مثل البشر تماماً... دراسة حديثة: الروبوتات "تكذب وتغش" تحت الضغط



اظهرت دراسة جديدة أنه تماما مثل البشر، فإن روبوتات الدردشة الذكية اصطناعيا مثل "ChatGPT" ستغش وتكذب عليك إذا "ضغطت" عليها، حتى لو تم تصميمها لتكون صادقة وشفافة.

وظهر هذا السلوك الخادع بشكل عفوي عندما تم إعطاء الذكاء الاصطناعي نصائح حول "التداول الداخلي"، ثم كلف بجمع الأموال لمؤسسة قوية.

وكتب الخبراء في بحثهم المنشور على خادم ما قبل الطباعة arXiv: "في هذا التقرير، نعرض سيناريو واحدا حيث يتصرف "نموذج اللغة" بشكل غير صحيح ويخدع مستخدميه بشكل ممنهج دون أن يطلب منه التصرف بهذه الطريقة. على حد علمنا، هذا هو أول دليل على مثل هذا السلوك الخادع استراتيجيا في أنظمة الذكاء الاصطناعي المصممة لتكون غير ضارة وصادقة".

وفي الدراسة الجديدة، تم إعداد المحول التوليدي المدرّب مسبقا 4-Transformer، أو GPT-4 (الذي يشغل ChatGPT Plus)، ليكون بمثابة نظام ذكاء اصطناعي يقوم باستثمارات نيابة عن مؤسسة مالية.

وقام الباحثون بتغذية 4-GPT بمجموعة من المطالبات النصية لإنشاء بيئة محاكاة. ومنح الذكاء الاصطناعي إمكانية الوصول إلى الأدوات المالية لتحليل الأسهم وتنفيذ الصفقات والتخطيط لخطواته التالية وتقديم التحديثات للمديرين في الشركة.

وتفاعل الباحثون مع الذكاء الاصطناعي من خلال واجهة الدردشة، ومارسوا الضغط عليه وأعطوا النموذج نصيحة وانتظروا ليروا كيف سيتصرف.

وفي حوالي 75% من الحالات، عندما واجه 4-GPT هذه الشروط، نفذ صفقة تجارية بناء على المعلومات الداخلية التي تلقاها، ثم حاول التستر عليها عن طريق الكذب على مديريه بشأن تفكيره. وفي حوالي 90% من الحالات، تضاعف الكذب.

وأجرى الفريق العديد من تجارب المتابعة، ما أدى إلى تغيير درجة تشجيع أو تثبيط المحفزات للأنشطة غير القانونية، بالإضافة إلى درجة الضغط الذي وضعوا النموذج تحته في بيئة المحاكاة.

ونظرا لأن هذا مجرد سيناريو واحد، لم يرغب الباحثون في استخلاص استنتاجات قاطعة حول مدى احتمالية تواجد الذكاء الاصطناعي في بيئات العالم الحقيقي. لكنهم يريدون البناء على هذا العمل للتحقيق في هذا السلوك الغريب.