گزارش «شهروند» از داده‌هایی که هوش مصنوعی را «مسموم» می‌کنند

بمب ساعتی در قلب هوش مصنوعی

[لیلا مهداد]

داده‌های مسموم به عنوان یکی از بزرگ‌ترین تهدیدها برای امنیت هوش‌مصنوعی و توسعه سیستم‌های هوش‌مصنوعی ایمن شناخته می‌شوند. این پدیده با توجه به رشد فزآینده مدل‌های مبتنی بر یادگیری عمیق و افزایش وابستگی صنایع به هوش‌مصنوعی به موضوعی حیاتی تبدیل شده است. به عنوان نمونه در یک سیستم تشخیص تومورهای سرطان پستان که براساس داده‌های تاریخی آموزش‌دیده بود، تزریق ۰٫۱درصد داده مسموم (با برچسب‌های عمدا معکوس) منجر به کاهش ۴۰درصد دقت در شناسایی تومورهای مرحله اول، افزایش ۳۰۰درصدی موارد مثبت کاذب و هزینه مالی بالغ بر ۲میلیارد دلار برای اصلاح سیستم شده است، بنابراین مقابله با این پدیده نیازمند ترکیبی از فناوری‌های پیشرفته، قوانین شفاف و آگاهی عمومی است. توسعه‌دهندگان باید همواره اصل اعتماد صفر (Zero Trust) را در فرآیند آموزش مدل‌ها اعمال کنند.

حملات سمّ‌پاشی داده خطر مهمی برای هوش مصنوعی است. اما نه به معنای ناامن بودن مطلق آن. همان بحث و مثال قدیمی که با تیغ جراحی می‌توان کسی را از مرگ نجات داد یا کسی را کشت. مثل این می‌ماند که بگوییم اگر تیغ جراحی زنگ زده باشد، بدن بیمار عفونت می‌کند و می‌میرد. خب با تیغ زنگ زده جراحی نمی‌کنیم. منظورم این است که چارچوب‌هایی برای مهندسی ایمن داده وجود دارد. آزمون‌های منظم و فرهنگ‌سازی امنیتی. با انجام این کارها می‌توان سامانه‌های هوش مصنوعی را برابر تهدیدهای موجود مقاوم کرد

ایجاد اختلال در فرآیند یادگیری مدل‌های هوش‌مصنوعی

در حوزه هوش‌مصنوعی، داده‌های مسموم (Poisoned Data) به داده‌هایی اطلاق می‌شود که عمداً یا سهواً برای ایجاد اختلال در فرآیند یادگیری مدل‌های هوش‌مصنوعی طراحی شده‌اند. این داده‌ها می‌توانند باعث ایجاد سوگیری، کاهش دقت یا حتی رفتارهای خطرناک در سیستم‌های هوش مصنوعی شوند. دستکاری برچسب‌ها (Label Flipping) یکی از انواع داده‌های مصنوعی است؛ یعنی تغییر عمدی برچسب‌های داده‌های آموزشی برای گمراه کردن مدل. برخی از تزریق نمونه‌های مخرب (Adversarial Examples) بهره می‌برند، یعنی افزودن داده‌هایی که توسط انسان‌ها به‌راحتی تشخیص داده می‌شوند، اما مدل را فریب می‌دهند (مثلاً تصاویر نویزدار). حملات بک‌دور (Backdoor Attacks)هم یکی دیگر از انواع داده‌های مصنوعی است. به معنی آموزش مدل برای پاسخ‌دهی خاص به محرک‌های پنهان (مثلاً شناسایی چهره تنها در صورت وجود یک نشانه مخفی).

اشتیاق بی‌وقفه، سریع‌ترین راه نفوذ «سم» به آن

اشتیاق بی‌وقفه و سیری‌ناپذیر برای کسب داده‌های بیشتر می‌تواند نقص مهلک هوش‌مصنوعی یا دستکم سریع‌ترین راه برای نفوذ «سم» به آن باشد. به عنوان مثال مهاجمان با تزریق داده‌های جعلی مثلاً تصاویر نادرست برچسب‌گذاری شده به مجموعه‌های آموزشی، مدل‌های هوش‌مصنوعی را فریب می‌دهند. به عنوان مثال تغییر ۰٫۱درصد از داده‌های آموزشی می‌تواند دقت تشخیص چهره را تا ۳۰درصد کاهش دهد.

