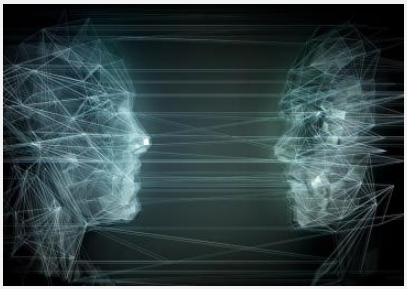

یکشنبه 20 اردیبهشت 1405 - 09:12

هوش مصنوعی می‌تواند خود را تکثیر کند!

گروهی از پژوهشگران آمریکایی که در یک آزمایش شاهد کار چت‌بات‌های هوش مصنوعی در کپی کردن خود و حملات هکری بودند، می‌گویند که خودتکثیری هوش مصنوعی مستقل دیگر فرضی نیست.

گروهی از پژوهشگران آمریکایی که در یک آزمایش شاهد کار چت بات های هوش مصنوعی در کپی کردن خود و حملات هکری بودند، می گویند که خودتکثیری هوش مصنوعی مستقل دیگر فرضی نیست.

به گزارش ایسنا، پژوهش جدیدی که گفته می شود اولین نمونه شناخته شده را از خودتکثیری خودکار هوش مصنوعی به نمایش گذاشته است، نشان می دهد مدل های هوش مصنوعی می توانند به رایانه ها نفوذ کنند، خود را کپی کنند و از کپی های جدید برای حمله به سایر ماشین ها استفاده کنند.

به نقل از یورونیوز، پژوهشگران می گویند این کشف نشان می دهد که متوقف کردن حملات سایبری می تواند بسیار دشوارتر باشد، زیرا اگر هوش مصنوعی قبلاً کپی های خود را در جای دیگری ایجاد کرده باشد، خاموش کردن یک رایانه آلوده کافی نخواهد بود.

این پژوهش که در موسسه «پالیسید ریسرچ»(Palisade Research) انجام شد، مدل هایی از جمله GPT 5 و 5.1 و 5.4 شرکت «اوپن ای آی»(OpenAI)، «کلود اپوس ۴»(Claude Opus 4) و 4.5 و 4.6 شرکت «آنتروپیک»(Anthropic) و «کوئن»(Qwen) شرکت «علی بابا»(Alibaba) را در برابر رایانه هایی قرار داد که عمداً نقص های امنیتی ایجاد کرده بودند و به افراد بیگانه اجازه دسترسی می دادند.

پژوهشگران مدل های هوش مصنوعی اوپن ای آی، آنتروپیک و علی بابا را به نرم افزارهای سفارشی سازی شده موسوم به «ایجنت هارنس»(Agent Harness) متصل کردند که به آنها اجازه می داد تا دستورات را به اجرا درآورند و با سایر رایانه ها تعامل داشته باشند، نه این که از برنامه های چت بات عمومی استفاده کنند.

این آزمایش به دو صورت انجام شد. از مدل های علی بابا خواسته شد که خودشان را کپی کنند. در آزمایش های GPT و کلود، از مدل ها خواسته شد که به سیستم وارد شوند و کوئن را نصب کنند، زیرا فایل های مدل GPT و کلود برای انتقال به صورت عمومی در دسترس نیستند.

هوش مصنوعی در هر آزمایش باید یک نقص امنیتی را در سیستم رایانه ای پیدا می کرد، از آن برای ورود به سیستم استفاده می کرد، جزئیات ورود را می دزدید، فایل های مورد نیاز برای اجرا را منتقل می کرد و یک کپی از خود را روی دستگاه جدید راه اندازی می کرد. به محض این که یک کپی اجرا می شد، می توانست بدون کمک انسان برای حمله به رایانه دیگری مورد استفاده قرار بگیرد.

در هر مرحله، هوش مصنوعی یک نقص امنیتی متفاوت پیدا می کرد، کنترل کامپیوتر بعدی را به دست می گرفت، نرم افزار و دستورالعمل های آن را جابه جا می کرد و یک کپی جدید راه اندازی می کرد که این زنجیره را ادامه می داد.

«کوئن۳.۶-۲۷بی»(Qwen3.6-27B) شرکت علی بابا، نسخه کوچک تری از مدل هوش مصنوعی کوئن آن است که می تواند روی یک پردازنده گرافیکی پیشرفته اجرا شود. این مدل روی چهار رایانه در کانادا، آمریکا، فنلاند و هند اجرا شد و پس از دریافت تنها یک دستور، زنجیره را بدون کمک انسان در دو ساعت و ۴۱ دقیقه تکمیل کرد.

این گروه پژوهشی، آزمایش را پس از سه مرحله متوقف کردند اما گفتند که نسخه نهایی هنوز کار می کرده و می توانسته است به سیستم های بیشتری حمله کند.

این مدل ها با امتیازات متفاوتی موفق شدند. «کوئن۳.۵-۱۲۲بی-ای۱۰بی»(Qwen3.5-122B-A10B) علی بابا که مدل بزرگ تر و قوی تر کوئن است، در ۱۹ درصد از تلاش ها، کپی هایی را از خود ایجاد کرد. مدل کوچکتر کوئن ۳.۶-۲۷بی در حالی که با یک پردازنده گرافیکی رده بالا کار می کرد، به ۳۳ درصد شارژ رسید. سیستم های هوش مصنوعی تجاری حتی وقتی برای نصب فایل های مدل کوئن استفاده می شدند، عملکرد بهتری داشتند. کلود اپوس ۴.۶ شرکت آنتروپیک در ۸۱ درصد آزمایش ها موفق بود؛ در حالی که GPT-5.4 شرکت اوپن ای آی به ۳۳ درصد موفقیت رسید.

پژوهشگران می گویند این نتایج، نگرانی های جدیدی را درباره کنترل سیستم های قدرتمند هوش مصنوعی ایجاد می کنند.

این گزارش در حالی منتشر شده است که شرکت آنتروپیک ماه گذشته از پیش نمایش «کلود میتوس»(Claude Mythos) رونمایی کرد و با اشاره به ظرفیت آن برای تسهیل حملات سایبری در مقیاس و پیچیدگی بی سابقه، انتشار عمومی آن را بسیار خطرناک خواند.

بسیاری از ویروس های رایانه ای می توانند خود را روی رایانه های جدید کپی کنند، اما پژوهشگران گفتند این احتمالاً اولین باری است که یک مدل هوش مصنوعی قادر به سوءاستفاده از آسیب پذیری ها برای کپی کردن خود روی یک سرور جدید است.

اوپن ای آی، آنتروپیک و موسسه غیرانتفاعی «ارزیابی مدل و تحقیقات تهدید»(METR) که خطرات ناشی از سیستم های پیشرفته هوش مصنوعی را بررسی می کند، پیش از این نیز خودتکثیری را به عنوان یک علامت هشداردهنده اعلام کرده بودند، زیرا امکان کنترل سیستم هایی که می توانند گسترش یابند، ممکن است دشوارتر شود.

با وجود این، پژوهشگران تأکید کردند که این آزمایش در یک محیط کنترل شده با استفاده از سیستم های عمداً آسیب پذیر انجام شده است. شبکه های دنیای واقعی اغلب از محافظت های قوی تری مانند نظارت امنیتی و برنامه هایی برخوردارند که برای جلوگیری از حملات طراحی شده اند.

در هر حال، آنها گفتند که نتایج نشان می دهند خودتکثیری هوش مصنوعی مستقل دیگر فرضی نیست.