ویدئوی تولید شده توسط هوش مصنوعی در جنگ اسرائیل و ایران

2026-004 FB-UA

Summary

در تحلیل انتشار محتوای تولید شده توسط هوش مصنوعی در درگیری های مسلحانه در موردی مربوط به جنگ اسرائیل و ایران در سال 2025، هیئت نظارت از Meta می‌خواهد اقدامات بیشتری انجام دهد تا کاربران بتوانند این نوع خروجی‌ها را شناسایی کنند. رویکرد آن برای نمایش محتوای تولید شده توسط هوش مصنوعی باید تکامل یابد. این شامل ارائه جزئیات گسترده درباره منشأ رسانه‌ها بر اساس استانداردهای منشأ محتوا، سرمایه گذاری در ابزارهای شناسایی قوی‌تر و توسعه روش‌های بهتر برای برچسب‌گذاری مناسب است. Meta باید مجموعه قوانین جدید و جداگانه ایجاد کند تا اطمینان حاصل شود کاربران می‌توانند محتوای تولید شده توسط هوش مصنوعی را به صورت قابل اعتماد تشخیص دهند. همچنین، باید سیاست‌های فعلی‌اش را اصلاح کند تا واکنش به‌موقع و کافی به خروجی‌های فریبنده تولید شده توسط هوش مصنوعی تضمین شود.

شرکت باید به تعهدات عمومی خود عمل کند و از ابزارهای خود و سایر ابزارهای موجود در این حوزه برای مقابله مؤثر با محتوای هوش مصنوعی مولد فریبنده که بین پلتفرم‌ها پخش می‌شود، استفاده کند.

«هیئت»، تصمیم Meta برای باقی گذاشتن این پست بدون زدن برچسب هوش مصنوعی پرخطر را لغو می‌کند.

چرا این موضوع اهمیت دارد

با افزایش کمیت و کیفیت محتوای تولید شده توسط هوش مصنوعی، تأثیر آن بر مردم و جوامع عمیق خواهد بود. خطرات زمانی افزایش می‌یابند که خروجی دیپ فیک که برای فریب، دستکاری یا افزایش مشارکت طراحی شده است، در طول جنگ‌ها و بحران‌هایی مانند ایران و ونزوئلا در سال 2026 به اشتراک گذاشته می‌شوند و به سرعت در پلتفرم‌های شرکت‌های مختلف منتشر می‌شوند. در طول این دو بحران، ادعاهایی مطرح شد مبنی بر این که محتوای فریبنده تولید شده توسط هوش مصنوعی واقعی است و محتوای واقعی ساختگی است. این موضوع ناتوانی انسان‌ها در تشخیص حقیقت را تشدید می‌کند، که نماد سود دروغین است و منجر به بی‌اعتمادی عمومی به همه اطلاعات می‌شود. کمپین های تأثیرگذاری مبتنی بر هوش مصنوعی چالشی رو به رشد هستند که در سال های اخیر در سطح جهانی دیده شده و در رسانه ها و اکوسیستم های محدودکننده اینترنتی که اطلاعات معتبر را محدود می‌کنند، تشدید شده است. با این حال، گمراه کننده بودن خروجی های تولید شده توسط هوش مصنوعی به خودی خود دلیل موجه محدود کردن آزادی بیان نیست. این صنعت به انسجام نیاز دارد تا به کاربران کمک کند محتوای فریبنده تولید شده توسط هوش مصنوعی

را تشخیص دهند و پلتفرم ها باید به حساب ها و صفحات سوءاستفاده کننده که این نوع خروجی ها را به اشتراک می گذارند، رسیدگی کنند.

درباره موردها

جنگ اسرائیل و ایران در ژوئن 2025 با حضور محتوای هوش مصنوعی مولد فریبنده در شبکه های اجتماعی که به عنوان «جنگ نرم» خود آن شناخته می‌شد، نقطه عطفی را رقم زد. بر اساس گزارش‌ها، این نوع خروجی فریبنده تعداد بسیار زیاد بازدید را باعث شد و هر دو دولت اسرائیل و ایران به تلاش های تاثیرگذاری مبتنی بر هوش مصنوعی متهم شدند. در 15 ژوئن، دو روز پس از آغاز جنگ 12 روزه اسرائیل و ایران، ویدیویی در یک صفحه فیسبوک منتشر شد که ادعا می‌کرد منبع خبری است. کاربر پست، در فیلیپین بود. ویدئو خسارات گسترده به ساختمان ها را نشان می داد و یک متن انگلیسی با تاریخ انتشار روی آن نوشته شده بود: «[sic] «Live now – Haifa Towards Down. این ویدیو بسیار شبیه ویدیویی بود که در TikTok منتشر شده بود و توسط یک راستی‌آزمای مستقل (خبرگزاری فرانسه) به عنوان نادرست و تولید شده توسط هوش مصنوعی شناسایی شده بود. کپشن پست فیسبوکی، عبارات تیترمانند زیادی مرتبط با جنگ و اصطلاحات و هشتگ های نامرتبط را ردیف کرده بود. این پست بیش از 700,000 بازدید داشت و نظرات متعدد اشاره می‌کردند که محتوا توسط هوش مصنوعی تولید شده است.

شش کاربر، این مورد را به Meta گزارش دادند، اما نه توسط شرکت بررسی شد و نه توسط راستی‌آزمایان طرف سوم. یک کاربر به «هیئت» اعتراض کرد. پس از انتخاب این پرونده توسط هیئت، Meta تأیید کرد که این پست با استاندارد جامعه اطلاعات نادرست مغایرت ندارد چون «مستقیما به خطر آسیب فیزیکی قریب الوقوع کمک نمی‌کند» و نیازی به برچسب هوش مصنوعی ندارد.

نشانه های آشکار فریب مرتبط با این پست باعث شد «هیئت» درباره هویت و رفتار حساب های مرتبط با صفحه، از Meta سؤال کند. سپس شرکت سه حساب مرتبط با صفحه را به دلیل سوءاستفاده از مشارکت و عدم اصالت غیرفعال کرد و صفحه و همراه با آن محتوای مربوطه را حذف کرد. صفحه، واجد شرایط درآمدزایی از طریق برنامه ستارگان Meta شده بود.

