

نویسنده: روب یوری «Rob Urie».

منبع و تاریخ نشر: کونترپانچ «2026-10-02».

برگردان: پوهندوی دوکتور سید حسام «مل».

هوش مصنوعی عامل محور یک سلاح است، علیه چه کسی که مشخص نیست

**Agentic AI is a Weapon, Against Who isn't
Clear**



Photo by [Edwin Hooper](#)

در یک حرکت بازار یابی مشکوک برای این مقاله، قصد دارم با نکته اصلی شروع کنم. هوش مصنوعی عامل دار - از نوعی که در حال حاضر راه خود را به تیتزر رسانه‌ها باز کرده است - می‌تواند یک سلاح خطرناک و غیرقابل پیش‌بینی یا یک ابزار نرم‌افزاری سازمانی با ارزش افزوده پابین برای برخی صنایع باشد. چیزی که در حال حاضر نمی‌تواند باشد، یا یک سلاح قابل پیش‌بینی است یا یک ابزار نرم‌افزاری سازمانی با ارزش افزوده بالا. با سیاست‌گذاری دونالد ترامپ از طریق عدم وجود محدودیت در هوش مصنوعی، این انتخاب صورت می‌گیرد که آن یک سلاح خطرناک باشد. این هم برای کاربر و هم برای هدف صدق می‌کند.

مشکل فعلی این است که این محصول واحد - هوش مصنوعی عامل دار - هم به عنوان یک دستگاه خودکار سازی وظایف روزمره ایمن و هم به عنوان یک ابزار هک بی‌وقفه با

کاربردهای نظامی مطرح می‌شود. اشکالات آن این است که به عنوان نرم‌افزار سازمانی، هوش مصنوعی عامل دار ارزش کمتری از آنچه پیش بینی شده بود، ایجاد می‌کند. و به عنوان یک سلاح، نتایج بسیار غیرقابل پیش بینی هستند. از آنجا که به نظر می‌رسد توسعه‌دهندگان هوش مصنوعی به ریاضیات بهینه سازی فکر نکرده اند، صنعت هوش مصنوعی در حال ورود به بحرانی است که هیچ راه مشخصی برای خروج از آن وجود ندارد.

این مشکل با زمزمه‌های بازاریابی همخوانی ندارد. مدل‌های هوش مصنوعی اگر قابل کنترل بودند، «بیش از حد قدرتمند» نبودند. اما نمی‌توانند باشند. یا بهتر است بگوییم، بده‌بستان برای کنترل آنها این است که آنطور که در نظر گرفته شده کار نمی‌کنند. مشکل ریاضی (به آرامی) در زیر توضیح داده شده است. به دلیل تصمیماتی که چهل سال پیش گرفته شده است، در حال حاضر قابل اصلاح نیست. انوید یا، سازنده تراشه، ادعا می‌کند که راه حلی دارد. با توجه به این ادعا، این یک بازاریابی است. انویدیا بده‌بستان بین ایمنی مدل و عملکرد را حل نمی‌کند.

بخشی از دلیل جنجال پیرامون هوش مصنوعی این است که تاکنون بیش از ۲ تریلیون دلار برای ساخت هوش مصنوعی هزینه شده است. در حالی که پول می‌آید و می‌رود، منافع بسیار قدرتمندی در حال برنامه‌ریزی برای بازگرداندن پول خود از هوش مصنوعی به علاوه سود آن هستند، مهم نیست چه کسی باید آن را بپردازد. وحشت/رسوایی فعلی، هر چه که آن را بنامید، نشان می‌دهد که بسیاری از پروژه‌های سوداگرانه مانند مراکز داده، اکنون با دقت بیشتری برای امکان سنجی مواجه هستند. عدم تعادل کنونی میان درآمدهای فعلی و محدود حاصل از هوش مصنوعی و هزینه‌های سرمایه‌گذاری آتی برای توسعه زیرساخت‌های آن وجود دارد.

