UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Revista Científica

http://revistas.udistrital.edu.co/ojs/index.php/revcie/index

EDITORIAL

Reflexiones sobre las perspectivas modernas de la IA con intervención humana

Andrés G. Soto-Rangel[1]
Diego H. Peluffo-Ordóñez[2]
Hector Florez[3]

La convergencia entre la cognición humana y la inteligencia artificial (IA) se ha convertido en un punto focal de la investigación en IA. El auge de las llamadas *metodologías con intervención humana* (HITL) destaca la importancia de integrar el razonamiento humano a los modelos de aprendizaje automático (ML) para una mejor toma de decisiones, interpretabilidad y adaptabilidad. La IA ha logrado avances significativos en el ámbito de la automatización, pero aún persisten desafíos en áreas que requieren razonamiento contextual, consideraciones éticas y generalización. Para cerrar estas brechas, los enfoques HITL aprovechan la experiencia humana en el entrenamiento, la validación y el refinamiento de modelos de IA, lo que permite sistemas más robustos y responsables. Los avances recientes en inteligencia artificial generativa (GenAI) han suscitado discusiones intrigantes sobre la delegación de tareas cognitivas y sus limitaciones inherentes. En cuanto a los modelos de lenguaje grandes (LLMs), algunos investigadores afirman que, a pesar de su impresionante habilidad para imitar el lenguaje humano, su base estadística de reconocimiento de patrones llega a impedir el entendimiento, el razonamiento y la perspectiva moral genuinos. Sin embargo, el hecho más interesante es que, independientemente de que estos algoritmos presenten razonamiento verdadero o no, su desempeño es tan convincente que ahora corresponde demostrar que no lo hacen. Dadas estas impresionantes capacidades, la dependencia de estas tecnologías por parte del trabajo y el conocimiento humanos está creciendo rápidamente, lo que hace aún más urgente abordar sus problemas relacionados, *i.e.*, limitaciones, sesgos y daño potencial —causado por ignorancia o de manera deliberada.

Para lograr el objetivo de los enfoques HITL, debe mejorarse el rendimiento de la interacción entre humanos y modelos de ML, sobre la premisa de que la mente humana aún es mejor para realizar varias tareas en diferentes dominios. Los modelos recientes se han vuelto más capaces y están realizando tareas cognitivas con éxito. Uno de los aspectos clave —aún en amplio debate— es hasta qué punto podemos confiar en que estos sistemas asuman tareas que tradicionalmente han realizado los humanos, así como hasta dónde puede llegar dicha confianza. A medida que crece la confianza en los modelos y estos sobrepasan las capacidades humanas, la línea entre la inteligencia de los humanos y las máquinas se vuelve más difusa. Esto implica muchos riesgos latentes, pues existe un acceso público a herramientas que pueden utilizarse para respaldar criterios profesionales e incidir en decisiones que involucran asuntos altamente sensibles. Esto surge de la publicación de dichas herramientas sin

1. SDAS Research Group, Colombia. andres.soto@sdas-group.com
2. Université Mohammed VI Polytechnique, Marruecos. peluffo.diego@uum6p.ma
3. Universidad Distrital Francisco Jose de Caldas, Colombia. haflorezf@udistrital.edu.co

ningún tipo de marco regulatorio, si bien las mismas podrían resultar bastante beneficiosas en algunos ámbitos tradicionalmente complejos de toma de decisiones.

Muchos años atrás, los enfoques de ML demostraron ser efectivos en tareas muy específicas, en las cuales sobrepasaban a los humanos, pero su campo de aplicación parecía ser muy limitado. Con los avances en el aprendizaje profundo (DL) y sus capacidades emergentes, los LLMs se han vuelto expertos en una amplia gama de tareas, demostrando cierta forma de inteligencia generalizada. Parece haber poco tiempo para descubrir qué hacer, dada la velocidad con la cual se está desarrollando esta tecnología. En esta carta, nos centramos en los desafíos clave de trabajar con los modelos más avanzados, los cuales han provocado una disrupción sin precedentes y a una escala aún incalculable.