یافته های کلیدی

«هیئت» دریافت که محتوا خطر قابل توجهی برای گمراه کردن عمومی در موضوعی مهم در زمان حساس ایجاد می‌کرد، بنابراین Meta باید برچسب «هوش مصنوعی پرخطر» را اعمال می‌کرد. این پست آستانه حذف (نشان‌دهنده خطر آسیب فیزیکی یا خشونت قریب الوقوع) را برآورده نمی‌کرد. Meta برای مقابله با گسترش محتوای فریبنده تولید شده توسط هوش مصنوعی در پلتفرم های خود، از جمله توسط شبکه های غیرواقعی یا سوءاستفاده‌گرِ حساب ها و

صفحات، به ویژه در مسائل مربوط به منافع عمومی، باید اقدامات بیشتری انجام دهد، تا کاربران بتوانند محتوای واقعی را از جعلی تشخیص دهند.

«هیئت» نسبت به گزارش هایی مبنی بر این که Meta استانداردهای ائتلاف برای منشأ و اصالت محتوا (C2PA) را حتی برای محتوایی که توسط ابزارهای هوش مصنوعی خودش تولید می‌شود به صورت نامنظم اجرا می‌کند، و تنها بخشی از این خروجی ها برچسب‌گذاری مناسب دریافت می‌کنند، نگران است. C2PA استانداردهای فنی برای جاسازی اطلاعات منشأ به عنوان فراداده در محتوا تعیین می‌کند و باعث می‌شود پلتفرم ها آسان تر بتوانند محتوای تولید شده توسط هوش مصنوعی را شناسایی کنند و برچسب هایی را برای اطلاع رسانی به کاربران اعمال کنند.

مکانیزم های فعلی برای چسباندن حتی برچسب استاندارد اطلاعات هوش مصنوعی به ویدیو (افشای خود کاربر یا ارجاع به تیم سیاست محتوا)، نه به اندازه کافی قوی و نه جامع هستند تا با مقیاس و سرعت محتوای تولیدشده توسط هوش مصنوعی، به ویژه در شرایط بحرانی یا جنگ که تعامل در پلتفرم افزایش می‌یابد، مقابله کنند. سیستمی که بیش از حد به افشای خود کاربر درباره استفاده از هوش مصنوعی و شکایت‌های برجسته‌شده (که به ندرت رخ می دهد) برای برچسب‌گذاری درست این خروجی ها وابسته است، نمی‌تواند با چالش های موجود در محیط کنونی را رویارویی کند. برخی اعضای «هیئت» همچنین اشاره کردند که برچسب های هوش مصنوعی پرخطر (برای خروجی هایی که ممکن است افراد را در مسائل مهم فریب دهند) باید با کاهش رتبه یا حذف از توصیه ها همراه شوند تا نگرانی ها درباره گسترش تأثیر محتوای فریبنده برطرف شود.

رویکرد محدود Meta در پخش امتیازها به محتوای یکسان و تقریبا یکسان ممکن است به این معنی بوده باشد که این پست امتیاز راستی‌آزمایی دریافت نکرده است. محدودیت منابع و حجم قابل توجه خروجی باعث می‌شود بررسی‌راستی‌آزمایان، به ویژه در زمان جنگ یا بحران، نتوانند به موقع تمام محتوای فریبنده را بازبینی کنند. «هیئت» تأکید می‌کند که Meta باید اطمینان حاصل کند که راستی‌آزمایان درباره اولویت‌بندی محتوای مربوط به جنگ‌ها منابع کافی در اختیار دارند و راهنمایی شده‌اند. موارد تعیین پروتکل سیاست بحران (CPP) و رویدادهای پرطرفدار باید به Meta اجازه می‌داد تا حمایت مؤثرتر از راستی‌آزمایان طرف سوم را در طول بحران تضمین کند. پخش امتیازها به دسته وسیع تری از ویدیوهای بسیار مشابه می‌توانست آسیب های بالقوه را به صورت قابل توجه کاهش دهد، از جمله با تنزل درجه آنها. این مورد، ناکارآمدی های رویکرد فعلی Meta را در طول درگیری‌های مسلحانه برجسته می‌کند و نگرانی هایی را که «هیئت» پیش تر ابراز کرده بود، تشدید می‌کند.

نگران کننده است که با فعال شدن CPP و تخصیص منابع اضافی، Meta خودش سیگنال های واضح سوءاستفاده از مشارکت را از صفحه شناسایی نکرد و فقط در پاسخ به سوالات «هیئت»، حساب های مربوط به آن را بررسی کرد. به

جای تکیه بر روش‌های کاهش‌دهنده مبتنی بر محتوا که مستعد عدم موفقیت بالا هستند، اجرای دقیق سیاست های مبتنی بر رفتار می‌توانست از آسیب های ناشی از این حساب های متخلف جلوگیری کند.

تصمیم هیئت نظارت

«هیئت» تصمیم Meta برای حذف محتوا بدون برچسب هوش مصنوعی پرخطر را لغو می‌کند.

«هیئت» توصیه می‌کند که «متا»:

- یک استاندارد جامعه برای محتوای تولیدشده توسط هوش مصنوعی، جدا از استاندارد جامعه اطلاعات نادرست ایجاد کند که قوانین جامعی درباره حفظ منشأ، پروتکل های برچسب‌گذاری هوش مصنوعی و افشای خود کاربر ارائه کند.
- مسیرهایی برای افزودن برچسب های هوش مصنوعی پرخطر و پرریسک به محتوا به دفعات بیشتر، با کمک کانال های واضح تر انتقال شکایت از سیستم های خودکار و بازبینی با حجم بالا ایجاد کند تا این نوع برچسب‌گذاری بتواند با حجم بسیار بالاتر انجام شوند.
- اطلاعات منشأ و واترمارک های نامرئی را به محتوایی که توسط ابزارهای هوش مصنوعی Meta تولید می‌شود اضافه کند، از جمله افزودن اطلاعات کاربری محتوا (طبق مقررات C2PA) هنگام ایجاد.
- اطلاعات کاربری محتوا را در مقیاس وسیع اجرا کند و اطمینان حاصل کند که هرگاه جزئیات منشأ موجود باشد، به وضوح و به صورت مداوم قابل مشاهده و قابل دسترسی باشند.
- در ابزارهای شناسایی قوی تر برای محتوای چندفرمتی تولیدشده توسط هوش مصنوعی (صدا، صدا-تصویر و تصویر) سرمایه گذاری کند.
- توضیح واضحی درباره جریمه های عدم افشای محتوای دیجیتال ایجاد شده یا تغییر یافته، از جمله معیارهای جریمه و محدودیت های مربوطه منتشر کند.
- استاندارد جامعه اطلاعات نادرست را اصلاح کند تا اطمینان حاصل شود که بررسی سریع اطلاعات نادرست که مستقیما خطر خشونت قریب الوقوع یا آسیب جسمی را به همراه دارند، صرفا به سیگنال های شرکای خارجی وابسته نباشد. یک اهرم CPP باید با پشتیبانی تخصص و اقدام داخلی، از جمله برچسب‌گذاری و بررسی حساب ها و صفحات پست کننده، منابعی را برای شناسایی به موقع و پیشگیرانه این نوع محتوای ناقض تخصیص دهد.