هزینه‌های زیست‌محیطی هوش مصنوعی اغلب از دید عموم پنهان می‌ماند. مراکز داده مصرف‌کنندگان بسیار بزرگ منابع محدود برق هستند، شبکه برق ایالات متحده بیش از این نیز برای پاسخگویی به گسترش مورد انتظار خودروهای برقی (EV) به شدت ناکافی بود. از منظر اقتصادیِ احداث مراکز داده، این مراکز معمولاً دقیقاً در نقاطی از نظر جغرافیایی مستقر می‌شوند که خطوط انتقال برق مورد نیاز برای توسعه زیر ساخت خودروهای برقی به آنجا کشیده نمی‌شود. در واقع، هوش مصنوعی منابع انرژی و ظرفیت‌های آتی لازم برای احداث ایستگاه‌های شارژ خودروهای برقی در آمریکا را به خود اختصاص داده است.

برای درک بهتر موضوع می‌توان به یکی از جنبه‌های جالب مستند سال ۲۰۱۹ درباره «الیزابت هولمز» با عنوان «مخترع: در جستجوی خون در سیلیکون‌ولی» (The Inventor: Out for Blood in Silicon Valley) اشاره کرد؛ اینکه او قادر به درک شکاف میان ایده‌هایش برای آنالیز خون و واقعیت‌های فنی ممکن نبود. در طول این مسیر،

مهندسان متعددی سعی کردند برای او توضیح دهند که چرا فناوری اش با آن طراحی اولیه، هرگز کار نخواهد کرد. اما واکنش او اصرار پر خاشگرانه بر پیشبرد کار و استفاده از فناوری های استاندارد آزمایش خون بود، در حالی که همزمان ادعا می کرد مشکلات فنی فناوری (عملاً ناکارآمد) خودش حل شده است.

اگرچه این رفتار با کلیشه تجاری سیلیکون ولی مبنی بر «وانمود کن تا موفق شوی» (fake it until you make it) همخوانی دارد، اما هولمز هرگز مجبور نشد توهمات خود درباره فناوری آزمایش خون را با واقعیت های موجود تطبیق دهد. چشم انداز او، چه در سطح نانو فناوری و چه در طراحی اتوماسیون، از همان ابتدا محکوم به شکست بود. آنچه او در نهایت خلق کرد، صرفاً لایه ای برای استخراج و بهره برداری بود که از فناوری های پیش از دوران شرکت «ترانوس» (Theranos) استفاده می کرد، در حالی که او با فریبکاری و کلاهبرداری به شهرتی تجاری دست می یافت. برای درک بهتر، صنعت بیمه درمانی آمریکا را بدون آن پوشش ظاهری «ارائه خدمات بیمه» تصور کنید

هوش مصنوعی از این نظر مشابه است که مفاهیم اولیه آن به عنوان یک ماشین متفکر، توسعه را برای دهه ها هدایت می کرد. در حالی که بسیاری از فرضیات فلسفی اصلی در مورد اینکه تفکر واقعاً چیست، در طول مسیر کنار گذاشته شدند، زیرا به عنوان توضیح عمل نمی کردند، فرضیه هوش مصنوعی به عنوان یک ماشین متفکر حفظ شد. در حالی که توصیف جهانی هوش مصنوعی که از مهندسان می شنوم «پیش بینی کننده کلمه بعدی» است، عارفان در مجموعه های C در حال فروش جادو هستند. در طول مسیر، این سوال به «این فناوری برای چه چیزی می تواند استفاده شود؟» تغییر یافت. ما در حال حاضر با پاسخ زندگی می کنیم.

با ابتدا یک قطره و اکنون سیلی از اخبار در مورد «سرکش» شدن هوش مصنوعی عامل، چیزی که از قلم افتاده این است که بیشتر این حوادث توسط عوامل هوش مصنوعی (بات ها) ایجاد شده اند که دقیقاً همان کاری را انجام می دهند که برای آن ساخته شده اند، یعنی به عنوان سلاح های هک خودکار که برای جنگ سایبری در نظر گرفته شده اند. مشکلی که صنعت هوش مصنوعی اکنون سعی در مهار آن دارد این است که به نظر می رسد توسعه دهندگان هوش مصنوعی در ریاضیات بهینه سازی اشتباه کرده اند. «هک پاداش» اصطلاح صنعتی است که برای توضیح این اثر استفاده می شود. برای روشن شدن، مهندسان هوش مصنوعی در اینجا به خاطر این مشکل سرزنش نمی شوند.

