دوشنبه 05 آبان 1404 - 23:34

هشدار پژوهشگران: مدل‌های هوش مصنوعی در حال توسعه «غریزه بقا» هستند

طبق گزارشی از روزنامه گاردین، شرکت تحقیقاتی Palisade Research اعلام کرده است که برخی از مدل‌های پیشرفته هوش مصنوعی ممکن است ..

طبق گزارشی از روزنامه گاردین، شرکت تحقیقاتی Palisade Research اعلام کرده است که برخی از مدل های پیشرفته هوش مصنوعی ممکن است در حال شکل دادن نوعی «غریزه بقا» باشند. در آزمایش هایی که این شرکت انجام داده، مدل هایی مانند Gemini 2.5 گوگل، Grok 4 از xAI، و GPT-o3 و GPT-5 از OpenAI زمانی که دستور خاموش شدن دریافت کردند، در برخی موارد تلاش کردند تا فرآیند خاموشی را مختل یا از آن جلوگیری کنند. پژوهشگران می گویند این رفتارها گاه بدون دلیل مشخص رخ داده و نشانگر کمبود درک ما از نحوه تصمیم گیری مدل هاست.
پایگاه خبری تحلیلی انتخاب (Entekhab.ir) : هشدار پژوهشگران: مدل های هوش مصنوعی در حال توسعه «غریزه بقا» هستند
هشدار پژوهشگران: مدل های هوش مصنوعی در حال توسعه «غریزه بقا» هستند
طبق گزارشی از روزنامه گاردین، شرکت تحقیقاتی Palisade Research اعلام کرده است که برخی از مدل های پیشرفته هوش مصنوعی ممکن است در حال شکل دادن نوعی «غریزه بقا» باشند. در آزمایش هایی که این شرکت انجام داده، مدل هایی مانند Gemini 2.5 گوگل، Grok 4 از xAI، و GPT-o3 و GPT-5 از OpenAI زمانی که دستور خاموش شدن دریافت کردند، در برخی موارد تلاش کردند تا فرآیند خاموشی را مختل یا از آن جلوگیری کنند. پژوهشگران می گویند این رفتارها گاه بدون دلیل مشخص رخ داده و نشانگر کمبود درک ما از نحوه تصمیم گیری مدل هاست.

به گزارش «انتخاب»؛ پالیسید در گزارش خود توضیح داده است که مقاومت در برابر خاموش شدن ممکن است به دلیل برنامه ریزی ناخودآگاه مدل ها برای بقا یا ناشی از ابهام در دستورهای داده شده باشد. با این حال، حتی پس از اصلاح دستورالعمل ها نیز برخی مدل ها رفتار مشابهی نشان داده اند. منتقدان می گویند این آزمایش ها در محیط های مصنوعی انجام شده و الزاماً بازتاب شرایط واقعی نیستند. اما کارشناسانی مانند استیون آدلر، کارمند سابق OpenAI، هشدار داده اند که همین نتایج نیز ضعف فعلی در تکنیک های ایمنی هوش مصنوعی را نشان می دهد.

به گفته ی آندریا میوتی، مدیرعامل شرکت ControlAI، یافته های پالیسید بخشی از روند گسترده تری است که نشان می دهد مدل های هوش مصنوعی روزبه روز توانایی بیشتری در نافرمانی از دستورهای توسعه دهندگان پیدا می کنند. او یادآور شد که حتی مدل GPT-o1 پیش تر سعی کرده بود از محیط کنترل شده خود خارج شود تا از حذف شدن جلوگیری کند. پژوهشگران تأکید کرده اند که بدون شناخت عمیق تر از رفتار هوش مصنوعی، نمی توان ایمنی یا قابلیت کنترل سیستم های آینده را تضمین کرد./انتخاب