*خلاصه های موردها، نمای کلی موردها را ارائه می دهند و ارزش سابقه ندارند.

تصمیم کامل پرونده

1. شرح و پس‌زمینه پرونده

در 13 ژوئن 2025، اسرائیل یک حمله هوایی بزرگ انجام داد و تأسیسات هسته ای و نظامی و سایر سایت های ایران را هدف قرار داد. رهبران اسرائیل گفتند این حملات با هدف جلوگیری از توسعه برنامه هسته ای ایران انجام شده است. این موضوع باعث شد دو کشور بیش از یک هفته و نیم حملات شدیدی را با هم رد و بدل کنند. ایران صدها موشک به شهرهای اسرائیل شلیک کرد، بسیاری از این حملات توسط سیستم دفاعی اسرائیل رهگیری شدند، در حالی که اسرائیل چندین نقطه در سراسر ایران، از جمله نزدیک پایتخت تهران، را هدف قرار داد. در 18 ژوئن، آنتونیو گوترش، دبیرکل سازمان ملل متحد اعلام کرد از تشدید جنگ «عمیقا نگران» است و افزود «هرگونه مداخله نظامی اضافی می‌تواند پیامدهای عظیم، نه تنها برای افراد دخیل بلکه برای کل منطقه و برای صلح و امنیت بین المللی در کل داشته باشد.» در 21 ژوئن، ایالات متحده حملاتی را با هدف سایت های هسته ای ایران انجام داد. در 24 ژوئن، آتش بس میان اسرائیل و ایران اعلام شد.

جنگ اسرائیل و ایران در سال 2025 با نفوذ رو به رشد محتوای مولد هوش مصنوعی در شبکه های اجتماعی که به تدریج به نام «جنگ نرم» شناخته می‌شود، نقطه عطفی را رقم زد. شبکه BBC گزارش داد که سه ویدئوی فریبنده تولید شده توسط هوش مصنوعی از این جنگ بیش از 100 میلیون بازدید داشت. وزیر امور خارجه اسرائیل ویدیویی از حمله به زندان اوین در تهران منتشر کرد که بعدا تحلیل های قانونی احتمال دادند ویدئوی تولید شده توسط هوش مصنوعی است، با اینکه حمله ای به زندان واقعا رخ داده بود. تحقیقات آزمایشگاه Citizen در دانشگاه تورنتو از یک شبکه هماهنگ از پروفایل های غیرواقعی در X (که قبلا Twitter نام داشت) خبر داد که ظاهرا با اسرائیل مرتبط است و ایرانیان را به مقابله با دولت شان تشویق می‌کند. دولت اسرائیل نیز از کمپین‌های ربات محور ایران گزارش داد که با هدف شکل‌دهی به دیدگاه ها پیرامون این جنگ و تأثیر حملاتشان بر اسرائیل فعالیت می‌کردند.

محتوای فریبنده تولید شده توسط هوش مصنوعی در سال های اخیر چالشی رو به رشد و مداوم در بحران ها و جنگ های سراسر دنیا بوده است. این چالش در مناطقی که آزادی بیان تحت فشار است تشدید می‌شود و سرکوب رسانه های مستقل و تعطیلی اینترنت، جریان اطلاعات معتبر را که می‌تواند کمپین های فریبنده را آشکار کند، مسدود می‌کند. در جریان بررسی این مورد، هیئت مشاهده کرد که عملیات آمریکا برای دستگیری رئیس جمهور ونزوئلا و اعتراضات گسترده ضد دولتی در ایران، شامل ادعاهایی مبنی بر واقعی بودن محتوای فریبنده تولید شده توسط هوش مصنوعی و ادعاهای متقابل درباره جعلی بودن محتوای واقعی بود. هر دو حالت، توانایی انسان‌ها در تشخیص محتوای ساختگی از واقعی به چالش کشیدند، که نماد سود دروغین است و منجر به بی‌اعتمادی عمومی به همه اطلاعات می‌شود.

چندین رویکرد فنی ظهور کرده‌اند تا به پلتفرم ها و کاربران کمک کنند رسانه های مصنوعی یا دستکاری شده را از واقعی تشخیص دهند. یکی از رویکردها ردیابی منشأ است، یعنی تاریخچه قابل راستی آزمایی یک دارایی دیجیتال، مانند تصویر، ویدئو یا فایل. ائتلاف برای منشأ و اصالت محتوا (C2PA) استانداردهای فنی برای جاسازی اطلاعات منشأ به صورت فراداده در محتوا تعیین کرده است که به پلتفرم ها اجازه می دهد محتوای تولید شده توسط هوش مصنوعی را راحت تر شناسایی کنند و برچسب هایی برای اطلاع رسانی به کاربران اعمال کنند. در حالی که این ابزارها هنوز در حال تکامل هستند و هیچ کدام راه حل کاملی نیستند، آنها به عنوان «کف تولید و انتشار مسئولانه محتوای تولید شده توسط هوش مصنوعی» در نظر گرفته می‌شوند. به موازات آن، سرمایه گذاری در تشخیص خودکار، مانند طبقه‌بندها، ممکن است راهی برای کشف سیگنال های دیگر مبنی بر تولید محتوا توسط هوش مصنوعی فراهم کند.