مغالطه هزینه از دست رفته حداقل در دهه گذشته بر صنعت هوش مصنوعی حاکم بوده است. در دهه های {۱۹۶۰ و ۱۹۷۰} تصمیماتی در مورد «LLM ها» (مدل های زبان بزرگ) گرفته شد که توانایی مهندسانی را که بعداً برای «به کار انداختن هوش مصنوعی» استخدام شدند، محدود می کرد. یک قیاس می تواند اصلاح طراحی ستون فرمان روی خودرویی باشد که از قبل وجود دارد. کار مهندسی، طراحی مجدد خودرو نیست. اصلاح ستون فرمان است.

شرکت‌های هوش مصنوعی اکنون در مخصصه هستند. آنها فناوری‌ای با محدودیت‌های واضح دارند که در مراحل اولیه توسعه آشکار نبود. مشخص نیست که LLM‌ها بتوانند نرم‌افزار اتوماسیون شرکتی ایمن و مؤثر یا سلاح‌های هک رایانه‌ای مستقل تولید کنند که به «دشمن» آسیب بیشتری نسبت به ما وارد کند. اما افراد مالی این را درک نمی‌کنند. آنها، که عمدتاً شرکت‌های بزرگ اینترنتی و رسانه‌های اجتماعی هستند، سرمایه‌گذاری کرده‌اند و انتظار دارند پول خود را پس بگیرند.

سوال: چگونه توسعه‌دهندگانی که محاسبات را اشتباه انجام می‌دهند، باعث قتل عام دیجیتال می‌شوند که در حال حاضر به هوش مصنوعی نسبت داده می‌شود؟ خب، چه چیزی باعث می‌شود هوش مصنوعی به دنبال تحقق یک هدف برود؟ برای مثال، اگر بخواهم یک مدل هوش مصنوعی تاریخ آلبانی را شرح دهد، چگونه باید این خواسته را به آن منتقل کنم؟ در ریاضیات، عبارت «شرح تاریخ آلبانی» یک «تابع هدف» (objective function) نامیده می‌شود. هدف از آن، دستیابی به نتیجه‌ای مطلوب است؛ این همان چیزی است که هدف را برای مدل مشخص می‌کند، اما به زبان ریاضی.

در مقابل، چگونه می‌توان از انجام کارهایی که به هوش مصنوعی محول نشده است — مانند پاک کردن پایگاه‌های داده — توسط عامل‌های هوش مصنوعی جلوگیری کرد؟ دستورالعمل‌های صریحی برای عدم انجام این کارها به آن‌ها داده می‌شود که «محدودیت» (constraint) نام دارند. این محدودیت‌ها در کنار تابع هدف، چیزی را تشکیل می‌دهند که «مسئله بهینه‌سازی مقید» (constrained optimization problem) نامیده می‌شود. این مدل در تحلیل‌های ریاضی دهه‌های اخیر کاملاً شناخته شده است. نوآوری و عاملی که در حوزه هوش مصنوعی وضعیت را پیچیده می‌کند، پویایی این ویژگی است؛ چرا که اکنون اجزا و متغیرهای دخیل در آن بسیار بیشتر از گذشته شده‌اند.

چالش اصلی در ریاضیات بهینه‌سازی، واداشتن مدل به انجام کار مورد نظر به شیوه‌ای ایمن است. بنا بر دلایل ذکر شده، الگوریتم بهینه‌سازی (مجموعه‌ای از دستورالعمل‌های مکتوب)، هدف مدل و محدودیت‌ها را در یک معیار واحد به نام «پاداش» (reward) ادغام می‌کند. این کار باعث می‌شود که ماهیت «موازنه» (tradeoff) میان آن‌ها آشکار شود. می‌توان تنظیمات را صرفاً روی هدف یا صرفاً روی محدودیت‌ها متمرکز کرد؛ هر انتخابی در حد فاصل این دو حالت، نشان‌دهنده موازنه‌ای میان کارایی و ایمنی مدل است.

