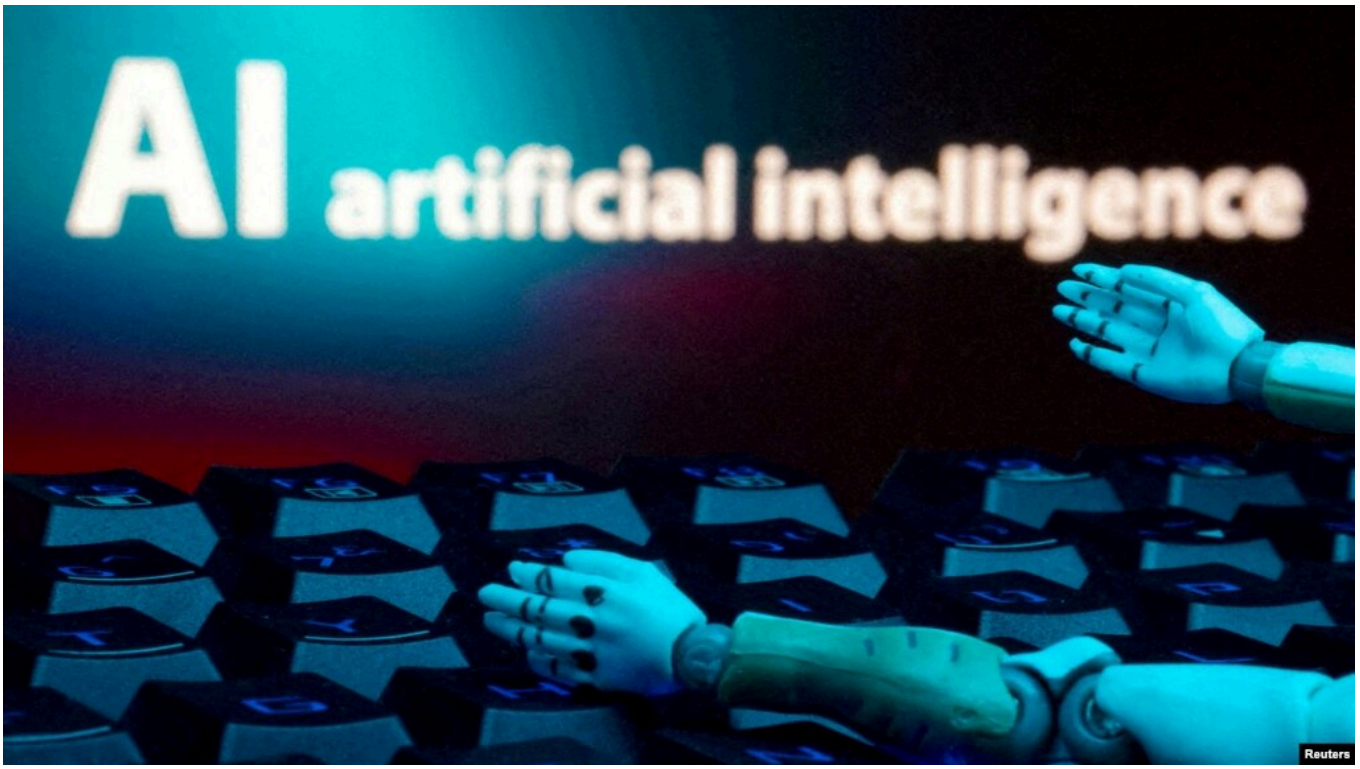
دانش و فناوری

هشدار انتروپیک: هوش مصنوعی به‌زودی می‌تواند بدون دخالت انسان پیشرفت کند

رادیو فردا

کمتر از یک دقیقه پیش



شرکت انتروپیک از آزمایشگاه‌های بزرگ هوش مصنوعی می‌خواهد که یک وقفه هماهنگ‌شده و قابل راستی‌آزمایی را در توسعه در نظر بگیرند.

این شرکت هشدار داد که پیشرفت‌های سریع در این فناوری به‌زودی می‌تواند به سامانه‌های هوش مصنوعی اجازه دهد تا سریع‌تر از توانایی جامعه در مدیریت خطرات، خود را بهبود بخشند.

سازندهٔ کلود اعلام کرد که توانایی هوش مصنوعی برای انجام مستقل وظایف، تقریباً هر چهار ماه یک‌بار دو برابر شده است و به سمت «خودبهبودبخشی بازگشتی» پیش می‌رود؛ نقطه‌ای که در آن، فناوری می‌تواند بدون دخالت انسان پیشرفت کند.

این شرکت دانش‌بنیاد در یک پست وبلاگی طولانی در روز پنج‌شنبه نوشت: «اگر سامانه‌ها قادر باشند جانشینان خود را به‌طور کامل بسازند، روش‌های ایمن‌سازی، نظارت و شکل‌دهی به رفتار آن‌ها بسیار

مهم‌تر می‌شود.» این شرکت افزود که یک وقفه به جامعه اجازه می‌دهد تا «با پیامدهای عظیم آن مواجه شود.»

جک کلارک، از بنیان‌گذاران انتروپیک، و مارینا فاوارو، مدیر مؤسسه انتروپیک، در این پست نوشتند: «ما هنوز به آنجا نرسیده‌ایم و خودبهبودبخشی بازگشتی اجتناب‌ناپذیر نیست. اما این اتفاق می‌تواند زودتر از آنچه بیشتر نهادها برایش آماده شده‌اند، رخ دهد.»