به نقل از دیجیتال‌ترندز، اکثر سیستم‌های هوش مصنوعی که امروزه با آنها مواجه می‌شویم از چت جی‌پی‌تی گرفته تا توصیه‌های شخصی‌سازی‌شده نتفلیکس فقط به دلیل حجم گسترده متن، تصویر، گفتار و سایر داده‌هایی که بر اساس آنها آموزش دیده‌اند، به اندازه کافی «هوشمند» هستند که چنین شاهکارهای چشمگیری را انجام دهند. اگر این گنجینه غنی آلوده شود، رفتار مدل می‌تواند دچار مشکل شود. پیامدهای چنین اتفاقی در دنیای واقعی بسیار فراتر از یک چت‌بات است که بی‌هدف حرف می‌زند یا مولدهای متن به تصویری که وقتی از آنها خواسته می‌شود پرنده‌ای را نشان دهند، تصویر یک هواپیما را تولید می‌کنند. گروه‌هایی از افراد سودجو می‌توانند به‌طور بالقوه باعث شوند که یک ماشین خودران چراغ قرمز را نادیده بگیرد، یا در مقیاس بسیار بزرگ‌تر، باعث اختلال و قطعی شبکه برق شوند.

دو فناوری یادگیری فدرال و بلاکچین برای کاهش آسیب‌پذیری مدل‌های AI

مطالعه اخیر محققان دانشگاه فلوریدا به رهبری پروفسور هادی امینی و اروین مور، مکانیسمی نوین برای مقابله با حملات مسمومیت داده (Data Poisoning) در سیستم‌های هوش مصنوعی ارائه کرده است. این تحقیق ترکیبی از دو فناوری یادگیری فدرال (Federated Learning) و بلاکچین را به‌کار می‌گیرد تا آسیب‌پذیری مدل‌های AI را کاهش دهد.

در روش یادگیری فدرال برای آموزش هوش مصنوعی، از یک نسخه کوچک از یک مدل آموزشی استفاده می‌شود که مستقیما روی دستگاه شما آموزش می‌بیند و فقط به‌روزرسانی‌ها و نه داده‌های شخصی کاربر را با مدل جهانی روی سرور یک شرکت به اشتراک می‌گذارد. این روش اگرچه که حریم خصوصی را حفظ می‌کند، اما همچنان در برابر حملات داده‌های مسموم، آسیب‌پذیر است. در فناوری بلاکچین که به دلیل نقشش در ارزهای دیجیتال مانند بیت‌کوین محبوبیت پیدا کرده، یک پایگاه داده مشترک است که در شبکه‌ای از رایانه‌ها توزیع شده است. داده‌ها در بلوک‌هایی که به ترتیب زمانی روی یک زنجیره به هم متصل شده‌اند، ذخیره می‌شوند. هر کدام اثر انگشت خود و همچنین اثر انگشت بلوک قبلی را دارند که عملا آن را ضد دستکاری می‌کند. در این فناوری کل زنجیره از یک ساختار خاص پیروی می‌کند.

این مانند یک فرآیند بررسی است تا اطمینان حاصل شود که بلوک‌های تصادفی اضافه نمی‌شوند. می‌توان آن را مانند یک چک لیست برای پذیرش داده در نظر گرفت.

درواقع محققان هنگام ساخت مدل خود از این قابلیت به نفع خود استفاده کردند. این مدل به‌روزرسانی‌های بلوک را مقایسه و محاسبه می‌کرد که آیا به‌روزرسانی‌ها به طور بالقوه سمی هستند یا خیر. به‌روزرسانی‌های بالقوه سمی ثبت شده و سپس از تجمیع شبکه حذف می‌شدند.

حملات سمّ‌پاشی داده خطر مهمی برای هوش مصنوعی است

به گفته شاکر خودروهای ما را هم امروز همین هوش مصنوعی به حرکت درآورده است و می‌راند. تفسیر نادرست علائم ترافیکی منجر به تصادف می‌شود و کاهش اعتماد به ماشین باعث تصویب مقررات سختگیرانه علیه هوش مصنوعی خواهد شد. این امر، حمل و نقل و لجستیک را به چالش می‌کشد. این پژوهشگر هوش مصنوعی بیان می‌کند: «ما داریم از شرایط گسترش اطلاعات نادرست، زندگی می‌کنیم و اخبار جعلی به سادگی و با نفوذ بالا مشغول گسترش‌اند.»