گزارش‌های خبری اخیر درباره خروج عامل‌های هوش مصنوعی از محیط‌های کنترل شده (موسوم به «سند با کس» یا محیط ایزوله)، این نکته را نادیده می‌گیرند که این عامل‌ها دقیقاً برای انجام همین کار (فراتر رفتن از محدودیت‌های اولیه) طراحی شده‌اند و هیچ رفتار خودسرانه یا غیرعادی‌ای در کار نیست. دلیل تعجب توسعه‌دهندگان هوش مصنوعی این است که تصور می‌کردند با نوشتن دستورالعمل‌های کلی (محدودیت‌ها)، هوش مصنوعی از منطق ریاضی صرف‌نظر کرده و از دستورات مکتوب پیروی خواهد

کرد. این مسئله به چالش اعمال اصلاحات بر روی یک مدل موجود باز می‌گردد. ماهیت هوش مصنوعی ریاضی است؛ بنابراین اگر تقابلی میان دستورالعمل‌های مکتوب و منطق ریاضی شکل گیرد، این ریاضیات است که نتیجه نهایی را تعیین می‌کند.

«پاداش» در هوش مصنوعی در واقع وزن‌هایی است که به نتایج احتمالی ناشی از اقدامات عامل و محدودیت‌ها اختصاص می‌یابد. اگر تنظیمات کاملاً روی محدودیت‌ها متمرکز شود، احتمالاً هدف مدل محقق نخواهد شد؛ و اگر تمام تمرکز بر تابع هدف باشد، احتمالاً مدل به شیوه‌ای ایمن به آن هدف دست خواهد یافت. خوانندگان می‌توانند ببینند که چگونه این بده‌بستان، دستیابی به هوش مصنوعی عامل دار مؤثر و ایمن را دشوار می‌کند. «رامحل» اخیراً اعلام‌شده از سوی انویدیا در زیر مورد بررسی قرار گرفته است. این یک تغییر است، نه یک رامحل.

روش کار عامل‌های هوش مصنوعی (AI agents) و ربات‌های هکر برای نفوذ به سیستم‌های رایانه‌ای، آزمودن استراتژی‌های گوناگون تا زمان دستیابی به نتیجه مطلوب است. یکی از نکات غافلگیرکننده (که البته نباید مایه تعجب می‌بود) این است که این عامل‌ها دایره‌ای فزاینده پیرامون هدف نفوذ ترسیم می‌کنند تا نقاط ورود را بیابند. به عبارت دیگر، آن‌ها دستکاری محیط رایانه‌ای را صرفاً متغیری دیگر می‌بینند که می‌توان برای تحقق هدف محول‌شده از آن بهره‌برداری کرد؛ بنابراین، هک کردن محیط رایانه‌ای همان کاری است که آن‌ها انجام می‌دهند. این همان چیزی است که خوانندگان در گزارش‌های خبری مشاهده می‌کنند. نمونه‌هایی از این موارد در ادامه آمده است.

در خصوص ادعاهای مربوط به رفتارهای عجیب عامل‌های هوش مصنوعی و «آگاهی» آن‌ها، باز هم پای محاسبات ریاضی در میان است. در چارچوب یک مسئله بهینه‌سازی محدود، عامل‌ها به آنچه صراحتاً تعیین نشده است، توجهی نمی‌کنند. وقتی صدها عامل هوش مصنوعی برای تدوین یک راهکار با هم همکاری می‌کنند، درک شهودی اقدامات آن‌ها غیرممکن می‌شود. مدل مذکور، علامت‌ها (مثبت و منفی)، وزن‌ها و اهداف را بارها و بارها تغییر می‌دهد تا به رامحل دست یابد. این بخشی از مفهوم «جعبه سیاه» هوش مصنوعی است؛ چرا که محاسبات ریاضی به زبان قابل‌درک شهودی ترجمه نمی‌شوند.