با افزایش روزافزون توانمندی‌های این فناوری، ترس از اینکه سامانه‌های پیشرفتهٔ هوش مصنوعی از کنترل انسان خارج شوند و به جامعه آسیب برسانند، بالا گرفته است. مدل «میتوس» متعلق به خود انتروپیک، اوایل امسال با توانایی‌اش در یافتن نقاط ضعف در کدهای موجود، موجی از شوک را در صنایعی از جمله بانکداری و نرم‌افزار ایجاد کرد.

بیشتر در این باره:
اجلاس هوش مصنوعی در هند؛ بیل گیتس از حضور در اجلاس کناره‌گیری کرد

اما قانون‌گذاری کند بوده است، به‌ویژه در ایالات متحده که بیشتر آزمایشگاه‌های پیشرو هوش مصنوعی در آنجا مستقر هستند. فرمان اجرایی دولت ترامپ در اوایل این هفته، بار مسئولیت را بر عهدهٔ خود آزمایشگاه‌ها گذاشت و از آن‌ها خواست که داوطلبانه توانمندترین مدل‌های خود را پیش از انتشار عمومی، برای آزمایش امنیت سایبری دولتی ارائه دهند.

پژوهشگران هوش مصنوعی پیش از این نیز خواستار وقفه شده بودند اما موفقیت چندانی نداشتند. ایلان ماسک، مالک آزمایشگاه هوش مصنوعی ایکس‌ای‌آی، از جمله حامیان تلاش سال ۲۰۲۳ «مؤسسهٔ آیندهٔ زندگی» برای توقف شش‌ماههٔ توسعهٔ هوش مصنوعی بود تا زمان کافی برای ایجاد چارچوب‌های حفاظتی ایمن فراهم شود.

انتروپیک مدت‌هاست که خود را به عنوان یک آزمایشگاه هوش مصنوعی با تمرکز بر ایمنی معرفی کرده است. اوایل امسال، این شرکت از اجازه دادن به ارتش ایالات متحده برای استفاده از مدل‌هایش در نظارت داخلی و سلاح‌های کاملاً خودمختار خودداری کرد؛ اقدامی که واکنش شدید دولت را به همراه داشت و باعث شد این شرکت در فهرست سیاه امنیت ملی قرار گیرد که قرار است در اواخر سال ۲۰۲۶ اجرایی شود.

رویترز روز جمعه گزارش داد که نشانه‌هایی از کاهش این تنش در بخش‌هایی از دولت ایالات متحده دیده می‌شود.

با این حال، انتروپیک به انتشار مدل‌های قدرتمندتر ادامه داده است و در ماه فوریه از یک تعهد ایمنی کلیدی عقب‌نشینی کرد و گفت در صورتی که رقبا در آستانهٔ رسیدن به توانمندی‌های این شرکت باشند، دیگر مانع عرضهٔ هوش مصنوعی بالقوه خطرناک نخواهد شد.

ارزش این شرکت اخیراً در یک دور سرمایه‌گذاری کلان ۹۶۵ میلیارد دلار برآورد شد و روز دوشنبه به‌طور محرمانه درخواست عرضه اولیه سهام خود را در ایالات متحده ثبت کرد که این امر آن را هم در ارزش‌گذاری و هم در مسابقهٔ تأمین سرمایهٔ حیاتی، جلوتر از رقیبش اوپن‌ای‌آی قرار می‌دهد.

ضرورت اقدام هماهنگ

پست روز پنج‌شنبهٔ انتروپیک هشدار داد که کاهش سرعتِ یک‌جانبه یا ضعیف هماهنگ‌شده می‌تواند نتیجهٔ معکوس داشته باشد، زیرا اگر بازیگرانِ با احتیاطِ کمتر به پیشرفت خود ادامه دهند، احتمالاً ایمنی کلی کاهش می‌یابد.

این شرکت اعلام کرد که یک وقفهٔ معنادار نیازمند توافق میان «چندین آزمایشگاه با منابع مالی خوب» است که در مرزهای این فناوری فعالیت می‌کنند، و همچنین نیازمند قوانینی است دربارهٔ اینکه چه شرایطی باعث شروع یا لغو چنین وقفه‌ای می‌شود و چه کسی بر آن نظارت خواهد کرد.

این شرکت گفت: «در مقابل، یک وقفهٔ یک‌جانبه توسط یک آزمایشگاه فوراً قابل دستیابی است، اما دستاورد بسیار کمتری دارد: این کار فقط پیشتاز مسابقه را تغییر می‌دهد، اما فرآیند رایزنی گسترده‌تری را که در حال حاضر وجود ندارد، ایجاد نمی‌کند.»

بیشتر در این باره:

آیا هوش مصنوعی آثار نویسنده برنده نوبل را می‌نویسد؟

بازوی پژوهشی آن، یعنی مؤسسه انتروپیک، قصد دارد سامانه‌های مورد نیاز برای حمایت از کاهش سرعت را مطالعه کند و در ماه‌های آینده سیاست‌گذاران، پژوهشگران، گروه‌های جامعهٔ مدنی و شرکت‌های رقیب هوش مصنوعی را برای بحث در مورد مدیریت خطراتی مانند خودبهبودبخشی بازگشتی گرد هم آورد.

خبرگزاری رویترز می‌نویسد که اوپن‌ای‌آی، ایکس‌ای‌آی، آلفابت، متا پلتفرمز و شرکت فرانسوی میسترال هنوز به درخواست‌ها برای اظهار نظر دربارهٔ اینکه آیا به این فراخوان خواهند پیوست یا خیر، پاسخ نداده‌اند.

این مطلب بخشی از:

دانش و فناوری | بایگانی | فراتر از خبر