El ML HITL tradicional combina el poder algorítmico con la perspectiva humana, en cierto intento por garantizar que los sistemas de IA estén alineados con las expectativas y los valores humanos. Sin embargo, con modelos más nuevos y sofisticados, nos enfrentamos a la automatización de tareas cada vez más complejas que usualmente requieren habilidades humanas adquiridas tras años de educación y experiencia. En vista de ello, actualmente se explora un gran número de tareas dentro del alcance de los agentes artificiales (AAs) a los que se asignan roles de trabajo. Si bien se han diseñado algunas estrategias para incrementar la robustez de estos agentes —haciéndolos menos propensos a ataques maliciosos—, un "desafío formidable" consiste en balancear utilidad, efectividad y robustez, donde se menciona el HITL como una medida adicional. Por supuesto, aún persisten muchos desafíos en cuanto a la automatización y la intervención humana, como el sesgo cognitivo, la escalabilidad y la fatiga del usuario, los cuales deben ser abordados para garantizar que los aportes humanos mejoren efectivamente el desempeño de la IA.

Hoy en día nos enfrentamos a las capacidades crecientes de las máquinas, las cuales se están insertando en los trabajos, la educación y la vida de la población general sin lineamientos estrictos respecto a su uso y limitaciones. Si, por ejemplo, los autores comienzan a pensar que sus herramientas son superiores a ellos en ciertos aspectos como la generación de ideas y conclusiones y la escritura, ¿cuál sería el punto de realizar estas tareas por su cuenta? Esto ocurrirá naturalmente a medida que la percepción de las capacidades del sistema gane confianza entre sus usuarios, y también es algo que nosotros como autores empezamos a cuestionar respecto a la relevancia de hacer algunos trabajos manualmente —no solo los ya mencionados, sino también muchas tareas relacionadas con la investigación en general. A este respecto, Joshua S. Gans dio un ejemplo interesante de las implicaciones inmediatas al agradecer ChatGPT-o1-pro al final de su manuscrito y, al explicar cómo se utilizó la herramienta para escribir en tiempo récord, resaltó la pregunta aún abierta sobre el futuro de la investigación de la mano de la IA.

El problema que esto supone es el siguiente: *¿cuándo debería considerarse el uso de herramientas de IA que sean lo suficientemente avanzadas y requieran una mínima intervención del usuario como un enfoque HITL?* A pesar de que los *prompts,* las revisiones y el refinamiento humanos son de gran valor, el trabajo intelectual más relevante y que más tiempo requiere es a veces realizado por una máquina. De alguna manera, la integración de sistemas de IA es inevitable, y aún así es tan poderosa y particular que, a casi dos años del lanzamiento de ChatGPT, aún no se observa una dirección clara en cuanto a su uso en contenido intelectual.

En cuanto a los enfoques HITL, si empleáramos algún tipo de aprendizaje proactivo (PL), el principal problema del factor humano sería el ruido que introduce la calidad variable del etiquetado. A medida que somos superados en más tareas y en un número creciente de campos, nos queda preguntarnos si el insumo humano debería abandonarse en favor de mejores resultados de IA. De hecho, podría ser que la enseñanza a la máquina (MT) se convierta en el siguiente foco de interés creciente, dadas la complejidad en aumento de las tareas y la necesidad de validación humana. A este respecto, se ha propuesto un marco interesante llamado *aprendizaje profundo por refuerzo basado en humanos como mentores de la IA* (HAIM-DRL) para aplicaciones de conducción autónoma, permitiendo que los humanos asuman el control en situaciones peligrosas y le enseñen al modelo cómo reaccionar. Siempre y cuando los humanos permanezcamos en el ciclo, concentrarnos en nuestra eficacia como profesores podría ayudar a enfrentar algunas de las limitaciones de enfoques anteriores. A modo de complemento, se ha planteado el aprendizaje curricular (CL) como un enfoque HITL más estrechamente relacionado con el aprendizaje activo (AL), si bien también podría verse como una extensión de la MT, pues busca reducir la carga sobre el profesor para optimizar todo el proceso, alcanzando una mayor efectividad mediante la eliminación de ruido.

Naturalmente, no existe un consenso universal sobre la clasificación de enfoques o metodologías —aunque debería. Es posible que conceptos como el *aprendizaje recíproco humano-máquina* (RHML) y el *aprendizaje humano-máquina* (HML) ganen relevancia en los próximos años, en un intento por favorecer las habilidades humanas. A medida que la interacción humano-computador continúa desarrollándose, los individuos más competentes dentro del ciclo para la mejora de las máquinas serán aquellos cuyas capacidades cognitivas sean aumentadas por la tecnología.

