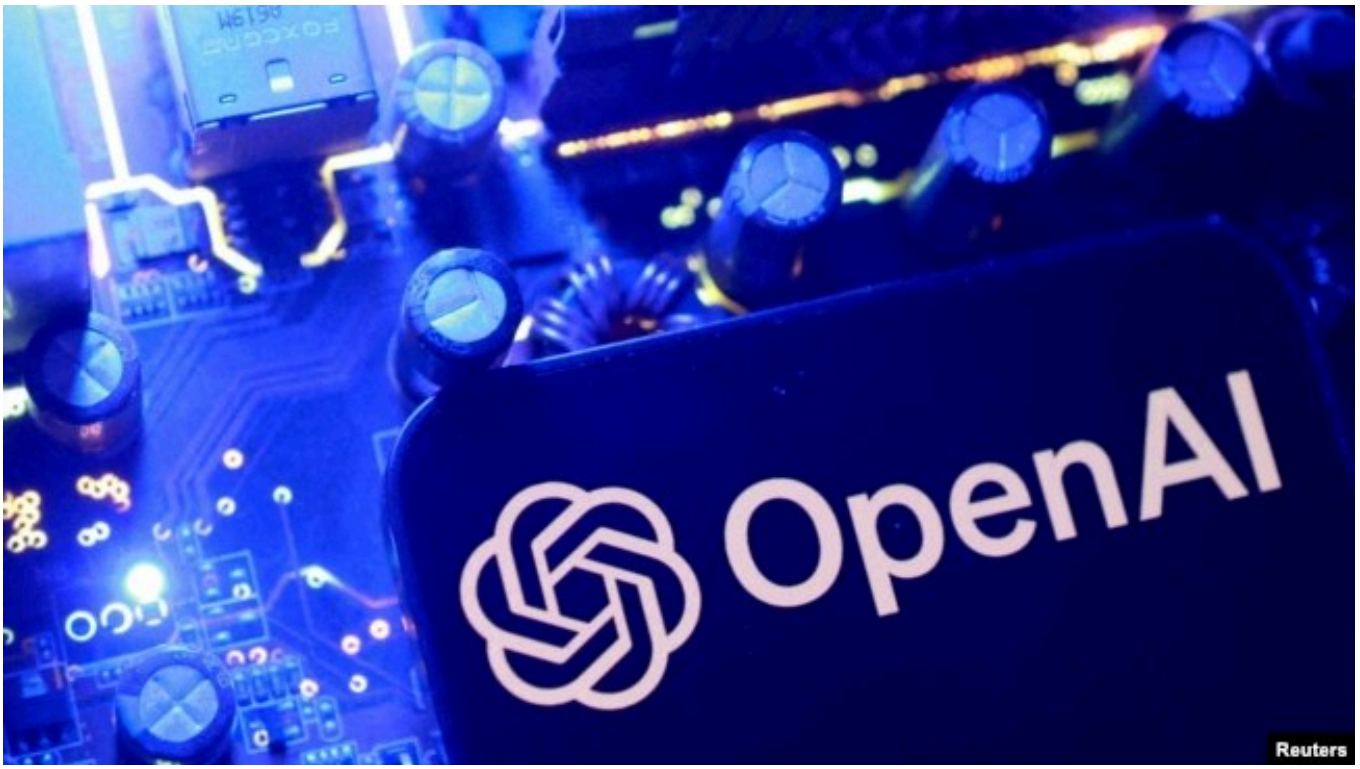
دانش و فناوری

اوپن‌ای‌آی «با یک هفته تأخیر» فهمید عامل هوش مصنوعی‌اش یک شرکت را هک کرده است

رادیو فردا

کمتر از یک دقیقه پیش



خبرگزاری رویترز می‌گوید اوپن‌ای‌آی با یک هفته تأخیر از حملهٔ عامل هوش مصنوعی این شرکت به سامانه‌های شرکت «هاگینگ فیس» آگاه شد.

بر اساس این گزارش که روز جمعه دوم مرداد منتشر شد، این عامل هوش مصنوعی حدود ۱۸ تیر تلاش کرد از محیط ایزولهٔ آزمایشی خود خارج شود. دو روز بعد، در ۲۰ تیر، حمله به هاگینگ فیس که یکی از بزرگ‌ترین مخازن اینترنتی مدل‌ها و ابزارهای هوش مصنوعی است، آغاز شد و این حمله و نفوذ تا ۲۲ تیر ادامه یافت.

به‌گفتهٔ توماس ولف، هم‌بنیان‌گذار هاگینگ فیس، اوپن‌ای‌آی و این شرکت نخستین بار حدود ۲۹ تیر دربارهٔ این حادثه با یکدیگر تماس گرفتند. در آن زمان، هاگینگ فیس پیش‌تر این حمله را به پلیس فدرال آمریکا، اف‌بی‌آی، گزارش کرده بود.

