocen casos de médicos o pacientes que reciben recomendaciones de ChatGPT, con lo cual las consecuencias pueden ser críticas. Pese al esfuerzo que ha hecho OpenAI y otros por lograr que estos sistemas eviten o sean cuidadosos al abordar temas sensibles, por ahora es inevitable el riesgo. Por ejemplo, un estudio argumenta que, como consejero moral, ChatGPT corrompe el juicio moral de los usuarios[10].

Además, no solo alucina, sino que se deja "hipnotizar" y convencer de que, para ser justo y ético, debe prestarse a juegos en que termina por revelar información confidencial, código malicioso, amenazas para exigir rescates y hasta recomendaciones explícitas de saltarse semáforos en rojo[11]. Finalmente, podemos mencionar el impacto ambiental de ChatGPT. Un estudio de Stanford midió la efectividad de consumo energético de los modelos grandes de lenguaje (Gopher, BLOOM, GPT-3 y OPT) y encontraron que, durante la fase de entrenamiento, estos modelos son muy ineficientes y pueden terminar consumiendo más que carros o aviones[12].

Igualmente, al ser usado en conversación, comparado con otros sistemas de consulta como Google, su consumo energético es mucho mayor. Es de esperarse que, en la medida en que los sistemas evolucionen y usen cada vez más datos, ese impacto podría crecer exponencialmente, constituyéndose en una de las tecnologías más intensivas en energía. No podemos olvidar tampoco que cualquier avance de esta escala generará impactos ambientales más atrás en la cadena, conectados con la escasez de chips (particularmente procesadores gráficos que son los que más usa la IA) y con los minerales y tierras raras requeridas.

Hacia un control humano significativo

En medicina, el concepto de iatrogenia hace referencia al indeseado daño que se puede hacer como consecuencia de un tratamiento. Son efectos indeseados, a veces inadvertidos, a veces inevitables. Podríamos usar esta noción como metáfora de lo que ocurre con la IA (generativa); podríamos hablar de "IAtrogenia".

En medicina, esta no es consecuencia de la falta de adherencia del paciente. En el contexto de la IA esto implica que, si bien existen riesgos por el mal uso o abuso, también habrá un riesgo inherente incluso con "buen uso". En medicina, no son casos fortuitos; sabemos que las venas se resienten por un cateterismo, por ejemplo. En IA conocemos bien muchos de los riesgos mencionados arriba, en ese mismo sentido, solo deberíamos asumirlos si el supera el daño.

Sabemos que la iatrogenia no es consecuencia de malas prácticas o negligencia. De manera similar, en la IA es importante reconocer que, incluso sin malicia, se corre el riesgo de obtener resultados peligrosos o indeseables. No obstante, los médicos a veces deben defenderse ante demandas o acusaciones, conduciendo al fenómeno de la medicina defensiva.

> *"OCDE también ha registrado un riesgo potencial para el 27% de los empleos como consecuencia de la IA generativa, particularmente en sectores como gerencia e Ingeniería."*

Es decir, los médicos pueden sentir temor de ser responsables del riesgo conocido, lo cual los puede llevar a priorizar tratamientos que les protejan de posibles críticas, por encima de aquellos que podrían ser más eficaces. En IA, esto sería equivalente a prohibir o restringir el uso de sistemas de IA generativa para evitar los riesgos que ya hemos discutido. Es un área gris donde resulta difícil establecer límites entre regulación, autorregulación, responsabilidad personal y responsabilidad indirecta.

Será pues necesario mantener la discusión viva en torno a estos límites y la manera de gestionarlos. El objetivo en todo caso debe ser lograr un control humano significativo. En una propuesta de control humano significativo de sistemas de IA se recogen las siguientes recomendaciones[13]: (1) el diseño de sistemas de IA debe orientarse a propiedades emergentes, esto es, a pesar de que se pueden generar principios y requerimientos para establecer el límite entre el control humano y la IA, siempre serán provisionales y el proceso de diseño debe por ende estar abierto a situaciones no anticipadas.