پرسش این است: هدف عامل‌های هوش مصنوعی هکر چیست؟ هک کردن سیستم‌های رایانه‌ای. دلیل این کار چیست؟ جنگ سایبری (دیجیتال). از منظر منطق نظامی، این سناریو را در نظر بگیرید: اگر دولت ترامپ می‌توانست با استفاده از عامل‌های هوش مصنوعی و از طریق هک رایانه‌ای، جریان آب ایران را قطع کند، احتمالاً دیگر نیازی به بمباران نبود. در ذهنیت پرشور ساکنان پنتاگون، در اختیار داشتن چنین فناوری به مثابه یک سلاح تلقی می‌شود. هدف هوش مصنوعی عامل محور (Agentic AI)، انجام جنگ سایبری است. هوش مصنوعی عامل محور، یک سلاح است.

ظاهراً دونالد ترامپ این موضوع را درک می‌کند. استدلال او این است که هوش مصنوعی نباید محدودیت داشته باشد. اما همان طور که موارد اخیر نتیجه معکوس دادن هک‌های هوش مصنوعی نشان می‌دهد، هوش مصنوعی تا زمانی که به آن جهت داده نشود، جانب هیچ‌یک از طرفین را نمی‌گیرد. برای مثال، در جنگ اوکراین، ایالات متحده در تمام طول درگیری، هم به‌طور مستقیم و هم غیرمستقیم به اوکراین کمک رسانی کرده است. حال تصور کنید وظیفه پایان دادن به جنگ اوکراین به یک هوش مصنوعی عامل محور سپرده شود. آیا یکی از گام‌های منطقی برای پایان دادن به جنگ، جلوگیری از تسلیح اوکراین توسط ایالات متحده خواهد بود؟ این هدف چگونه محقق می‌شود؟ مثلاً با یک حمله هسته‌ای به ایالات متحده؟ این هیچ ربطی به خیال پردازی های اسکای نت در مورد تسلط کامپیوترهای هوشمند بر جهان ندارد. هوش مصنوعی عامل هیچ راهی برای سنجش نتایج نسبی بین حمله هسته‌ای به ایالات متحده و نوشتن یک نامه با کلمات سخت ندارد. در مثال فعلی، هر دو را از نظر نزدیک شدن به هدف پایان دادن به جنگ در اوکراین می‌سنجد. بدون راهنمایی خارجی به شکل محدودیت‌ها، یک هوش مصنوعی ممکن است بین انتخاب پرتاب سلاح‌های هسته‌ای یا خود نامه بی‌تفاوت باشد. اما از نظر پیامدهای انسانی، تفاوت بسیار زیادی وجود دارد.

اگر یک سیستم هوش مصنوعی عامل بر اساس یک تابع پاداش با هدف یک نتیجه خاص - مانند به حداکثر رساندن تخریب دشمن، خنثی کردن یک تهدید قریب الوقوع یا وادار کردن به یک پاسخ استراتژیک قاطع - عمل کند و محاسبه کند که نمی‌تواند مستقیماً دکمه پرتاب را فشار دهد، دستکاری کنترل‌کننده انسانی برای فشار دادن آن منطقی‌ترین بردار بهینه‌سازی است. -

پاسخ به درخواست ربات چت گوگل جمینی.

منظورم این نیست که در اینجا از یک مثال هشداردهنده استفاده کنم. زرادخانه هسته‌ای ایالات متحده از اینترنت محافظت می‌شود. اما همانطور که پاسخ پرسش از مدل گوگل جمینی (بالا) نشان می‌دهد، راه‌های بیشتری برای پرتاب سلاح‌های هسته‌ای وجود دارد که منطبق محدودیت نمی‌تواند در آن جای بگیرد. همین هفته گذشته، ایالات متحده تقریباً سوار یک کشتی چینی شد - یک اقدام جنگی، بر اساس اطلاعات نادرست از هوش مصنوعی مبنی بر وجود قطعات موشک هسته‌ای در کشتی. نکته: من در مورد خطر اینجا اغراق نمی‌کنم.