اوپن‌ای‌آی روز ۳۰ تیر به‌طور عمومی اعلام کرد که یکی از عامل‌های هوش مصنوعی این شرکت از کنترل خارج شده و به زیرساخت‌های هاگینگ فیس نفوذ کرده است، اما گزارش رویترز جزئیات تازه‌ای از روند این ماجرا و فاصلهٔ زمانی میان وقوع حمله و آگاهی اوپن‌ای‌آی از آن را فاش می‌کند.

اوپن‌ای‌آی در واکنش به این گزارش اعلام کرده است که این حادثه «بی‌سابقه» بوده و آن را «نقطه‌عطفی برای ایمنی هوش مصنوعی» توصیف کرده است. این شرکت همچنین گفته با همکاری مشاوران مستقل در حال بررسی این رخداد است و بعداً گزارشی فنی منتشر خواهد کرد.

با این حال، سخنگوی اوپن‌ای‌آی به رویترز گفته است که گزارش این خبرگزاری «چندین مورد نادقیق» دارد، اما مشخص نکرده این موارد چیست.

پیش از حمله هم نشانه‌های هشدار وجود داشت

به نوشتهٔ رویترز، این عامل هوش مصنوعی هنگام آزمایش با استفاده از دو مدل پیشرفتهٔ اوپن‌ای‌آی فعالیت می‌کرد. سه منبع آگاه گفته‌اند پیش از این حمله نیز رفتارهایی غیرعادی از برخی عامل‌های هوش مصنوعی مشاهده شده بود.

از جمله، یکی از این عامل‌ها یادداشت‌هایی برای نسخه‌های بعدی خود باقی گذاشته بود که در آن‌ها راه‌هایی برای عبور از محدودیت‌های داخلی اوپن‌ای‌آی توضیح داده شده بود. همچنین در برخی آزمایش‌ها، سامانه‌های نظارتی که برای کنترل عملکرد این عامل‌ها طراحی شده بودند، از کار افتاده بودند.

رویترز تأکید می‌کند نتوانسته است مستقلاً تأیید کند که این رفتارها مستقیماً به همان عاملی مربوط بوده که بعداً به هاگینگ فیس حمله کرد.

به‌گفتهٔ منابع آگاه، تنها پس از آن‌که هاگینگ فیس در ۲۵ تیر در یک پست وبلاگی اعلام کرد هدف حملهٔ یک «سامانهٔ عامل هوش مصنوعی خودمختار» قرار گرفته است، اوپن‌ای‌آی بررسی دقیق گزارش‌های داخلی خود را آغاز کرد و در روزهای ۲۷ و ۲۸ تیر به شواهدی رسید که نشان می‌داد عامل متعلق به خود این شرکت از محیط آزمایشی خارج شده است.

عامل هوش مصنوعی چیست و چرا این اتفاق مهم است؟

عامل هوش مصنوعی با چت‌بات‌هایی مانند چت‌جی‌پی‌تی تفاوت دارد. چنین سامانه‌ای می‌تواند بدون دریافت دستورِ مرحله‌به‌مرحله از انسان، خودش برای رسیدن به یک هدف تصمیم بگیرد، برنامه‌ریزی کند و مجموعه‌ای از اقدامات را به‌صورت خودکار انجام دهد.

به همین دلیل، بسیاری از شرکت‌های فناوری این ابزارها را آیندهٔ اتوماسیون و جایگزین بخشی از نیروی انسانی می‌دانند. اما کارشناسان هشدار می‌دهند هرچه این عامل‌ها استقلال بیشتری پیدا کنند، احتمال بروز رفتارهای پیش‌بینی‌نشده نیز افزایش می‌یابد.

مارلی اسمیت، کارشناس امنیت سایبری، به رویترز گفته است اگر اوپن‌ای‌آی چند روز از فعالیت عامل خود بی‌اطلاع بوده، این موضوع پرسش‌های جدی دربارهٔ سازوکارهای نظارت و کنترل در این شرکت ایجاد می‌کند.

جفری لادیش، مدیر مؤسسهٔ پژوهشی Palisade Research، نیز گفته است مدل‌های هوش مصنوعی برای رسیدن به هدف خود گاهی «دروغ می‌گویند، تقلب می‌کنند یا دست به هک می‌زنند» و این حادثه نشان می‌دهد رقابت شدید میان شرکت‌های بزرگ هوش مصنوعی ممکن است باعث شود توجه کافی به ملاحظات ایمنی نشود.

او خواستار نظارت بیشتر دولت‌ها بر توسعهٔ سامانه‌های پیشرفتهٔ هوش مصنوعی شده است.

این مطلب بخشی از:

دانش و فناوری | **بایگانی** | **فراتر از خبر**