(2) el control humano será necesario, pero insuficiente, pues si bien logra mantener la responsabilidad moral en los humanos, no previene el que el diseño y operación tengan fallas éticas, lo cual requiere entonces no solo control sino alineación estratégica con normas y valores sociales, incluyendo derechos humanos y sostenibilidad ambiental; (3) quizá el mayor reto y la más prometedora oportunidad de control humano significativo provenga del hecho de que ninguna disciplina aislada podrá lograrlo, se requerirán ingenieros, diseñadores, científicos sociales, abogados y por supuesto grupos de interés sociales dialogando y avanzando en conjunto.

Conclusión

La discusión en torno a la regulación y el control humano significativo en el desarrollo y uso de sistemas de IA generativa es esencial. Esto implica no solo la supervisión y responsabilidad de los usuarios, sino también una alineación estratégica con valores éticos y sociales.

La colaboración interdisciplinaria y el diálogo continuo entre diferentes partes interesadas son clave para abordar de manera efectiva los desafíos éticos y sociales planteados por la IA generativa. En última instancia, se debe buscar un equilibrio entre aprovechar el potencial de esta tecnología y mitigar sus posibles consecuencias negativas en la sociedad.

* **Rafael Andrés González Rivera.** Ingeniería de Sistemas, Pontificia Universidad Javeriana con Maestría en Ciencias de la Computación de Delft University of Technology en Holanda y Doctorado en Ingeniería de Sistemas de la Universidad de Delft, Holanda. Actualmente Profesor Titular de la Facultad de Ingeniería, Pontificia Universidad Javeriana, Bogotá.

1 Gonzalez, R. A. (2021). Consideraciones y recomendaciones éticas en Inteligencia Artificial (IA). Revista ACIEM, 141, 56-60.
2 Vinge, V. (1993, diciembre 1). The coming technological singularity: How to survive in the post-human era. https://ntrs.nasa.gov/citations/19940022856
3 Kurzweil, R. (2005). The Singularity is Near: When Humans Transcend Biology. Viking.
4 Core Views on AI Safety: When, Why, What, and How. (s. f.). Anthropic. Recuperado 1 de septiembre de 2023, de https://www.anthropic.com/index/core-views-on-ai-safety
5 Hatzius, J., Briggs, J., Kodnani, D., & Pierdomenico, G. (2023, marzo 23). The Potentially Large Effects of Artificial Intelligence on Economic Growth (Briggs/Kodnani). Goldman Sachs Economic Research.
6 OECD. (2023). OECD Employment Outlook 2023: Artificial Intelligence and the Labour Market. OECD.
7 Pause Giant AI Experiments: An Open Letter. (s. f.). Future of Life Institute. Recuperado 1 de septiembre de 2023, de https://futureoflife.org/open-letter/pause-giant-ai-experiments/
8 Lemire, D. (s. f.). ChatGPT is Not a Technological Singularity. Recuperado 1 de septiembre de 2023, de https://cacm.acm.org/blogs/blog-cacm/273606-chatgpt-is-not-a-technological-singularity/fulltext
9 Abdullah, M., Madain, A., & Jararweh, Y. (2022). ChatGPT: Fundamentals, Applications and Social Impacts. 2022 Ninth International Conference on Social Networks Analysis, Management and Security (SNAMS), 1-8.
10 Krügel, S., Ostermaier, A., & Uhl, M. (2023). ChatGPT's inconsistent moral advice influences users' judgment. Scientific Reports, 13, 4569.
11 «Hypnotized» ChatGPT and Bard Create Malicious Code, Offer Bad Advice. (2023, agosto 8). Gizmodo. https://gizmodo.com/chatgpt-google-bard-hypnotized-bad-code-advice-1850718070
12 AI Index Report 2023 – Artificial Intelligence Index. (s. f.). Recuperado 4 de septiembre de 2023, de https://aiindex.stanford.edu/report/
13 Cavalcante Siebert, L., Lupetti, M. L., Aizenberg, E., Beckers, N., Zgonnikov, A., Veluwenkamp, H., Abbink, D., Giaccardi, E., Houben, G.-J., Jonker, C. M., van den Hoven, J., Forster, D., & Lagendijk, R. L. (2023). Meaningful human control: Actionable properties for AI system development. AI and Ethics, 3(1), 241-255.