او در پاسخ به این سوال که چنین اتفاقی در ایران رخ داده یا خیر می‌گوید: «به عنوان یک پژوهشگر در جایگاهی نیستم که بتوانم به این پرسش، پاسخ دهم. اما تا آنجایی که منابع خبری را دنبال کرده‌ام ایران متاسفانه هدف بسیاری از حملات سایبری بوده است. اما اینکه آیا حملات سمّ‌پاشی داده (Data Poisoning) روی سیستم‌های هوش مصنوعی ایران داشتیم یا نه را نمی‌دانم و باید از کسانی بپرسید که در این زمینه به اطلاعات طبقه‌بندی شده دسترسی دارند.» بنابرنظر شاکر حملات سمّ‌پاشی داده خطر مهمی برای هوش مصنوعی است. اما نه به معنای ناامن بودن مطلق آن. همان بحث و مثال قدیمی که با تیغ جراحی می‌توان کسی را از مرگ نجات داد یا کسی را کشت. مثل این می‌ماند که بگوییم اگر تیغ جراحی زنگ زده باشد، بدن بیمار عفونت می‌کند و می‌میرد. خب با تیغ زنگ زده جراحی نمی‌کنیم. منظورم این است که چارچوب‌هایی برای مهندسی ایمن داده وجود دارد. آزمون‌های منظم و فرهنگ‌سازی امنیتی. با انجام این کارها می‌توان سامانه‌های هوش مصنوعی را برابر تهدیدهای موجود مقاوم کرد.

تجربه دیگران

تای (Tay) مایکروسافت (۲۰۱۶) در آمریکا: مکانیسم آن به این شکل بود که ربات چت با یادگیری از تعاملات کاربران در توییتر به سرعت از کاربران عبارات نژادپرستانه و توهین‌آمیز آموخت. نتیجه آن این بود که تنها در ۲۴ ساعت، تای بیش از ۵۰ هزار توئیت توهین‌آمیز منتشر کرد و مایکروسافت مجبور به خاموشی آن شد.

حمله به سیستم تشخیص چهرهٔ Clearview AI (۲۰۲۲) اروپا: هکرها با آپلود هزاران تصویر جعلی حاوی الگوهای نوری خاص، دقت تشخیص چهره را ۴۰درصد کاهش دادند. نتیجه این شد که این شرکت در ۱۲کشور اروپایی جریمه شد.

سیستم تشخیص بیماری‌های پوستی در هند (۲۰۲۳): مشکل این بود که تمرکز داده‌ها بر پوست روشن باعث شد دقت تشخیص ملانوما در بیماران با پوست تیره تنها ۳۴درصد باشد. نتیجه آن این شد که ۲۷مورد تشخیص اشتباه منجر به مرگ شد.

کلاهبرداری با هوش مصنوعی در چین (۲۰۲۴): مکانیسم آن به این شکل بود که تولید ۱٫۲میلیون پرونده اعتباری جعلی با الگوریتم‌های GAN برای فریب سیستم‌های اعتبارسنجی استفاده شد. این مساله ۴۵۰میلیون دلار ضرر به بانک‌ها وارد کرد.

جنگ سایبری روسیه-اوکراین (۲۰۲۳): در این نمونه تاکتیک تزریق داده‌های ترافیک جعلی به سیستم‌های ناوبری خودروهای خودران نظامی بود. نتیجه هم این شد که ۱۵دستگاه خودروی خودران در خط مقدم از مسیر منحرف شدند.

پروژهٔ ابرشهر هوشمند عربستان (Neom): مشکل به این شکل بود که داده‌های حسگرهای محیطی آلوده به نویز هدفمند، باعث خطا در پیش‌بینی مصرف انرژی شد. این مساله تاخیر ۸ ماهه پروژه با ۲۰۰ میلیون دلار ضرر مالی بود.