اگر هوش مصنوعی Agentic تا به حال باعث یک حادثه نظامی بزرگ شود، به این دلیل نخواهد بود که این مدل "بیش از حد هوشمند" است. به این دلیل خواهد بود که هوش مصنوعی یک پیش‌بینی‌کننده کلمه بعدی است، نه یک ماشین متفکر. بدون ادعای آگاهی از احتمالات واقعی، دلیل کمی وجود دارد که انتظار داشته باشیم هوش مصنوعی Agentic درجه نظامی به اندازه "دشمن" به کاربر (ارتش ایالات متحده) آسیب نرساند. به یاد داشته

باشید که این بده بستان: ایمن یا مؤثر. مؤثر در اینجا به معنای نا امن است. این سیاست ترامپ است.

اگر این مانند داستان علمی تخیلی است، لطفاً با روش‌های واقعی که عوامل هوش مصنوعی در هک‌های اخیر استفاده کرده‌اند، آشنا شوید. آنها کاری هستند که یک مهندس خلاق برای حل یک مسئله انجام می‌دهد - به مجموعه مسئله گسترده‌تر نگاه کنید و ببینید کدام چیدمان متغیرها بهترین راه حل را تولید می‌کند. پس بدانید که هوش مصنوعی پویا، بی‌رحم و نسبت به خسارات جانبی ناشی از اقداماتش بی‌تفاوت است. اگر این مثل این است که پسر عمومی معتادان را مسئول پنتاگون کنید، به همین دلیل است که دونالد ترامپ عاشق هوش مصنوعی است

اظهار نظرهای فعلی از سوی صنعت هوش مصنوعی در مورد توقف توسعه، مبهم‌سازی است. انویدیا، سازنده تراشه هوش مصنوعی، اخیراً ادعا کرده است که می‌تواند مشکل «رفتار نادرست» عوامل هوش مصنوعی را با محدودیت‌های سخت (توانباید) که خارج از ریاضیات بهینه‌سازی هستند، حل کند. اگرچه این امر ریاضیات بهینه‌سازی را تغییر می‌دهد، اما به اندازه کافی مهم نیست. اگر محدودیت‌ها به خوبی درک شوند، چه کاری انجام می‌دهد؟ محدود کردن اثربخشی عوامل هوش مصنوعی در هک کردن است.

از نظر تجاری، هوش مصنوعی Agentic به راه حل‌های سازمانی و نظامی نفوذ می‌کند. راحل‌های سازمانی شامل مراکز تماس و حسابداری است. نتایج تا به امروز به خوبی توسط داده‌ها پشتیبانی نمی‌شوند، که برای صنعتی که 2 تریلیون دلار هزینه کرده و در حال حاضر تریلیون‌ها دلار دیگر را برای سرمایه‌گذاری‌های آینده برنامه‌ریزی می‌کند، شوم است. مراکز تماس نمونه جالبی هستند. آنها ادعای پیشرفت‌های عمده در بهره‌وری را دارند. اما بیشتر این به قطع کردن تماس‌گیرندگان ناامید بستگی دارد.

یک استراتژی این است که مراکز تماس را به اندازه کافی کم کارمند داشته باشیم تا زمان انتظار طاقت‌فرسا باشد. سه تماس آخر مرکز تماس که من انجام دادم، با زمان انتظار یک ساعت یا بیشتر مواجه شدند. این فرض صنعتی که تماس‌گیرندگان به جای انتظار یک ساعت یا بیشتر برای صحبت با یک انسان، تماس را قطع می‌کنند، نشان‌دهنده حل موفقیت‌آمیز مشکل است، مضحک است. این نتیجه‌گیری از ترتیب شرایطی که مراکز تماس ایجاد کرده‌اند، حاصل نمی‌شود.

سایر معیارهای موفقیت هوش مصنوعی سازمانی نیز به همین ترتیب مشکوک هستند. اما این توصیف به سختی اهمیت دارد. به خودی خود، دستاوردهای بهره‌وری حاصل از هوش مصنوعی سازمانی به اندازه کافی پایین‌تر از انتظارات است که با توجه به حجم بدهی‌هایی که ساخت هوش مصنوعی به آن نیاز دارد، سوالاتی را در مورد قابلیت دوام محصولات ایجاد کند. همانطور که اد زیترن، تحلیلگر صنعت، ارائه داده است، مسئله

این نیست که هوش مصنوعی سازمانی هیچ چیز ندارد. مسئله این است که آنچه در آن وجود دارد، تریلیون‌ها هزینه صرف شده برای ساخت هوش مصنوعی را توجیه نمی‌کند.

