ویدیو تولیدشده با هوش مصنوعی در درگیری ایران-اسرائیل

2026-004 FB-UA

Summary

در بررسی گسترش محتوای تولیدشده توسط هوش مصنوعی در درگیری‌های مسلحانه، در یک پرونده مربوط به جنگ اسرائیل و ایران در سال ۲۰۲۵، هیئت نظارت از متا می‌خواهد اقدامات بیشتری انجام دهد تا کاربران بتوانند چنین محتوایی را شناسایی کنند. رویکرد این است که بالا آمدن محتوای تولیدشده با هوش مصنوعی باید استنتاج شود. این امر شامل ارائه جزئیات دردسترس درباره منبع رسانه براساس محتوای استانداردهای سرمنشأ، بازرسی با ابزارهای شناسایی قوی‌تر، و توسعه روش‌های بهتر برای برچسب‌زنی‌های مناسب‌تر است. Meta باید مجموعه قوانین جدید و جداگانه‌ای تنظیم کند تا مطمئن شود کاربران می‌توانند بااطمینان محتوای تولیدشده با هوش مصنوعی را تشخیص دهند. به‌علاوه، باید خطمشی‌های کنونی اش را تصحیح کند تا مطمئن شود به برونداد گمراه‌کننده تولیدشده با هوش مصنوعی پاسخ به‌موقع و مناسب می‌دهد.

شرکت باید به تعهدات عمومی خود پایبند باشد و از ابزارهای خود و دیگرانی که در سراسر این صنعت دردسترس هستند استفاده کند تا بتواند به‌طور مؤثر محتوای تولیدشده با هوش مصنوعی را که در پلتفرم‌ها منتشر می‌شوند شناسایی کند

هیئت نظارت تصمیم متا را برای باقی گذاشتن این پست بدون برچسب «هوش مصنوعی با ریسک بالا» در این مورد لغو می‌کند.

چرا این مسئله مهم است

همچنان که کیفیت و کمیت محتوای تولیدشده با هوش مصنوعی بالا می‌رود، تأثیر آن بر مردم و جوامع عمیق‌تر می‌شود. وقتی این خطرات بیشتر می‌شوند که خروجی آن‌ها باعث گمراهی، دستکاری یا افزایش تعامل در اشتراک‌گذاری درطول نزاع‌ها و بحران‌هایی مثل ایران یا ونزوئلا در ۲۰۲۶ شود، این محتوا به سرعت در پلاتفرم‌های شرکت‌های مختلف منتشر می‌شوند. درطول این دو بحران، ادعاهایی بود که معتقد بود محتوای گمراه‌کننده تولیدشده با هوش مصنوعی واقعی است و محتوای واقعی جعلی است. این امر توانایی تشخیص حقیقت توسط عموم را کم می‌کرد، سهم دروغ‌را نمادین می‌کرد و منجر به بی‌اعتمادی عمومی به همه اطلاعات می‌شد. در سال‌های اخیر، پویش‌های تأثیرگذاری با نیروی هوش مصنوعی به چالشی تبدیل شده است که در همه جا دیده می‌شود و در اکوسیستم اینترنت و رسانه‌های محدودی که در آن‌ها اطلاعات معتبر محدود است تشدید شده است. با این حال، گمراه‌کننده بودن برونداد‌های تولیدشده با هوش مصنوعی به‌تنهایی دلیل معتبری برای محدود کردن آزادی بیان نیست. این صنعت باید به‌طور

یکپارچه‌ای به کاربران کمک کند تا بتوانند محتوای تولیدشده با هوش مصنوعی را تشخیص دهند و پلاتفرم‌ها باید حساب‌های کاربری سوءاستفاده کننده و صفحه‌های اشتراک‌گذاری‌کننده این نوع بروندادها را شناسایی کنند.

درباره این پرونده‌ها

جنگ ایران-اسرائیل در ژوئن ۲۰۲۵ نشان‌دهنده نقطه عطفی بود برایحضورمحتوای گمراه‌کننده تولیدشده با هوش مصنوعی در رسانه‌های اجتماعی، و معروف شد به «جنگ نرم». چنین بروندادهای گمراه‌کننده‌ای بازدیدهای بسیار زیادی هم دریافت کرد و هر دو دولت ایران و اسرائیل برای تأثیرگذاری مبتنی بر هوش مصنوعی مورد اتهام قرار گرفتند. در ۱۵ ژوئن، دو روز پس از شروع جنگ دوازده روزه ایران و اسرائیل، ویدیویی در صفحه فیس‌بوک پست شد که ادعا داشت منبع خبری است. کاربر پست‌گذار در فیلیپین بود. ویدیو ویرانی گسترده‌ای از ساختمان‌ها را نشان می‌داد که متنی انگلیسی روی آن بود که می‌گفت «اکنون، زنده - حیفا درحال سقوط» [sic] با تاریخ پست. این ویدیو بسیار شبیه به ویدیویی بود که از TickTok سرچشمه گرفته بود و یک حقیقت‌یاب مستقل (Agence France-Presse) آن را به‌عنوان جعلی و تولیدشده با هوش مصنوعی شناسایی کرده بود. شرح پست فیس‌بوک، عبارت‌هایی به سبک عناوین خبری را که به این درگیری و اصطلاحات و هشتگ‌های نامرتبط با آن پیوند داشت را فهرست کرده بود. این پست بیش از ۷۰۰,۰۰۰ بازدید گرفته بود و نظرات بسیاری داشت که هیچکدام نمی‌گفت که این تولیدشده با هوش مصنوعی است.

