La inteligencia artificial generativa y la evaluación: ¿Qué pasará con los exámenes?

Generative artificial intelligence and assessment: What will happen to exams?

"Con mucho, el mayor peligro de la inteligencia artificial es que las personas concluyen demasiado pronto que la entienden".

Eliezer Yudkowsky

Recientemente, Daisy Christodoulou, una académica de Gran Bretaña, preguntó en su blog: *"Si estamos haciendo exámenes que un robot puede contestar bien, ¿qué dice eso de nuestras evaluaciones?"*[1]. Las sofisticadas herramientas de la inteligencia artificial generativa (IAGen), como ChatGPT, continúan creando muchas inquietudes en docentes e instituciones, ya que introducen aspectos éticos, técnicos y académicos que siguen sin tener una respuesta definida que satisfaga a todos los actores del proceso educativo[2,3]. Una de las áreas más controversiales es la que tiene que ver con los retos de evaluación del y para el aprendizaje que surgen con la IAGen. El proceso de evaluación del y para el aprendizaje, aún antes de la pandemia y de la revolución actual de inteligencia artificial, distaba de ser el ideal y, por diversas circunstancias, no era aplicado satisfactoriamente en los escenarios escolares y clínicos por el profesorado y las instituciones de salud y educativas. En esta Editorial comentaremos brevemente las implicaciones positivas y negativas de la IAGen para la evaluación, así como algunas posibles estrategias para enfrentarlas.

Tradicionalmente la evaluación del y para el aprendizaje es una tarea compleja, infravalorada, que con frecuencia no está alineada con el currículo, y que continúa teniendo un gran énfasis en exámenes y evaluaciones sumativas en contextos artificiales y poco auténticos. En tiempos recientes la pandemia causó una severa crisis en el mundo educativo en todos los niveles, al obligar a realizar evaluaciones a distancia con exámenes en línea en casa. Como consecuencia hubo cambios importantes en la educación en línea y en las formas de realizar las evaluaciones formativas y sumativas. Desafortunadamente de forma gradual parece ser que estamos regresando a la actitud prepandemia de minusvaloración de la evaluación, subestimando su rol en el proceso educativo.

Este es un artículo Open Access bajo la licencia CC BY-NC-ND (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Aunado al proceso postpandemia de regreso a nuestras zonas de comodidad, aparece un nuevo jugador en educación, la IAGen, que está generando cambios potencialmente disruptivos en la educación superior[4,5]. Reflexionemos sobre los tipos de instrumentos de evaluación que pueden ser más afectados por la IAGen. Actualmente los tipos de evaluación que se encuentran en la categoría de "evaluaciones escritas", como el ensayo, las preguntas de opción múltiple, de respuesta corta, de compleción, y las disertaciones, son los que tienen mayor riesgo de ser influenciados por el uso de la IAGen por docentes o estudiantes. Sin embargo, existen otras formas de evaluación que podemos utilizar sin preocuparnos demasiado porque sean afectadas a corto plazo por la IAGen, como portafolios, reportes de pares, observación directa, listas de cotejo, registros de desempeño, exámenes clínicos objetivos estructurados, exámenes orales, entre otras. Es importante tomar en cuenta que el ritmo inaudito con el que están avanzando estas herramientas tecnológicas nos hace pensar que en un futuro cercano la IAGen modificará profundamente el escenario de evaluación, sobre todo cuando la inteligencia artificial multimodal (texto, audio y video) sea de uso cotidiano por la comunidad. Tal vez lo importante a considerar es que, aún con los instrumentos que pueden ser más afectados por la IAGen, podemos trabajar en equipo en un diálogo constante con el estudiantado, para aprovechar su potencial y mejorar el aprendizaje[6].