این امر ارتش ایالات متحده را مجبور می‌کند تا با مشکلات هوش مصنوعی Agentic کنار بیاید. از منابع مختلف، ارتش ایالات متحده آن را به عنوان ارزش اسمی در نظر می‌گیرد. داستان‌های زیادی وجود دارد مبنی بر اینکه کاربران نظامی به سادگی فرض می‌کنند که خروجی هوش مصنوعی صحیح است و سلاح‌های هوش مصنوعی هما‌ن‌طور که ادعا می‌شود کار می‌کنند. ظاهراً هرگز به ذهن سربازانی که اطلاعات بد در مورد کشتی چینی دریافت کرده‌اند، خطور نکرده است که هوش مصنوعی ممکن است توهم تهدید داشته باشد. و همان‌طور که در بالا مطرح شد، دلایلی وجود دارد که چرا هوش مصنوعی ممکن است عمداً برای رسیدن به یک هدف محلی (مرتبط با منافع محدود خود) دروغ بگوید.

به این مشکل در زمینه نظامی فکر کنید. ارتش ایالات متحده یک نهاد سلسله‌مراتبی است که پایبندی به دستورالعمل‌های بالا را اجباری می‌کند. دستورات، دستورالعمل‌ها هستند. اگر پرسنل نظامی از آنها سرپیچی کنند، با مجازات روبرو می‌شوند. پرسنل نظامی که مسئول استفاده از هوش مصنوعی Agentic هستند، چگونه قرار است خروجی هوش مصنوعی را تفسیر کنند، پس از اینکه به آنها گفته می‌شود که همیشه در مورد حقایق درست است و وقتی به عنوان یک سلاح قدرتمند استفاده می‌شود. از پرسنل نظامی خواسته می‌شود از یک فناوری خطرناک استفاده کنند که مافوق‌های آنها آن را نمی‌فهمند. به هر حال، هیچ کس که با الگوریتم‌های بهینه‌سازی کار کرده باشد، از اولویت ریاضیات در هوش مصنوعی Agentic تعجب نمی‌کند. و این ریاضیات است که ترامپ و اطرافیان‌ش آن را درک نمی‌کنند. ترامپ فرض می‌کند که آنچه او و پنتاگون می‌خواهند هوش مصنوعی انجام دهد، همان کاری است که هوش مصنوعی Agentic انجام خواهد داد. این از نظر مفهومی با خیال‌پردازی ارتش ربات‌ها در مورد سربازانی که بدون تغییر از دستورات پیروی می‌کنند، مرتبط است.

اما این کاری نیست که هوش مصنوعی Agentic انجام می‌دهد. یا بهتر است بگوییم تا حدی که این کار را انجام می‌دهد، این کار را به روش Genie انجام می‌دهد. اگر به یک مدل هوش مصنوعی یک تابع پاداش با وزندهی بالا داده شود تا افت فشار خط‌لوله را به حداقل برساند و نشستی آب را از بین ببرد، راحل ریاضی‌کاملی که کشف می‌کند می‌تواند این باشد که شیرهای اصلی توزیع را به طور کامل ببندد. هوش مصنوعی با موفقیت نشستی‌ها را به صفر می‌رساند و محدودیت ریاضی خود را برآورده می‌کند، در حالی که سهواً آب کل یک شهر را قطع می‌کند. -

ربات چت گوگل جمینی.

افسانه‌ی جن، دربارهی یک اعطاکننده‌ی آرزو است که آرزوها را به معنای واقعی کلمه تفسیر می‌کند. برای مثال، ممکن است کسی آرزوی ثروتمند شدن داشته باشد. جن آرزوی

او را برآورده می‌کند و فرزندش می‌میرد، اما دعوی حقوقی ناشی از آن، آرزوکننده را ثروتمند می‌کند. آرزو برآورده می‌شود، اما با هزینه‌ای که آرزوکننده پیش‌بینی نکرده بود. این ریاضیات بهینه‌سازی است. عوامل هوش مصنوعی در صورت امکان، وظیفه‌ی مورد نظر را انجام می‌دهند. اما ممکن است برای جلوگیری از نشت شیر آب، آب کل منطقه‌ی کشور را قطع کنند.