شش کاربر این مورد را به Meta گزارش کردند ولی هیچ کدام از این گزارش‌ها توسط شرکت یا حقیقت‌یاب طرف سوم بازبینی نشد. کاربری این مورد را برای بازنگری به «هیئت» فرستاد. پس از اینکه «هیئت» این مورد را انتخاب کرد، Meta تأیید کرد که این پست «استاندارد انجمن برای اطلاعات نادرست» را نقض نکرده است زیرا این پست «مستقیماً در آسیب فیزیکی حتمی مشارکت نداشته است»، و نیازی به برچسب هوش مصنوعی ندارد.

نشانه‌های آشکاری از فریبکاری در این پست باعث شد تا «هیئت» از Meta بخواهد هویت و رفتار حساب‌های مرتبط با آن صفحه پرس‌وجو کند. درنتیجه، شرکت سه حساب مرتبط با صفحه را به‌دلیل سوءاستفاده از تعامل و اصیل نبودن غیرفعال کرد، صفحه را برداشت و همراه با آن، محتوای مربوطه را هم برداشت. صفحه برای درآمدزایی ازطریق برنامه ستاره‌های شرکت Meta فاقدشرایط شناخته شد.

یافته‌های کلیدی

«هیئت» دریافته است که محتوا خطر ملموس گمراه کردن عموم را در زمینه مسئله‌ای مهم در زمانی حساس داشته است بنابراین Meta باید برچسب «هوش مصنوعی با خطر بالا» را اعمال می‌کرده است. این پست به آستانه حذف به‌دلیل (مشارکت در خشونت یا خطر فیزیکی حتمی نداشته است) نرسیده است. Meta باید برای مقابله با گسترش

محتوای گمراه‌کننده تولیدشده با هوش مصنوعی در پلاتفرم‌هایش، ازجمله توسط شبکه‌ها و صفحه‌های غیراصیل یا سوءاستفاده‌کننده، و به ویژه در مسائل مربوط به منافع عمومی بیشتر کار کند تا کاربران بتوانند بین واقعیت و جعل تمیز قائل شوند.

نگرانی «هیئت» برای این گزارش‌ها اینجاست که Metaاستانداردهای «ائتلاف برای سرمنشا و اصالت محتوا» (C2PA) را حتی بر محتوای تولیدشده توسط ابزارهای هوش مصنوعی خود هم به طور ناسازگار اعمال کند و تنها بخشی از این برونداد‌ها برچسب‌گذاری مناسب دارند. استانداردهای C2PA مجموعه‌ای از استانداردهای فنی را ارائه می‌کند تا اطلاعات سرمنشا به عنوان فراداده‌هایی در محتوا تعبیه کند که اجازه می‌دهند پلاتفرم‌ها راحت‌تر محتوای تولیدشده با هوش مصنوعی را شناسایی کنند و برچسب‌هایی را برای آگاه‌سازی کاربران اعمال کنند.

مکانیسم‌های کنونی حتی برای برچسب‌گذاری استاندارد «اطلاعات هوش مصنوعی» به ویدیو (خوداظهاری کاربر یا ارجاع به تیم سیاست محتوا) هم نه قدرتمند و نه به اندازه کافی جامع هستند تا بتواند در مقیاس و سرعت تولید محتوای هوش مصنوعی با آن رقابت کنند، به ویژه در زمان بحران یا تنش که تعامل‌ها را در پلاتفرم افزایش می‌دهد. سیستمی که به شدت به خوداظهاری استفاده از هوش مصنوعی و بررسی‌های ارتقایافته (که به ندرت رخ می‌دهد) وابسته باشد، برای برچسب‌گذاری درست این برونداد نمی‌تواند به چالش‌هایی که در محیط کنونی رخ می‌دهد پاسخ دهد. برخی از اعضای «هیئت» همچنین اشاره کرده‌اند که برچسب‌های «هوش مصنوعی با خطر بالا» (برای برونداد‌هایی که می‌تواند افراد را در مسائل مهم گمراه کند) باید همراه با کاهش رتبه یا حذف از پیشنهادها باشد تا نگرانی‌های مربوط به گسترش تأثیر محتوای گمراه‌کننده را رفع کند.