Es un hecho que las herramientas de IAGen como ChatGPT tienen un enorme potencial de generar una disrupción real en evaluación educativa[4,5]. La comunidad docente y administradores de la educación estamos en un estado que pudiera considerarse de "pánico moral", por la irrupción de IAGen en las evaluaciones tradicionales, como exámenes y ensayos. Hay quien dice que "¡se acabaron los exámenes y las tareas!", ya que las nuevas herramientas son capaces de responder mejor algunos instrumentos de evaluación que los seres humanos, como ensayos y algunos exámenes de certificación profesional, poniendo a profesores y universidades en una posición difícil: ¿cómo estar seguro de que el resultado de la evaluación es válido y refleja realmente lo que sabe y sabe hacer el educando? Por ejemplo, ChatGPT 4 es capaz de obtener puntuaciones similares o superiores en exámenes sumativos de alto impacto dirigidos a médicos, abogados, entre otros, lo que nos hace preguntarnos ¿cuál es el futuro de este tipo de exámenes?[7].

Las respuestas iniciales a IAGen en evaluación que ocurren en el profesorado generalmente caen en estas tres categorías[8]:

- **Vetarla.** Prohibir el uso de IAGen en universidades no funciona, aunque estemos tentados a ello. Cuando se dice a los estudiantes que no usen una herramienta tecnológica porque serán castigados, puede ser un incentivo para que la usen más. El rechazo es motivado por miedo y desconocimiento de instituciones y docentes, basados en la premisa cuestionable de que los estudiantes lo usarán para hacer trampa y que no se puede confiar en ellos. Usar IA para vigilar exámenes en línea es invasivo, caro y estresante, por lo que hay que reflexionar cuidadosamente sobre su uso con este fin.
- **Evitarla.** Meter la cabeza en la arena, como el avestruz, para darle la vuelta al problema y regresar a evaluaciones cara a cara, con exámenes escritos en clase, con vigilancia humana. En otras palabras, regresar al pasado prepandemia y pretecnología. Creo que esta actitud no debe ser considerada, de otra manera estaríamos dejando de usar una innovación que puede ayudarnos a mejorar el aprendizaje de los estudiantes. Hay que hacer notar que se están utilizando las mismas herramientas de IA para detectar contenido creado por IA, pero hasta ahora la evidencia es bastante contundente: ¡ninguna sirve!, existen demasiados falsos positivos y negativos. De forma empírica y teórica se ha demostrado que los detectores para textos de IA no son confiables en escenarios realistas[9]. Al basarse en la premisa equivocada de que hay que hacer lo posible para que los estudiantes no hagan trampa, en lugar de modificar nuestros paradigmas de evaluación, terminamos dando vueltas en círculos, sin avanzar en la forma como nos relacionamos con el estudiantado en estos temas.
- **Adoptarla.** Probablemente esta debe ser la principal estrategia a utilizar. Abrazar la IAGen como

herramienta y explorarla para desarrollar las competencias digitales de los estudiantes y sus habilidades analíticas, al mismo tiempo que se genera confianza y se dialoga con ellos como compañeros de la aventura del aprendizaje. De esta forma se genera un equipo colaborativo profesorado-estudiantado que se involucra en un proceso continuo de aprendizaje sobre el uso efectivo, ético y sensato de las herramientas, al mismo tiempo que se mantiene actualizado en el tema. Por supuesto que deben tomarse en cuenta las limitaciones de la IAGen, como son las alucinaciones y la confianza con la que afirma datos que pueden ser falsos e incorrectos[2,3].

Las siguientes sugerencias pueden ser de utilidad para las y los docentes de hospitales y universidades:

- ¡No es necesario asustarse! Mantengan la calma y abracen estas herramientas, familiarícense con ellas, tengan claridad de su gran potencial e importantes limitaciones.
- Chequen preguntas y exámenes con IAGen y reflexionen sobre lo que quieren hacer. No pierdan de vista el objetivo principal de la evaluación del y para el aprendizaje, las preguntas principales son: ¿qué evaluar, para qué evaluar y cómo queremos/podemos evaluar? El contexto es fundamental, así como el currículo del curso a evaluar.
- Amplíen su abanico de instrumentos: exámenes orales, proyectos, observación de discusiones, elaboración de diagramas. Es fundamental integrar la evaluación formativa frecuente en el proceso educativo, y disminuir el énfasis en evaluaciones sumativas de alto impacto. La IAGen puede ser excelente para dar retroalimentación frecuente y personalizada.
- Platiquen con los estudiantes sobre su uso, lleguen a acuerdos de consenso, de preferencia siguiendo las políticas y normatividad institucionales. Si no existen políticas locales, es importante sugerirles a las autoridades institucionales que creen grupos de trabajo sobre esta temática.
- Usen la herramienta para promover creatividad y pensamiento crítico, por ejemplo, pueden pedir a los estudiantes que critiquen las respuestas de ChatGPT y que fundamenten sus argumentos.
- Pidan a los estudiantes que siempre chequen las fuentes y referencias. No confíen ciegamente en la IAGen.
- Piensen sobre el futuro inmediato y a mediano plazo: ¿cómo podemos empoderarnos los docentes, mientras nosotros y los estudiantes usamos estas herramientas?, ¿cómo integrarlas en las tareas cotidianas, aprovechando sus bondades con conciencia de sus limitaciones?