Otros puntos de vista más filosóficos sobre las metas y el mejoramiento de los sistemas plantean preocupaciones sobre la búsqueda de la inteligencia artificial general (AGI). Todo esto, desde una perspectiva antropocéntrica, proponiendo diferentes objetivos como el rol de asistente de los humanos por sobre el reemplazo o el desplazamiento en favor de la AGI "similar a la humana" como el estándar de referencia, siempre y cuando el sistema inteligente pueda "colaborar" como un humano. En este sentido, la discusión principal gira en torno al mejoramiento de la alfabetización en IA como una base fundamental para trabajar con estos sistemas de manera efectiva. Lo que podemos esperar en los años venideros es una demanda creciente de profesores humano-máquina, así como la demanda de cualquier trabajo o habilidad relacionada con la supervisión, auditoría o entrenamiento de sistemas para su mejora. Existirá la posibilidad de que esto suceda siempre que haya áreas en los que la IA presente dificultades. Con este objetivo en mente, la investigación debería explorar enfoques híbridos que integren el razonamiento simbólico con el DL, fomentando el surgimiento de sistemas de IA capaces de adaptarse a escenarios novedosos a la vez que se mantiene la supervisión humana. Adicionalmente, mejorar las interfaces humano-IA por medio de una mejor visualización y el procesamiento del lenguaje natural (NLP) mejorará la accesibilidad y la usabilidad.

La cuestión principal que es relevante para cualquier entidad de toma de decisiones corresponde a las consideraciones éticas. Estos temas, como las entidades artificiales, se han tratado tradicionalmente en el ámbito de la ciencia ficción, y es por ello que las

consideraciones formales son tan nuevas y abiertas. La discusión ya está en curso, con diferentes enfoques que parecen abordar el problema como cierto tipo de característica emergente de las soluciones implementadas que debe ser controlada, en vez de centrarse en el problema humano transversal. Los enfoques más reflexivos consideran esto como el bloque fundamental, como es el caso de la ética por diseño (EbD). Dado el profundo impacto de estos sistemas —con el potencial para incidir en decisiones delicadas— es esencial contar con una mayor exploración a nivel de desarrollo. Con la creciente disponibilidad de librerías de ML y herramientas de IA listas para usar, un desarrollo de sistemas inteligentes centrado en la ética (EDD) podría ser beneficioso para complementar la EbD desde la perspectiva de los desarrolladores. Esto podría tener un impacto significativo, dado el rol creciente de los LLMs en la programación y el potencial de las máquinas para enseñar a otras máquinas. Es crucial establecer mejores prácticas a seguir para los modelos de IA en el desarrollo de *software* y futuros modelos. Además, aún hay una gran cantidad de trabajo por hacer en lo que respecta al refinamiento de los marcos éticos para la IA, pues la simplificación excesiva implica múltiples tipos de riesgos.

A pesar del argumento de que los humanos somos éticamente inmaduros, esta podría ser la última cosa que querríamos delegar, y tener la última palabra bien podría ser uno de nuestros intereses principales. No obstante, no podemos negar que las capacidades analíticas están expandiéndose gracias a la IA. Así, para al menos apoyar a los expertos en ética en la evaluación de los sistemas de valores de los modelos de IA, debe considerarse el trabajo de los agentes morales artificiales (AMAs), especialmente en relación con el aumento de las capacidades de evaluación humanas.

Otra gran preocupación es que la tecnología militar es secreta por naturaleza, y que con seguridad no estamos informados sobre los últimos avances, los cuales inherentemente representan graves riesgos éticos y existenciales. Adicionalmente, en la actualidad el concepto de *transparencia algorítmica* es más una propuesta que una realidad en los gobiernos, como es el caso de los sistemas de votación o administrativos propietarios, lo cual podría extrapolarse a los sistemas internos y las políticas de la IA si no hay presión pública para un cambio. Hoy en día, las arquitecturas de DL son lo suficientemente opacas para cubrir también los algoritmos y conjuntos de datos subyacentes, por lo que la manera más efectiva de promover la transparencia es fomentar la apertura en relación con estas tecnologías y su funcionamiento interno.