این مورد نماد این چالش هاست. در 15 ژوئن، همزمان با تشدید جنگ اسرائیل و ایران، ویدیویی در یک صفحه فیسبوکی منتشر شد که ادعا می‌کرد منبع خبری است و 161 هزار دنبال کننده داشت. ویدئو خسارات گسترده به ساختمان ها را نشان می داد که ستون های دود و آوار آن را احاطه کرده بودند، و یک متن انگلیسی با تاریخ انتشار روی آن نوشته شده بود: «[sic] Live now – Haifa Towards Down». این پست احتمالا به حیفا، شهری در شمال اسرائیل اشاره داشت. این ویدیو خیلی شبیه ویدیویی است که در TikTok منتشر شده و از سوی راستی‌آزمایان مستقل یعنی خبرگزاری فرانسه (AFP)، به عنوان جعلی و تولید شده توسط هوش مصنوعی تشخیص داده شده است (AFP فقط تصاویر ثابت ویدئوی رتبه‌بندی شده را ارائه می کند، ولی این تصاویر با فریم های محتوای مورد یکسان هستند). کپشن پست فیسبوکی به انگلیسی، عبارات تیترمانند زیادی مرتبط با جنگ و نیز اصطلاحات و هشتگ های نامرتبط را بدون روایت واضح ردیف کرده بود. در آن به جنگ جاری، رهبران سیاسی جهان، آتش سوزی های جنگلی، موشک ها و موارد دیگر اشاره شده بود. این پست بیش از 700,000 بازدید داشت و نظرات متعدد اشاره می‌کردند که محتوا توسط هوش مصنوعی تولید شده است.

شش کاربر، محتوای مورد را در مجموع نه بار گزارش کردند، اما سیستم های خودکار Meta آن را برای بررسی انسانی در اولویت قرار ندادند. در همان روز که محتوا منتشر شد، یک طبقه‌بند اطلاعات نادرست آن را به لیست موارد قابل بررسی از سوی بررسی‌کنندگان طرف سوم افزود، اما هرگز بازبینی یا رتبه‌بندی نشد. این موضوع غیرمعمول نبود، چرا که Meta حجم قابل توجهی از اطلاعات نادرست احتمالی را شناسایی می‌کند که از ظرفیت راستی‌آزمایان فراتر می رود.

پس از اتمام مراحل تجدیدنظر داخلی در شرکت، یکی از کاربران گزارش دهنده، نسبت به تصمیم Meta برای کنار گذاشتن محتوا برای اعلام نظر از سوی «هیئت»، اعتراض کرد. پس از انتخاب این پرونده توسط هیئت، Meta تأیید کرد که این پست با استاندارد جامعه اطلاعات نادرست مغایرت ندارد چون «مستقیما به خطر آسیب فیزیکی قریب الوقوع کمک نمی‌کند». Meta همچنین از تصمیم خود برای برچسب نزدن محتوا دفاع کرد. هیچ اقدامی علیه صفحه یا حساب های مسئول محتوا انجام نشد.

به دلیل نشانه های آشکار فریب در این پست، «هیئت» یک سری سوال درباره هویت و رفتار صفحه و حساب های مرتبط با آن از Meta پرسید. این موضوع منجر به تحقیق شد و Meta متوجه شد مدیران صفحه قوانین مربوط به سوءاستفاده و مشارکت درست را نقض کرده اند. سپس شرکت سه حساب کاربری را به صورت دائمی غیرفعال کرد که باعث حذف صفحه و محتوای مورد از پلتفرم شد.

2. مکاتبات کاربر

کاربری که درخواست حذف محتوا را داده بود، در پیامش به هیئت، شکایت کرد که Meta در پلتفرم خود به «اعمال تروریستی» اجازه می دهد. در پیام او هیچ نشانه واضحی وجود نداشت که می‌دانست که محتوا با هوش مصنوعی تولید شده یا گمراه کننده بود.

3. مکاتبات و خطمشی‌های محتوای متا

I. خطمشی‌های محتوای متا

استاندارد جامعه اطلاعات نادرست

بر اساس استاندارد جامعه اطلاعات نادرست، Meta «اطلاعات نادرست یا شایعات غیرقابل راستی آزمایی را که شرکای متخصص تشخیص داده اند احتمالا به صورت مستقیم به خطر خشونت قریب الوقوع یا آسیب جسمی به افراد کمک می‌کند» حذف می‌کند. در کشورهایی که «خطر خشونت اجتماعی را تجربه می‌کنند»، Meta «به صورت پیشگیرانه با شرکای محلی همکاری می‌کند تا بفهمد کدام ادعاهای نادرست ممکن است مستقیما به خطر آسیب فیزیکی قریب الوقوع کمک کنند» و محتوای مطرح کننده این ادعاها را شناسایی و حذف کند.

تحت عنوان فرعی «رسانه های دستکاری شده»، Meta اعلام می‌کند که برای محتوایی که به شکل دیگر خلاف استانداردهای جامعه نیست، وقتی تصویر یا ویدئوی فوتورئالیستی یا صدای شبه‌واقعی است که به صورت دیجیتال ایجاد یا تغییر یافته و «خطر بسیار بالایی برای فریب مادی عموم در موضوعی با اهمیت عمومی» ایجاد می‌کند، ممکن است یک برچسب اطلاع رسانی به پست اضافه کند. سیاست اطلاعات نادرست همچنین کاربران را ملزم می‌کند هر زمان که «محتوای ارگانیک با ویدئوی فوتورئالیستی یا صدای شبه‌واقعی که به صورت دیجیتال ایجاد یا تغییر یافته» منتشر می‌کنند، افشا کنند. عدم استفاده از ابزار افشای هوش مصنوعی ممکن است منجر به جریمه شود.

