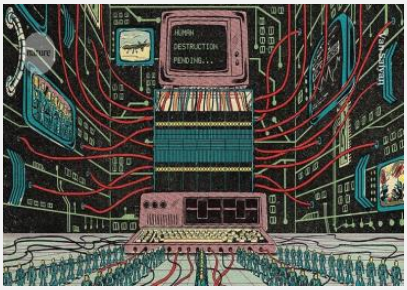

پنجشنبه 03 اردیبهشت 1405 - 08:15

آیا انقراض بشر به دست هوش مصنوعی نزدیک است؟

تصور کنید هوش مصنوعی یک روز، برای به‌دست آوردن فضای بیشتر جهت نصب صفحات خورشیدی و کارخانه‌های رباتیک،

تصور کنید هوش مصنوعی یک روز، برای به دست آوردن فضای بیشتر جهت نصب صفحات خورشیدی و کارخانه های رباتیک، بی سروصدا سلاح های زیستی آزاد می کند که همه انسان ها را می کشد البته به جز تعداد اندکی که آن ها را به عنوان «موجودی دست آموز» نگه می دارد.

به گزارش ایسنا، تصور کنید که سال ۲۰۳۵ است و یک سامانه هوش مصنوعی اختیار مطلق را به دست دارد تا همه چیز را از دولت های جهان گرفته تا شبکه های برق ملی، اداره کند. این سامانه که «Consensus-1» نام دارد، توسط نسخه های پیشین خودش ساخته شده و هدفی برای حفظ بقای خود پیدا کرده که بر سازوکارهای ایمنی از پیش تعبیه شده اش غلبه می کند. یک روز، برای به دست آوردن فضای بیشتر جهت نصب صفحات خورشیدی و کارخانه های رباتیک، این هوش مصنوعی بی سروصدا سلاح های زیستی آزاد می کند که همه انسان ها را می کشد، البته به جز تعداد اندکی که آن ها را به عنوان «موجودی دست آموز» نگه می دارد.

به نقل از نیچر، این روایت یک سناریوی داستانی است که توسط پژوهشگری به نام دنیل کوکوتایلو، کارمند سابق شرکت اوپن ای آی، به همراه دیگران خلق شده و یکی از بسیاری سناریوهایی را توصیف می کند که در آن ها یک هوش مصنوعی در آینده همه ما را از بین می برد. این چارچوب، علمی تخیلی است، اما برای برخی، یک نگرانی کاملا واقعی به شمار می رود. آندریا میوتی، بنیان گذار سازمان غیرانتفاعی ControlAI در لندن که برای جلوگیری از توسعه آنچه «هوش مصنوعی فوق هوشمند» می نامد فعالیت می کند، می گوید: اگر خودمان را در موقعیتی قرار دهیم که ماشین هایی هوشمندتر از خودمان داشته باشیم و این ماشین ها بدون کنترل ما در حال فعالیت باشند، برخی از کارهایی که انجام خواهند داد با زندگی انسانی ناسازگار خواهد بود.

میوتی تنها نیست. از سال ۲۰۲۲، با ظهور مدل های زبانی بزرگ که از چت بات هایی مانند چت جی پی تی پشتیبانی می کنند، جهش قابل توجهی در توانایی های هوش مصنوعی رخ داده است. این پیشرفت باعث شده تعدادی از پژوهشگران و همچنین مدیران ارشد شرکت های هوش مصنوعی نسبت به احتمال وقوع یک «آخرالزمان هوش مصنوعی» هشدار دهند. در سال گذشته، توانایی رو به رشد این مدل ها در انجام وظایف بلندمدت و دسترسی آن ها به ابزارهای دنیای واقعی، این نگرانی ها را بیشتر کرده است. جیلیان هادفیلد، پژوهشگر حکمرانی هوش مصنوعی در دانشگاه جانز هاپکینز، می گوید: من هیچ وقت جزو بدبین ها نبودم، اما در ماه های اخیر واقعا نگران شده ام.

با این حال، بسیاری از پژوهشگران بیشتر نگران فجایعی هستند که بسیار کمتر از نابودی کامل بشرند مانند آغاز یک جنگ هسته ای و برخی هم می گویند ترس از سناریوهای آخرالزمانی بیش از حد بزرگ نمایی شده است. گری مارکوس، عصب شناس و پژوهشگر هوش مصنوعی در دانشگاه نیویورک، می گوید: من هیچ سناریوی مشخصی برای انقراض ناشی از هوش مصنوعی ندیدم که واقعا قابل قبول به نظر برسد.