Reflexionemos intensamente sobre estas herramientas, y comencemos a realizar trabajos de investigación educativa sobre IAGen y educación en profesiones de la salud. Por el momento la cantidad y calidad de trabajos sobre el tema están en sus albores, en su mayoría son "*preprints*" que no han pasado el arbitraje por pares, por lo que se requiere que con rigor académico documentemos investigaciones y experiencias locales sobre el tema.

Regresando al presente número de la revista, tenemos artículos originales sobre los siguientes temas: ambientes educativos y éxito académico; educación de geriatría en México; examen nacional del internado en Perú; flexibilidad cognitiva y rendimiento académico en estudiantes de medicina; simulación virtual en enfermería; curso virtual de lectura crítica en la pandemia; inteligencia emocional en médicos residentes; exposición oral en clases de docentes y estudiantes. Además, contamos con un artículo de revisión sobre el desarrollo y validación de las actividades profesionales a confiar (EPAs en inglés), así como un ensayo sobre la formación médica y un escenario político contingente, tema especialmente relevante en la época actual.

El advenimiento de la IAGen y sus múltiples herramientas, de las que ChatGPT es una de las más conocidas, están propiciando un escenario educativo global proclive a la innovación, por lo que es necesario explorarlas y combinarlas con las que ya utilizamos, en beneficio de los estudiantes.

Melchor Sánchez Mendiola
Editor en Jefe
Facultad de Medicina, UNAM

REFERENCIAS

1. Christodoulou D. If we are setting assessments that a robot can complete, what does that say about our assessments? The No More Marking Blog. Feb. 5, 2023. https://blog.nomoremarking.com/if-we-are-setting-assessments-that-a-robot-can-complete-what-does-that-say-about-our-assessments-cbc1871f502
2. Pearce J, Chiavaroli N. Rethinking assessment in response to generative artificial intelligence. Med Educ. 2023; 57(10):889-891. https://doi.org/10.1111/medu.15092
3. Swiecki Z, Khosravi H, Chen G, Martinez-Maldonado R, Lodge JM, Milligan S, Gašević D. Assessment in the age of artificial intelligence. Computers and Education: Artificial Intelligence. 2022;3:100075. https://doi.org/10.1016/j.caeai.2022.100075
4. Susnjak T. ChatGPT: The End of Online Exam Integrity? ArXiv, 2022, abs/2212.09292. https://doi.org/10.48550/arXiv.2212.09292
5. Farazouli A, Cerratto-Pargman T, Bolander-Laksov K, McGrath C. "Hello GPT! Goodbye home examination? An exploratory study of AI chatbots impact on university teachers' assessment practices", Assessment & Evaluation in Higher Education. 2023:1-13. https://doi.org/10.1080/02602938.2023.2241676
6. Karolinska Institutet. ChatGPT and assessment. 2023. https://staff.ki.se/chatgpt-and-assessment
7. Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of GPT-4 on Medical Challenge Problems. ArXiv 2023. https://arxiv.org/abs/2303.13375v2 https://doi.org/10.48550/arXiv.2303.13375
8. Lindsay K. ChatGPT and the Future of University Assessment. Kate Lindsay Blogs. Jan 16, 2023. https://katelindsayblogs.com/2023/01/16/chatgpt-and-the-future-of-university-assessment/
9. Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, Soheil Feizi. Can AI-Generated Text be Reliably Detected? ArXiv 2023. https://arxiv.org/abs/2303.11156v2 https://doi.org/10.48550/arXiv.2303.11156