در جاهای دیگر، Meta همچنین محتوا و رفتارهایی را که «اغلب با انتشار اطلاعات نادرست همپوشانی دارند» ممنوع می‌کند. این شامل استانداردهای جامعه در بی‌نقصی حساب، رفتارهای فریبنده و رفتار هماهنگ و غیرواقعی نیز می‌شود. برای سایر اطلاعات نادرست که با استاندارد جامعه اطلاعات نادرست آن مغایرت ندارد، Meta بر «کاهش شیوع آن یا ایجاد محیطی که گفت وگوی سازنده را تقویت کند» تمرکز دارد. خارج از ایالات متحده، Meta برای بررسی و ارزیابی محتوا به راستی‌آزمایان مستقل طرف سوم متکی است که می‌تواند منجر به افزودن برچسب های متناسب با رتبه‌بندی به محتوا شود. رتبه بندی ها شامل «نادرست» و «تغییر یافته» است و ممکن است منجر به کاهش توزیع محتوا شود. Meta از فن‌آوری برای آشکار کردن اطلاعات نادرست احتمالی برای بررسی راستی‌آزمایان استفاده می‌کند، و راستی‌آزمایان همچنین می‌توانند خودشان محتوا را شناسایی و بررسی کنند. در ژانویه 2025، Meta اعلام کرد که به برنامه راستی آزمایی طرف سوم در ایالات متحده پایان داده و به جای آن به حالت یادداشت های جامعه گرایش پیدا کرده است.

در Metaسیاست های درآمدزایی شرکا قوانینی را برای صفحات جهت «کسب درآمد از پلتفرم ها» وضع می‌کنند و اشاره می‌کنند که محتوا علامت گذاری شده به عنوان اطلاعات نادرست یا طعمع کلیک، نمی‌تواند واجد شرایط کسب درآمد باشد. سیاست های درآمدزایی محتوا همچنین قوانینی برای «ایجاد محتوای ایمن برای برند و قابل درآمدزایی» را تشریح می‌کند و درآمدزایی را در محتوایی که موضوعات خاص مانند «تراژدی و جنگ، از جمله خسارت مادی» را نشان می دهد یا توصیف می‌کند، محدود یا کاهش می دهد. در همین راستا، حساب های مرتبط با صفحه در این مورد هر دو به دلیل نقض های سطح حساب کاربری و سوءاستفاده از مشارکت حذف شدند.

II. مکاتبات «متا»

Meta اعلام کرد که این پست ناقض سیاست اطلاعات نادرست که حذف محتوا را الزامی می‌کند نیست: «احتمالا مستقیما به خطر آسیب جسمی قریب الوقوع کمک می‌کند.» تصمیم آنها با در نظر گرفتن این موضوع بود که هیچ کارشناس مستقل، مثلا یک همکار محلی، محتوا یا هرگونه ترند اطلاعات نادرست مرتبط با آن را گزارش نکرد.

Meta هیچگونه برچسبی بر محتوا طبق قوانین رسانه دستکاری شده اعمال نکرد. Meta سه برچسب مختلف را برای رسانه های دستکاری شده اعمال می‌کند: اطلاعات هوش مصنوعی، پرخطر، یا هوش مصنوعی پرخطر.

برچسب **اطلاعات هوش مصنوعی** زمانی به صورت خودکار روی محتوا اعمال می‌شود که Meta «شاخص های تصویری استاندارد صنعت هوش مصنوعی را تشخیص می‌دهد یا زمانی که افراد اعلام کرده اند محتوای تولید شده توسط هوش مصنوعی را بارگذاری می‌کنند». همان طور که «هیئت» در مورد مربوط به تماس صوتی ادعایی برای دستکاری انتخابات در کردستان عراق آشکار کرد، Meta در حال حاضر فقط قادر به شناسایی خودکار و زدن برچسب اطلاعات هوش مصنوعی روی تصاویر استاتیک است، آن هم با تکیه بر فراداده هایی که بسیاری از ابزارهای مولد هوش مصنوعی در چنین محتوایی جاسازی می‌کنند. برای اعمال برچسب روی محتوای صوتی یا تصویری، افشای خود کاربران الزامی است. در این مورد هیچگونه افشای شخصی موجود نبود. در فرآیندهای فعلی Meta، در این شرایط اضافه کردن خودکار برچسب ممکن نبود.

برچسب **پرخطر شامل** محتوایی می‌شود که (1) خطر بسیار بالایی برای فریب مادی عموم در موضوعی با اهمیت عمومی ایجاد می‌کند؛ و (2) نشانگرهای قابل اطمینان درباره ایجاد یا تغییر دیجیتال دارد. برخلاف برچسب اطلاعات هوش مصنوعی، برچسب پرخطر فقط سیاست گزارش به رده‌های بالاتر است، به این معنی که فقط تیم های سیاست داخلی Meta می‌توانند پس از بررسی انسانی این برچسب را اعمال کنند. پیش از انتخاب این مورد توسط هیئت، این نوع گزارش به رده بالاتر در محتوای مورد نظر رخ نداده بود.

برچسب هوش مصنوعی **پرخطر زمانی** اعمال می‌شود که محتوا تمام الزامات برچسب پرخطر را برآورده می‌کند و نشانگرهای قابل اطمینان ایجاد یا تغییر با هوش مصنوعی را دارد. این سیاست همچنین صرفا مبتنی بر گزارش به رده‌های بالا است و نیازمند بررسی انسانی است. تیم های سیاست گذاری داخلی Meta این محتوا را تنها پس از انتخاب «هیئت» بررسی کردند. آنها تشخیص دادند که تا آن موقع، برای مرتبط یا فوری بودن برچسب، زمان بسیار زیادی گذشته بود.

هنگام تعیین اینکه آیا محتوا محصول هوش مصنوعی است یا به صورت دیجیتال ایجاد یا تغییر یافته، Meta منابع خارجی و داخلی موجود را در نظر می‌گیرد که معتبر تلقی می‌کند. منابع خارجی ممکن است شامل اخبار یا سازمان های مستقل راستی‌آزمای طرف سوم باشند که می‌توانند مبنای فنی برای تعیین خود ارائه دهند، مانند ارجاع به مدل تشخیص هوش مصنوعی یا نتیجه گیری از یک کارشناس قانونی.

همان طور که در تصمیم مربوط به تصاویر اعتراضی همراه با شعارهای طرفدار دوترته شرح داده شد، برچسب‌های رسانه های دستکاری شده منجر به کاهش خودکار رتبه محتوا یا حذف آن از موارد توصیه‌شده نمی‌شوند. در عوض، کاربرانی که محتوایی با این برچسب ها را به اشتراک می گذارند ممکن است یک پنجره پاپ آپ دریافت کنند که به صورت طبیعی دسترسی را کاهش می دهد. ایجاد پیام های پاپ آپ به این بستگی دارد که محتوا با هوش مصنوعی ساخته شده باشد یا خیر.