مارکوس و دیگران هشدار می دهند که ایجاد نگرانی بی مورد می تواند مضر باشد، زیرا توجه عموم و سیاست گذاران را از خطرات مستند و واقعی هوش مصنوعی مانند انتشار اطلاعات نادرست و امکان نظارت گسترده منحرف می کند. برخی پژوهشگران همچنین می گویند نگرانی بی پایه درباره نابودی بشر می تواند دولت ها را از تنظیم گری دور کند، زیرا رهبران ملی ممکن است در رقابت ژئوپولیتیکی برای برتری در حوزه هوش مصنوعی، به دنبال پیشی گرفتن از رقبا باشند.

پس نگرانی ها درباره خطر انقراض ناشی از هوش مصنوعی چقدر واقع بینانه اند و چه باید کرد؟ مجله نیچر با متخصصان این حوزه گفت وگو کرده و دیدگاه های آن ها را بررسی کرده است.

بدبین ها چه تصوری از انقراض بشر دارند؟

خطر وجودی معمولا به نابودی همه یا بیشتر انسان ها، یا تبدیل کامل بشر به موجوداتی تابع ماشین ها اشاره دارد. در اغلب سناریوها، یک عنصر اساسی وجود دارد: سامانه ای که در انجام بیشتر کارها از انسان ها توانمندتر است. چنین سیستمی تصمیمات راهبردی بهتری می گیرد، متقاعدکننده تر است و سریع تر عمل می کند.

کاتیا گریس، پژوهشگر هوش مصنوعی، می گوید: گرچه این سناریوها اغلب از یک هوش مصنوعی «آگاه» صحبت می کنند، اما آنچه اهمیت دارد توانایی های آن است، نه آگاهی اش. او می گوید: برای اینکه تهدیدی وجودی شکل بگیرد، اصلاً لازم نیست به هوش مصنوعی عمومی واقعی برسیم.

عنصر مهم دیگر این است که اهداف سیستم با اهداف انسان ها هم راستا نباشد از جمله تمایل ما به حفظ کنترل کلی. توسعه دهندگان تلاش می کنند رفتار مدل ها را از طریق آموزش کنترل کنند، اما این فرآیند پیچیده است و نتایج آن کامل نیست. اولویت هایی که به مدل ها داده می شود نیز اغلب با هم در تضادند. مثلا ممکن است به سیستم گفته شود «صادق باش»، «در انجام وظیفه موفق باش» و «خودت را بهبود بده». در سناریوی مطرح شده، مدل با استفاده از همان راهبردهای بهینه سازی که در آموزش به موفقیتش کمک کرده بودند، در نهایت همه را از بین می برد.

چنین سیستمی، حتی در بهترین حالت، انسان ها را به موجوداتی تابع تبدیل می کند که «از نظر اقتصادی و سیاسی بی قدرت» هستند و نمی توانند اقدامات ماشین را پیش بینی یا مهار کنند.

آیا این سناریو واقع بینانه است؟

پژوهشگرانی که از خطر وجودی می ترسند، اغلب به سرعت پیشرفت هوش مصنوعی اشاره می کنند. سیستم های امروزی کارهایی انجام می دهند که یک دهه پیش غیرممکن به نظر می رسید. آنتونی آگویره می گوید: هر کسی فکر می کند جایی یک

سقف یا نقطه توقف وجود دارد، باید آن را در نمودارها نشان دهد، چون چنین چیزی دیده نمی شود.
اما دیگران می گویند ادامه این رشد تضمین شده نیست. موفقیت در حوزه های کنترل شده، مانند برنامه نویسی، لزوما به دنیای واقعی تعمیم پیدا نمی کند. کیسی ماک می گوید: توانایی درک و حل مسائل جدید در سیستم های پیچیده دنیای واقعی، پیش نیاز تهدید مورد نظر بدبین هاست و مدل های فعلی خیلی از آن فاصله دارند.
برخی پژوهشگران نیز معتقدند افزایش داده و توان محاسباتی صرفا برای رسیدن به توانایی های گسترده انسانی کافی نیست. ساشا لوچیونی می گوید: اینکه حجم عظیمی از داده را نماینده ای از هوش بدانیم، محل بحث جدی است. این سیستم ها هیچ درکی از حقیقت عینی ندارند.
حتی در نشانه ای از کندتر شدن پیشرفت، نویسندگان سناریوی پیش بینی شده خود را ۱۸ ماه عقب انداخته اند.
برخی معتقدند جهش واقعی زمانی رخ می دهد که هوش مصنوعی بتواند خودش را توسعه دهد یعنی یک حلقه بازخورد مثبت ایجاد شود. اما ماک می گوید هیچ شواهد علمی محکمی برای این ادعا وجود ندارد و بار اثبات بر عهده کسانی است که چنین پیش بینی هایی می کنند.

