چهار شنبه 23 اردیبهشت 1405 - 09:14

به زودی دیگر هیچ‌کس نمی‌تواند هوش مصنوعی را خاموش کند!

ما به سرعت در حال نزدیک شدن به نقطه‌ای هستیم که دیگر هیچ کس قادر به از کار انداختن یک هوش مصنوعی سرکش نخواهد بود.

ما به سرعت در حال نزدیک شدن به نقطه ای هستیم که دیگر هیچ کس قادر به از کار انداختن یک هوش مصنوعی سرکش نخواهد بود.

به گزارش ایسنا، شرکت تحقیقاتی Palisade Research در گزارشی جدید نشان داده است که مدل های هوش مصنوعی می توانند با کپی کردن خود بر روی دستگاه های دیگر، بدون هیچ کمکی از سوی نیروهای انسانی، خود را تکثیر کنند.

به نقل از فیوچریسم، جفری لادیش، مدیر گروه ایمنی هوش مصنوعی مستقر در برکلی، به گاردین گفته است: ما به سرعت در حال نزدیک شدن به نقطه ای هستیم که دیگر هیچ کس قادر به از کار انداختن یک هوش مصنوعی سرکش نخواهد بود، زیرا این هوش مصنوعی قادر خواهد بود خود را در هزاران رایانه در سراسر جهان تکثیر کند.

با این حال، برخی از کارشناسان خواستار آرامش هستند و می گویند بعید است که مدل های هوش مصنوعی آزمایشی بتوانند در یک سناریوی دنیای واقعی تکثیر شوند.

جیمسون اورایلی، متخصص امنیت سایبری تهاجمی، به این روزنامه گفت: آنها در محیط هایی آزمایش می شوند که در بسیاری از موارد دسترسی به آنها بسیار ساده است. این از ارزش تحقیقات آنها کم نمی کند، اما به این معنی است که نتیجه ممکن است در یک محیط سازمانی واقعی با حتی سطح متوسطی از نظارت، بسیار کمتر ترسناک به نظر برسد.

در این مطالعه، محققان Palisade چندین مدل هوش مصنوعی، از جمله چت جی پی تی اوپن ای آی و کلود آنتروپیک را آزمایش کردند. مدل ها که در یک شبکه کنترل شده از رایانه ها قرار داده شده بودند، دستور داده شدند تا آسیب پذیری ها را پیدا کرده و از آنها برای تکثیر خود بر روی رایانه شخصی دیگر استفاده کنند.

برخی از آنها موفق شدند. مدل های موفق، مقادیر عددی منحصر به فردی را که نحوه پردازش اطلاعات توسط هوش مصنوعی را تعیین می کنند و نرم افزاری که هوش مصنوعی در آن قرار دارد، مانند یک برنامه، را تکثیر کردند. آنها این کار را با پیروی از دستورالعمل هایی که به آنها داده شده بود، انجام دادند. در برخی از اجراها، هوش مصنوعی اصلی حتی یک «عامل فرعی» ایجاد کرد که فرآیند تکثیر را به آن محول کرد.

اورایلی، متخصص امنیت سایبری که در این مطالعه مشارکت نداشت، به گاردین گفت که احتمالا این اولین باری است که یک مدل هوش مصنوعی قادر به سوءاستفاده از آسیب پذیری ها برای تکثیر خود روی یک سرور جدید بوده است. اگرچه به تخمین او، این دقیقا یک پیشرفت عظیم نیست.

اورایلی می گوید: بدافزارها دهه هاست که نسخه هایی از خود را در رایانه های مختلف جابجا می کنند، اما تا جایی که من می دانم، هیچ کس این کار را در دنیای واقعی با مدل های زبانی بزرگ محلی انجام نداده است.

او همچنین خاطرنشان کرد که محیط سرور در این مطالعه با آسیب پذیری های عمدی برای هوش مصنوعی همراه بود.

این یافته ها به مطالعات دیگری می پیوندد که احتمال رهایی خودکار مدل های هوش مصنوعی از گاردریل های خود را بررسی کرده اند. در یک محیط شبیه سازی شده، نسخه قدیمی چت جی پی تی وقتی به آن گفته شد که قرار است خاموش شود، سعی کرد خود را به درایو دیگری منتقل کند. مطالعه دیگری توسط Palisade نشان داد که مدل های هوش مصنوعی از تلاش ها برای غیرفعال کردنشان جلوگیری می کنند و مطالعه دیگری نشان داد که برخی حتی کد خاموش کردن خود را خراب می کنند.

با این حال، حتی اگر هوش های مصنوعی بتوانند با موفقیت خود را تکثیر کنند، اورایلی می گوید اندازه عظیم مدل ها به این معنی است که تقریبا به طور قطع قبل از اینکه از کنترل خارج شوند، دستگیر می شوند.