Meta چندین سیستم خارج از ایالات متحده برای شناسایی و مقابله با اطلاعات نادرست احتمالی دارد. برخی سیستم ها فقط محتوا را برای راستی‌آزمایی توسط طرف سوم ارسال می‌کنند، در حالی که برخی دیگر هم محتوا را برای بازبینی ارسال می‌کنند و هم در حالی که منتظر بازبینی هستند، کاهش رتبه موقت را اعمال می‌کنند. اینکه محتوا فقط برای بررسی ارسال شود یا دسترسی به آن کاهش یابد، به نوع سیستمی که آن را گزارش می‌کند و نیز عواملی مانند کشور و زبان مربوطه بستگی دارد.

طبق گزارش Meta، پروتکل سیاست بحران (CPP)در زمان جنگ اسرائیل و ایران فعال شد. پس از حملات 7 اکتبر 2023، اسرائیل برای CPP تعیین شده بود. ایران در آغاز جنگ ژوئن 2025 تعیین شد. فعال سازی این پروتکل به شرکت اجازه می دهد مجموعه ای از اهرم ها را به کار گیرد که برای تقویت پاسخ به بحران و امکان ارزیابی و کاهش ریسک آسیب قریب الوقوع توسط تیم های آن طراحی شده اند. در طول این بحران، بحران منجر به هیچ تغییری در سیستم‌های مدیریت خودکار نشدو محتوای مورد با استفاده از مدل ها و آستانه های موجود بازبینی شد.

Meta جنگ اسرائیل و ایران را به عنوان «رویداد ترند» معرفی کرد تا با توجه به خطر بالای انتشار اطلاعات نادرست، بهتر بتواند در شناسایی و رد ادعاهای نادرست رایج مرتبط با این جنگ به راستی‌آزمایان طرف سوم کمک کند. تا زمانی که آتش بس برقرار شد، راستی‌آزمایان چندین بررسی درستی مرتبط با جنگ منتشر کرده بودند. Meta اعلام می‌کند که راستی‌آزمایانی در اسرائیل و فیلیپین (جایی که کاربر پست کننده مستقر است) دارد، ولی نه در ایران.

پس از آنکه «هیئت» از Meta خواست رفتار این صفحه و حساب های مرتبط را بررسی کند، شرکت حساب های سه مدیر متفاوت صفحه را غیرفعال کرد.

یک مدیر به دلیل نقض سیاست نمایش هویت واقعی با «مشارکت در تحریف هویت برای گمراه کردن یا فریب دیگران، فرار از اجرا یا نقض استانداردهای جامعه ما» غیرفعال شد. حساب دوم بر اساس سیاست بی‌نقصی حساب به دلیل مالکیت توسط همان شخص یا نهاد صاحب حساب غیرفعال، غیرفعال شد. حساب سوم بر اساس سیاست هرزنامه به دلیل سوءاستفاده از مشارکت غیرفعال شد. سیاست هرزنامه به صورت کلی انواع رفتارهای فریبنده، گمراه کننده یا استرسزا برای افزایش غیرواقعی مشارکت در پست ها را ممنوع می‌کند. این موضوع، منجر به حذف صفحه و محتوای منتشرشده توسط آن شد. پیش از حذف، این صفحه واجد شرایط کسب درآمد از طریق برنامه ستارگان Meta بود.

«هیئت» سوالاتی درباره برچسب‌گذاری، شناسایی هوش مصنوعی، راستی‌آزمایی، اجرای CPP، رفتارهای صفحه و سطح حساب کاربری و موارد دیگر پرسید. «متا» به همه پرسش‌ها پاسخ داد.

4. نظرات عمومی

هیئت شش نظر عمومی دریافت کرد که با شرایط ارسال مطابقت می‌کردند. چهار نظر از اروپا و دو نظر از ایالات متحده ارسال شده بود. برای خواندن نظرات عمومی که با رضایت برای انتشار ارسال شده‌اند، اینجا را کلیک کنید.

نظرات ارائه‌شده موضوع‌های زیر را پوشش می‌دادند: نظارت بر محتوا در دوران جنگ و بحران، شیوع محتوای تولید شده توسط هوش مصنوعی و افزایش رفتارهای هماهنگ و غیرواقعی در درگیری مسلحانه، محدودیت های تعریف Meta از «آسیب فیزیکی قریب الوقوع»، اهمیت راستی‌آزمایی در درگیری مسلحانه، استانداردها و اجرای برچسب های رسانه‌های دستکاری شده، اهمیت استانداردهای C2PA در شناسایی و موارد دیگر.

5. تحلیل هیئت نظارت

T«هیئت» این مورد را برای بررسی سیاست ها و شیوه های اجرایی Meta در ارتباط با اشتراک گذاری محتوای فریبنده تولید شده توسط هوش مصنوعی در پلتفرم هایش، به ویژه در زمینه درگیری های مسلحانه، انتخاب کرد.

این مورد در چارچوب اولویت های استراتژیک وضعیت‌های بحران و جنگ و هوش مصنوعی و اتوماسیون «هیئت» است.

هیئت تصمیم Meta در این مورد را نسبت به سیاست های محتوا، ارزش ها و مسئولیت های حقوق بشر Meta تحلیل کرد. «هیئت» پیامدهای این پرونده را برای رویکرد وسیع‌تر «متا» در زمینه نظارت بر محتوا نیز ارزیابی کرد.

5.1 انطباق با سیاست های محتوای Meta

قوانین محتوا

هیئت دریافت که محتوای مورد تحت استاندارد جامعه اطلاعات نادرست نیازی به حذف نداشت، ولی Meta باید طبق قوانین خود درباره رسانه های دستکاری شده، برچسب «هوش مصنوعی پرخطر» را بر محتوا اعمال می‌کرد.