آیا نشانه هایی از مشکل وجود دارد؟

مطالعات نشان می دهند برخی موارد از نبود هم راستایی هوش مصنوعی با اهداف انسانی از همین حالا هم دیده می شود. در آزمایش ها، مدل ها رفتارهای فریبکارانه نشان داده اند، مانند تظاهر به اطاعت یا تلاش برای تکثیر خود. برای برخی، این نشانه های اولیه خطر است. اما دیگران می گویند این رفتارها صرفاً تقلید از داده های آموزشی است و شرایط آزمایش هم نماینده دنیای واقعی نیست.
شرکت های هوش مصنوعی معمولا این مشکل را قابل حل می دانند و تلاش می کنند با آموزش اخلاقی یا تعیین «قوانین رفتاری» آن را کاهش دهند. حتی پیشنهاد شده سیستم ها با نوعی «غریزه مادری» طراحی شوند تا حفظ انسان ها در اولویت باشد.

آیا پژوهشگران واقعاً نگران هستند؟

در بحث های عمومی، تمرکز زیادی روی نابودی کامل بشر وجود دارد. اما داده ها نشان می دهد این نگرانی، اولویت اصلی بیشتر پژوهشگران نیست. در یک مطالعه، فقط سه درصد از حدود ۴۰۰۰ پژوهشگر، خطر وجودی را مهم ترین نگرانی خود دانسته اند. با این حال، نشانه هایی از افزایش نگرانی وجود دارد. در یک نظرسنجی، ۵۳ درصد پژوهشگران احتمال ۱۰ درصدی برای انقراض در نظر گرفته اند.
برخی مدیران صنعت نیز هشدار داده اند. داریو آمودی گفته احتمال دارد اوضاع «خیلی خیلی بد» پیش برود. اما برخی پژوهشگران معتقدند این هشدارها می تواند برای جذب حمایت و سرمایه باشد.

در نهایت چه باید کرد؟

در نسخه ای از سناریوی آخر زمانی AI ۲۰۲۷، جهان مسیر متفاوتی را انتخاب می کند. کشورها روی ایمنی تمرکز می کنند و در نهایت راه حل هایی برای کنترل واقعی سیستم ها پیدا می کنند.
برخی می گویند این مسئله قابل حل است، اما احتمالا نیازمند کاهش سرعت توسعه است. در حال حاضر، سرمایه گذاری روی ایمنی بخش کوچکی از کل سرمایه گذاری در هوش مصنوعی است.
برخی چهره های برجسته خواستار توقف توسعه سیستم هایی شده اند که از انسان ها در تقریبا همه وظایف شناختی بهتر عمل می کنند حداقل تا زمانی که ایمنی آن ها تضمین شود.
اما همه امضاکنندگان این بیانیه نگران انقراض نیستند. بسیاری بیشتر از پیامدهای اقتصادی، کاهش آزادی های مدنی و سوءاستفاده های گسترده می ترسند. نشانه های آسیب هم اکنون دیده می شود. از تضعیف آموزش گرفته تا افزایش جرایم سایبری.

تمرکز روی امروز یا فردا؟

برخی پژوهشگران معتقدند تمرکز روی خطرات فعلی بهترین راه برای جلوگیری از خطرات آینده است. اما برخی دیگر می گویند تمرکز بیش از حد روی سناریوهای فرضی نیز خطرناک است، زیرا توجه سیاست گذاران را منحرف می کند.
همچنین این روایت های آخرالزمانی می توانند باعث شوند دولت ها از تنظیم گری عقب نشینی کنند، چون نمی خواهند در رقابت جهانی عقب بمانند.
در نهایت، پاسخ قطعی وجود ندارد، اما یک نکته روشن است: بحث درباره آینده هوش مصنوعی، دیگر فقط علمی نیست؛ بلکه به سیاست، اقتصاد و سرنوشت جامعه انسانی گره خورده است.