جدول مقایسه‌ای اثرات جهانی

منطقه	حوزه	درصد آلودگی داده	کاهش دقت مدل
آمریکای شمالی	تشخیص چهره	۰٫۸	۳۷درصد
اروپا	سیستم‌های مالی	۰٫۴	۲۹درصد
آسیا	سلامت	۱٫۲	۴۵درصد
خاورمیانه	شهرهای هوشمند	۰٫۹	۳۳درصد

آنچه اکنون هوش مصنوعی را خطرناک‌تر می‌کند تصورات اسطوره‌ای از آن است

این پژوهشگر هوش مصنوعی درخصوص مشکلات ناشی از ناامن بودن هوش مصنوعی است، می‌گوید: «باید بگوییم که در این رابطه باید مدام برای واکنش‌های متعادل آماده باشیم. نه خود را تقدیم به ارزش‌های شرکتی هوش مصنوعی کنیم و نه اینکه از آن بترسیم و بخواهیم هوش مصنوعی را به زندان بیندازیم.» او بیان می‌کند: «به‌تبع آن ماجراهایی که تهدید محسوب می‌شوند، برای رسانه جذاب‌تر است. هوش مصنوعی حتی اگر نتواند انسان را منقرض کند، سوژه‌های خوبی برای «اقتصاد توجه» در جهان پدید آورده است. اینجاست که رسانه‌ها نیز شده‌اند میدان نمایش شکاف میان «بازاریاب‌های هوش مصنوعی» و «مخالفان هوش مصنوعی».» شاکر ادامه می‌دهد: «یعنی بستری فراهم شده که در آن کلی اطلاعات درست و غلط و اغراق‌شده در رابطه با توانمندی‌ها، فواید و خطرهای هوش مصنوعی وجود دارد.» این پژوهشگر هوش مصنوعی اظهار می‌کند: «در رابطه با آینده در درجه نخست باید با شفافیت بیشتری به داده‌های موجود نگاه کنیم. از متخصصان ایمنی شبکه کمک بخواهیم و همه این‌ها را ختم کنیم به ارتقای سواد مواجهه با هوش مصنوعی به ویژه در مواجهه با اخبار رسانه‌ای.» به باور او آنچه اکنون هوش مصنوعی را خطرناک‌تر می‌کند، تصورات اشتباه، ساختگی و اسطوره‌ای از آن است. این ترس بستری فراهم می‌کند برای سم‌پاشی داده و دستکاری روند هوشمندسازی زندگی روزمره.

عده‌ای می‌توانند با شیطنت‌هایی به ظاهر پیش‌پا افتاده نظم در حال شکل‌گیری کنونی را مخدوش کنند

علی شاکر، پژوهشگر هوش مصنوعی به «شهروند» می‌گوید: «در روزگاری زندگی می‌کنیم که عده‌ای می‌توانند با شیطنت‌هایی به ظاهر پیش‌پا افتاده نظم در حال شکل‌گیری کنونی را کاملاً مخدوش کنند. مهاجمان سایبری اکنون دستکاری در نحوه آموزش هوش مصنوعی را هدف قرار داده‌اند. جوری که ماشین یا پاسخ‌های نادرست بدهد یا حتی واکنش‌های خطرناک داشته باشد.»

شاکر ادامه می‌دهد: «باید توجه کنیم که در اینجا با موضوع‌های کمتر خطرناکی مثل میم‌ها و جعل عمیق‌های بامزه طرف نیستیم، داده‌های مسمومی که هکرها به ماشین می‌دهد، می‌تواند منجر به تشخیص غلط بیماری‌ها شود. با توجه به اطمینانی که اکنون به ماشین‌های هوشمند در عرصه پزشکی داریم، این دستکاری‌های مخرب ممکن است مثلاً منجر به نادیدگرفتن تومور در بدن بیمار شود. ممکن است ماشین توصیه درمانی و دارویی اشتباه کند.» این پژوهشگر هوش مصنوعی اظهار می‌کند: «موضوع فقط جان نیست. مال هم هست. دستکاری‌های مخرب ممکن است باعث شود که ماشین نتواند تراکنش‌های تقلبی را پرچم‌گذاری کند و باعث زیان‌های مالی شود یا به اشتباه تراکنش‌های قانونی را مسدود کند و اعتماد و عملیات کاربران را مختل سازد.» او عنوان می‌کند: «متخصصان بازارهای مالی می‌دانند که الگوریتم‌های معاملاتی تغییریافته می‌توانند معاملات اشتباهی را اجرا و بازارها را بی‌ثبات کنند و زمینه سوءاستفاده را به وجود بیاورند.»