رویکرد محدود Meta برای اعمال رتبه‌بندی‌ها به محتوای مشابه و تقریباً مشابه می‌تواند به این معنی باشد که برای این پست رتبه‌بندی‌های بررسی واقعیت انجام نشده است. محدودیت منابع و حجم قابل توجه برونداد باعث می‌شود کار حقیقت‌یاب‌ها برای اطمینان از بررسی به‌موقع تمام محتوای گمراه‌کننده دشوار شود، به‌ویژه در دوران نزاع یا بحران. «هیئت» تأکید دارد که Meta باید مطمئن باشد که حقیقت‌یاب‌ها به اندازه کافی منبع داشته باشند و دستورالعمل‌هایی برای اولویت‌بندی محتوای مناقشات در دسترس آن‌ها باشد. «پروتکل سیاست بحران» (CPP) و تشخیص «رویدادهای پرطرفدار» باید به Meta اجازه دهد تا در زمان بحران، پشتیبانی مؤثرتری از حقیقت‌یاب‌های طرف سوم ارائه دهد. گسترش دادن ارزیابی‌ها به دسته‌های وسیع‌تری از ویدیوهای بسیار مشابه (مثلاً با کاهش رتبه این موارد) می‌تواند به طور قابل توجهی از آسیب‌های احتمالی کم کند. این مورد، ناکارآمدی‌های رویکرد فعلی Meta را در زمان مناقشات نظامی برجسته می‌کند و نگرانی‌های قبلی «هیئت» را تشدید می‌کند.

با فعال شدن CPP، و اختصاص منابع اضافه، Meta خودش مستقلاً نشانه‌های واضح سوءاستفاده از تعاملات این صفحه را شناسایی نکرد و فقط در پاسخ به پرسش‌های «هیئت» به بررسی حساب‌های کاربری پشت این پست اقدام کرد. به‌جای اتکا به اقدامات تعدیلی مبتنی بر محتوا پس از انتشار و گسترش پست که می‌تواند نرخ شکست بالایی داشته باشد، اجرای دقیق سیاست‌های مبتنی بر رفتار می‌توانست آسیب‌های ناشی از این حساب‌های متخلف جلوگیری کند.

تصمیم نظارتی «هیئت»

این هیأت تصمیم متا را برای باقی گذاشتن این پست بدون برچسب «هوش مصنوعی با ریسک بالا» در این مورد لغو می‌کند.

«هیئت» توصیه می‌کند که «متا»:

- «استاندارد انجمن» برای محتوای تولیدشده با هوش مصنوعی تدوین شود که از «استاندارد انجمن برای اطلاعات گمراه‌کننده» مجزا بوده و شامل قوانین جامع درباره حفظ سرمنشا، پروتکل‌های برچسب‌گذاری هوش مصنوعی و خوداظهاری باشد.
- مسیرهایی برای برچسب‌گذاری «خطر بالا» و «هوش مصنوعی با خطر بالا» به محتواهایی که فراوانی بیشتری دارند ایجاد شود تا با استفاده از کانال‌های ارتقا با شفافیت بیشتر در سیستم‌های خودکار و بازبینی در مقیاس بالا اعمال شود به‌طوری که چنین برچسب‌گذاری‌هایی بتوانند با حجم بسیار بیشتری انجام شوند.
- در محتوای تولیدشده با ابزارهای هوش مصنوعی Meta، اطلاعات سرمنشا و واترمارک‌های نامرئی، ازجمله به کارگیری «اعتبارنامه‌های محتوا» (همان‌طور که در C2PA تعریف شده است) درج شود.
- «اعتبارنامه‌های محتوا» در مقیاس بزرگ اعمال شود و از اینکه این موارد در هرجا که اطلاعات سرمنشا در دسترس است به وضوح و به‌طور مداوم قابل مشاهده است اطمینان حاصل شود.
- در ابزارهای شناسایی قوی‌تری برای محتوای تولیدشده با هوش مصنوعی چند قالبی (صوتی، صوتی تصویری و تصویر) سرمایه‌گذاری شود.
- توضیحات واضحی درباره جریمه‌های خوداظهاری نکردن برای محتوای دیجیتالی تولیدشده یا تغییریافته، ازجمله معیارهایی برای جریمه‌ها و محدودیت‌های ناشی از آن‌ها، منتشر شود.
- «استاندارد انجمن برای اطلاعات گمراه‌کننده» اصلاح شود تا اطمینان حاصل شود که اطلاعات نادرستی که به‌طور مستقیم خطر خشونت فوری یا آسیب جسمانی را به‌دنبال دارد بررسی شود و این بررسی فقط به سیگنال‌های شرکای خارجی وابسته نباشد. ابزار CPP باید برای شناسایی به‌موقع و فعال چنین محتوای نقض‌کننده‌ای منابعی را اختصاص دهد و با تخصص داخلی و اقداماتی از قبیل برچسب‌زنی و تحقیق درباره حساب‌ها و صفحات ارسال‌کننده، آن را پشتیبانی کند.

*خلاصه‌های مورد ارائه‌دهنده نمای کلی موارد است و ارزش سابقه‌سازی ندارد.