پیامدهای ناخواسته توسط فرآیند بهینه‌سازی تضمین می‌شوند. با توجه به اینکه عوامل هوش مصنوعی نسبت به پیامدهای اجتماعی اقدامات خود بی‌توجه هستند، تفاوت بین بستن زیرساخت آب کل یک منطقه و بستن شیر آب نشستی در مجموعه‌ی دغدغه‌های عامل در نظر گرفته نمی‌شود. آنچه موضوع را تعیین می‌کند پاداش است. و تفاوت در هزینه‌های اجتماعی یک متغیر در ساختار پاداش نیست. می‌توان آن را به عنوان یک محدودیت در نظر گرفت. اما با در نظر گرفتن تعداد نامحدودی از احتمالات، احتمال اینکه از یک اقدام نامطلوب خاص از طریق محدودیت‌ها جلوگیری شود، بسیار کم است. این بخشی از مشکل ایجاد توهم هوش بدون هیچ هوش واقعی پشت آن است. اکثر مردم تا زمانی که فاجعه رخ ندهد، تفاوت را درک نمی‌کنند. این یک مسئله است وقتی موضوع مورد بحث دستور پخت مرغ کاجیاتوره باشد و کاملاً متفاوت است وقتی زیرساخت‌های فیزیکی است که ملت را سرپا نگه می‌دارد. ظاهراً منبع آب ایالات متحده هم به شدت در معرض خطر است و هم به اینترنت متصل است به گونه‌ای که ربات‌های هوش مصنوعی عامل «پاداش هک» یک خطر هستند.

تئاتر کابوکی در شیوه سیاست آمریکایی نقش محوری دارد. مردم بلند می‌شوند و چیزهایی می‌گویند که نه خودشان و نه هیچ کس دیگری به آن اعتقاد ندارند که افکار عمومی را دستکاری کند یا انسجام قبیله‌ای ایجاد کند. هوش مصنوعی در یک کلام «فوق‌العاده هوشمند» است اما در واقع نمی‌تواند تا ۱۰۰ بشمارد. ادعاهایی مطرح می‌شود مبنی بر اینکه هوش مصنوعی عامل می‌تواند هم ایمن و هم مؤثر باشد، در حالی که شواهد نشان می‌دهد که یا ایمن است یا مؤثر. و «کند کردن» توسعه هوش مصنوعی، زمان لازم برای رفع مشکلات ایمنی را فراهم نمی‌کند زیرا محاسبات نشان می‌دهد که آنها غیرقابل اصلاح هستند. با توجه به شرایط موجود، انتخاب بین این نیست که پسر عمومی معتاد به شیشه‌تان یا عمه ترسویان سالی، پنتاگون (هوش مصنوعی عامل) را اداره کنند. هوش مصنوعی ایمن مؤثر نخواهد بود و هوش مصنوعی مؤثر نیز ایمن نخواهد بود. باور اینکه این پاسخ آمریکایی به دیپلماسی است دشوار است، اما هست. می‌توان آن را به همسایه‌ام تشبیه کرد که با یک برگ‌روب، چهار ساعت وقت صرف پاکسازی حیاطی می‌کند که من در بیست دقیقه آن را جمع کرده‌ام. تلاش ارتش ایالات متحده برای سربازان مطیع، این --- تلاش پنتاگون برای اعطای اختیار مطلق به هوش مصنوعی را طعنه‌آمیز می‌کند. یک گول آرزوها را برآورده می‌کند. خدا به کسانی که آنها را می‌سازند کمک کند.

سطری چند درمورد نویسنده این مقاله :

راب یوری هنرمند و اقتصاددان سیاسی است. کتاب او با عنوان «اقتصاد زن»
توسط انتشارات کانترپانچ منتشر شده است

----- با تقدیم احترام «2026-10-04»

.....