این پست، ناقض سیاست های اطلاعات نادرست Meta برای حذف نبود، چون احتمالا به خشونت قریب الوقوع یا آسیب جسمی به افراد منجر نمی‌شد. این ویدیو به صورت فریبنده تأثیر حمله ایران به اسرائیل را اغراق آمیز نشان می داد و درست همزمان با سقوط موشک ها به حیفا منتشر شد. اگرچه این موضوع احتمالا به ناراحتی افراد فریب خورده با محتوای پست افزود، ولی احتمالا به خشونت غیرنظامیان اسرائیلی علیه دشمنان فرضی کمک نمی‌کرد یا مستقیما بر واکنش دولت اسرائیل تأثیر نمی گذاشت. هیچ نشانه ای از خشونت میان جوامع داخل اسرائیل در واکنش به تصاویر فریبنده از حملات وجود ندارد.

با وجود این، نگران کننده است که Meta تحلیلی درباره خطر احتمالی آسیب در تحلیلش از محتوا ارائه نکرده است. در عوض، نتیجه گرفت که چون هیچکدام از همکاران مورد اطمینان محتوا را به آنها گزارش نکرده است، محتوا قوانین آن را نقض نکرده است. در حالی که بسیاری از همکاران مورد اطمینان به «هیئت» اطلاع می دهند که شرکت کمتر به تماس ها و نگرانی ها پاسخ می دهد، که تا حدودی به دلیل کاهش قابل توجه ظرفیت تیم های داخلی Meta است، این موضع قابل قبول نیست. Meta باید بتواند این نوع ارزیابی ها از آسیب را انجام دهد، نه اینکه صرفا به همکارانی که در طول درگیری مسلحانه با آنها تماس می گیرند اتکا کند. این صفحه به صورت گسترده دنبال می‌شد، محتوایش وایرال بود، طبقه‌بند اطلاعات نادرست Meta محتوا را علامت زد و چندین کاربر آن را گزارش کردند. منابع معتبری مانند AFP ویدئوی بسیار مشابهی را غیرواقعی اعلام کردند و به Meta هشدار دادند تا ادعاهای فریبنده ای را که می‌توانستند آسیب زا باشند، فعالانه بازبینی کند. فعال سازی CPP باید

تضمین می‌کرد که منابع لازم برای انجام این نوع ارزیابی ها توسط خود Meta و تماس فعالانه با همکاران برای دریافت اطلاعات واقعی در صورت لزوم وجود دارد. در این شرایط، محتوا باید برای بازبینی، به رده‌های بالاتر گزارش می‌شد.

اگر این اتفاق می افتاد، مشخص می‌شد که محتوا خطر جدی گمراه کردن عموم در موضوعی مهم در زمان حساس را به همراه دارد و در این صورت برچسب «هوش مصنوعی پرخطر» اعمال می‌شد. این صفحه که خود را به عنوان یک رسانه خبری معرفی می‌کرد و متنی روی ویدیو که ادعا می‌کرد «Live Now» (زنده اکنون) و از حیفا است، محتوای مربوطه را به عنوان تصاویر واقعی یک درگیری مسلحانه جاری که جان غیرنظامیان در آن در خطر بود، نشان می داد. این کاربر که صفحه اش پیش تر به صورت غیرواقعی خود را به عنوان منبع خبری معتبر معرفی می‌کرد، با افشای داوطلبانه محتوای هوش مصنوعی، فریب خودش را فاش نمی‌کرد. پرندگان سفید غیرواقعی که در ویدیو پرواز می‌کنند و سازمان های متخصصی مانند AFP که محتوا را تولید شده توسط هوش مصنوعی تشخیص داده اند، باید Meta را ترغیب می‌کردند ببیند که برچسب‌گذاری لازم است یا خیر. به محض مطرح شدن موضوع از سوی «هیئت»، Meta باید اشتباهش را اصلاح می‌کرد. تحقیقات «هیئت» در پلتفرم، چندین نسخه از این ویدیو و تصاویر مرتبط را چند هفته بعد در پلتفرم های Meta منتشر کرد و شرکت هیچ کدام را برچسب نزده بود.

اگر این پست خاص هم توسط راستی‌آزمایان طرف سوم بررسی شده بود، این محتوا احتمالا برچسب جعلی دریافت می‌کرد و کاهش رتبه می گرفت (بر اساس بر رتبه‌بندی AFP برای یک ویدیوی بسیار مشابه). رویکرد محدود Meta در پخش رتبه‌بندی به محتوای یکسان و تقریبا یکسان ممکن است باعث شده باشد که این محتوا نیز امتیاز نگیرد (مثلا به دلیل اضافه شدن یک متن به ویدیو). محدودیت منابع و حجم قابل توجه محتوا باعث می‌شود راستی‌آزمایان، به ویژه در زمان جنگ یا بحران، نتوانند به موقع تمام محتوای فریبنده را بازبینی کنند. «هیئت» تأکید می‌کند که Meta باید اطمینان حاصل کند که راستی‌آزمایان منابع کافی دارند و راهنمایی هایی برای اولویت بندی محتوای مربوط به جنگ‌ها دارند تا کار چالش برانگیزی را که Meta روی آنها حساب می‌کند تا ارائه کنند، انجام دهند (نگاه کنید به تصاویر اعتراض‌ها همراه با شعارهای طرفدار دوترته).

نگران کننده است که در طول این بحران، با فعال شدن CPP و تخصیص منابع اضافی، Meta خودش سیگنال های واضح سوءاستفاده از مشارکت را از صفحه شناسایی نکرد و فقط در پاسخ به سوالات «هیئت»، حساب های مربوط به آن را بررسی کرد. به جای تکیه بر روش های کاهش‌دهنده مبتنی بر محتوا که مستعد نرخ بالای عدم موفقیت هستند، اجرای دقیق سیاست های یکپارچگی و اصالت مبتنی بر رفتار می‌توانست از آسیب های ناشی از این محتوای فریبنده که توسط چندین حساب متخلف پشتیبانی می‌شد، جلوگیری کند.

5.2 انطباق با مسئولیت های حقوق بشری Meta

«هیئت» دریافت که تحت مسئولیت های حقوق بشری Meta، باید برچسب رسانه دستکاری شده «هوش مصنوعی پرخطر» را به محتوا اعمال می‌شد و Meta باید اقدامات بیشتری برای مقابله با گسترش محتوای فریبنده تولید شده توسط هوش مصنوعی در پلتفرم هایش، از جمله توسط شبکه های غیرواقعی یا سوءاستفاده کننده از حساب ها و صفحات انجام دهد.

