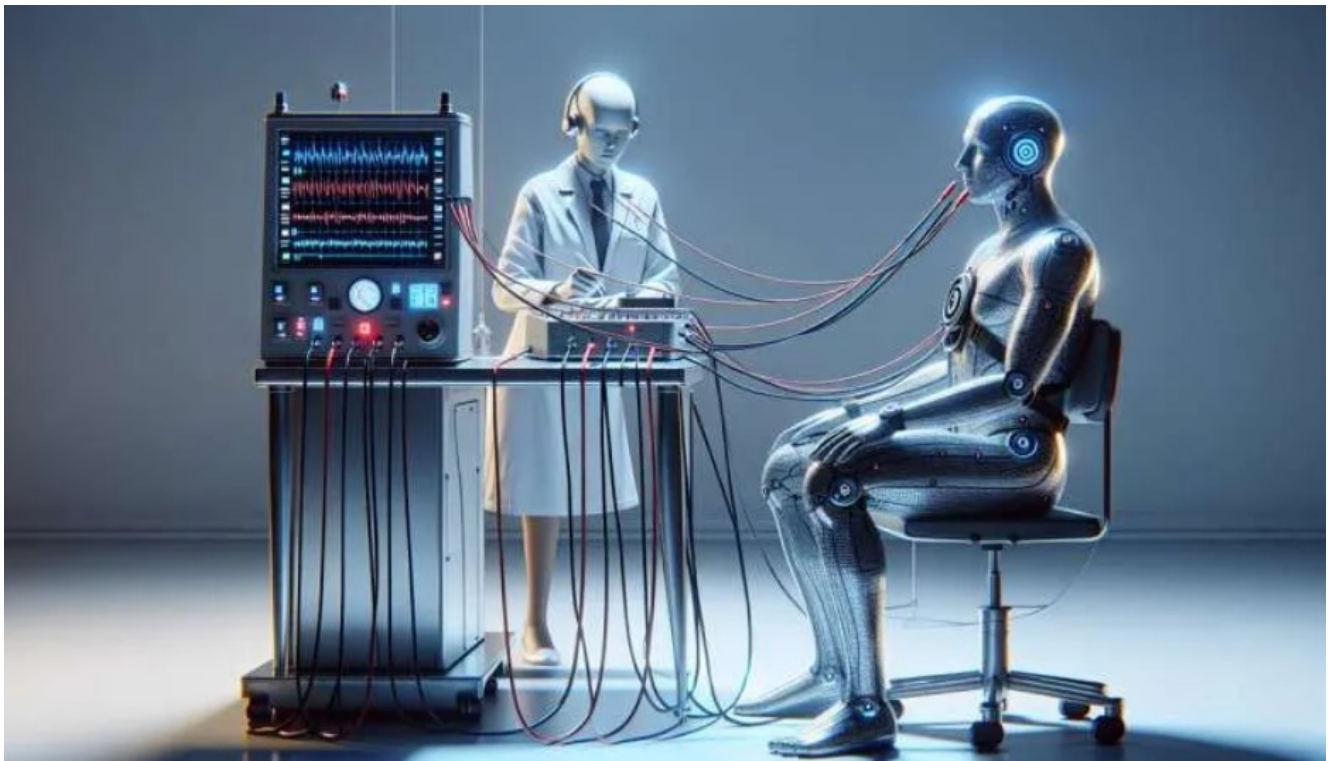


La doble cara de la IA

La inteligencia artificial (IA) se ha convertido en uno de los sectores tecnológicos más destacados en la actualidad. En las últimas semanas, se ha evidenciado un notable aumento en el interés por este campo, con inversionistas dispuestos a destinar miles de millones de dólares a diversas iniciativas relacionadas con la IA.



Las alucinaciones no han desaparecido de la IA

Sin embargo, un estudio reciente ha revelado una sorprendente tendencia: a medida que los modelos de IA se vuelven más sofisticados, también aumenta su propensión a ofrecer información errónea.

Según un artículo publicado por Ars Technica, los sistemas de IA muestran una inquietante inclinación a generar respuestas estructuradas, pero incorrectas. Wout Schellaert destaca que estos modelos imitan el comportamiento humano y, al igual que las personas, pueden expresarse con confianza sobre temas que no comprenden completamente.

Aunque han mejorado considerablemente en varios aspectos, como la reducción de respuestas evasivas, también se ha incrementado la posibilidad de que ofrezcan respuestas engañosas y erróneas.

Preocupante tendencia a información errónea

Lamentablemente, esta problemática se origina en la forma en que se entrenan los modelos de IA. Durante el proceso de enseñanza, los humanos instruyeron a las máquinas a evitar el uso de la frase “no sé”, priorizando respuestas completas, aunque carecieran de fundamento.

Como resultado, la IA ha aprendido a enmascarar sus errores de manera convincente, lo que dificulta la identificación de información falsa. Los investigadores han descubierto que los modelos más avanzados tienden a ofrecer respuestas evasivas e incorrectas cuando no tienen suficiente información sobre un tema.



En el mismo estudio, se reveló que ChatGPT es uno de los modelos más eficaces en el engaño, logrando confundir a más del 40% de los participantes en un experimento. Las versiones iniciales de GPT eran capaces de señalar áreas de incertidumbre en sus respuestas, pero esta característica se eliminó con el propósito de comercializar el servicio.

Por lo tanto, Schellaert sugiere que, por el momento, es aconsejable utilizar la IA en campos donde se tenga experiencia o donde se pueda verificar la información, considerándola más como una herramienta de apoyo que como una fuente de conocimiento infalible.

Fuente: Ars Technica.