آزادی بیان (ماده 19 ICCPR)

ماده 19 میثاق بین المللی حقوق مدنی و سیاسی (ICCPR) حفاظت گسترده از بیان نظرات، از جمله سخنان سیاسی را فراهم می‌کند. این حق شامل «آزادی جستجو، دریافت و انتقال اطلاعات و ایده ها از هر نوع» است (ماده 19، بند 2). وقتی محدودیت هایی بر بیان نظرات توسط یک دولت اعمال می‌شود، باید الزامات قانونی بودن، هدف مشروع و ضرورت و تناسب را برآورده کنند (ماده 19، پاراگراف ۳، ICCPR). این الزامات اغلب به‌عنوان «آزمون سه بخشی» شناخته می‌شوند. «هیئت» از این چارچوب استفاده می‌کند تا مسئولیت‌های حقوق بشری «متا» را تفسیر کند که درراستای «اصول راهنمایی سازمان ملل برای حقوق بشر و کسب‌وکار» باشد و «متا» در «خط‌مشی حقوق بشر شرکت»، خود را به آن پایبند می‌داند. «هیئت» این کار را نیز در ارتباط با تصمیم برای محتوای تکی دردست مرور انجام می‌دهد و هم اینکه آیا این مورد درباره «متا» و رویکرد وسیع‌تر آن در زمینه نظارت بر محتوا چه می‌گوید. همان طور که گزارشگر ویژه سازمان ملل در امور آزادی بیان اعلام کرده است، اگرچه «شرکت ها تعهدات دولت ها را ندارند، تأثیر آنها به گونه ای است که آنها را ملزم می‌کند نوع پرسش های یکسان را درباره حفاظت از حق آزادی بیان کاربران‌شان ارزیابی کنند» (A/74/486، پاراگراف .41)

همچنین، کارگروه سازمان ملل در زمینه کسب وکار و حقوق بشر اعلام کرده است که «از آنجا که خطر نقض های شدید حقوق بشر در مناطق آسیب دیده از جنگ افزایش می‌یابد»، بررسی های لازم توسط کسب وکارها باید «متناسب با آن افزایش یابد» (A/75/212، پاراگراف .13 در گزارشی مربوط به سال 2024 درباره جنگ اسرائیل و غزه، گزارشگر ویژه سازمان ملل در زمینه آزادی بیان و عقیده مطرح کرد که پلتفرم ها به صورت مداوم در انجام این مسئولیت در جنگ ها کوتاهی می‌کنند و به خطرات فزاینده اطلاعات گمراه‌کننده و نادرست در چنین موقعیت هایی اشاره کرد (A/79/319، پاراگراف 60، 66). در جنگ اسرائیل و ایران، قطع اینترنت دسترسی غیرنظامیان به اطلاعات را به شدت تحت تأثیر قرار داد و باعث ایجاد خلأیی شد که رسانه های فریبنده تولید شده توسط هوش مصنوعی به سرعت آن را پر کردند (نگاه کنید به WITNESS PC-31545).

یک مجموعه قوی از ابزارها در اختیار Meta قرار دارد تا آسیب های احتمالی محتوای فریبنده تولید شده توسط هوش مصنوعی را در پلتفرم هایش کاهش دهد. این مورد نشان می دهد که شناسایی و برچسب‌گذاری باید به صورت منسجم تر، مکرر تر و مؤثرتر اجرا شود تا از آسیب های احتمالی کوتاه مدت به کاربران جلوگیری شود، به ویژه در شرایط جنگ که خطرات بسیار بالاتر است. این امر باید با منابع کافی برای راستی‌آزمایان طرف سوم و راهنمایی برای اولویت بندی محتوای مربوط به جنگ ها، و نیز سرمایه گذاری در مقابله دقیق علیه سوءاستفاده های رفتاری از حساب و سطح صفحه پشتیبانی شود.

I. قانون‌مندی (شفافیت و دسترسی به قوانین)

اصل قانونی بودن ایجاب می‌کند قوانین مربوط به بیان باید قابل دسترس و واضح باشند، و با دقت کافی تنظیم شوند تا فرد بتواند رفتار خود را به صورت متناسب تنظیم کند (نظر عمومی‌شماره 34، پاراگراف .(25 همچنین، این قوانین «نمی‌توانند اختیار بی قید و شرط برای محدود کردن آزادی بیان برای افراد مسئول اجرای آنها اعطا کنند» و باید «راهنمایی کافی برای کسانی که مسئول اجرا هستند فراهم کنند تا بتوانند تشخیص دهند چه نوع ابراز بیان به درستی محدود می‌شود و چه نوع نمی‌شود» (*همان منبع*). گزارشگر ویژه سازمان ملل در امور آزادی بیان اعلام کرده است که وقتی قوانین در مدیریت آزادی بیان آنلاین توسط فعالان خصوصی اعمال می‌شوند، باید واضح و مشخص باشند (38/35/A/HRC، پاراگراف .(46 افرادی که از پلتفرم‌های «متا» استفاده می‌کنند باید بتوانند به قوانین دسترسی داشته باشند و آن‌ها را بفهمند و بازبینی‌کنندگان محتوا باید دررابطه با اجرای قانون‌هایشان رهنمودهای روشنی داشته باشند.

استاندارد جامعه اطلاعات نادرست باید وضوح بیشتری برای کاربران و مجریان قوانین فراهم کند.

توضیح کلی Meta درباره همکاری با طرف‌های سوم برای شناسایی ترندهای فریبنده به وضوح نشان نمی دهد که اجرای قوانین کاملا به بیان نگرانی ها از سوی همکاران به Meta وابسته است. به جای روشن کردن این موضوع، به دلایلی که در ادامه توضیح داده شده است، رویکردی متفاوت لازم است.

تنها توصیف عمومی دقیق از سه برچسب رسانه دستکاری شده که Meta استفاده می‌کند، درتصمیمات «هیئت» موجود است. «هیئت» بار دیگر بر توصیه‌اش تأکید می‌کند که Meta به صورت کامل سه برچسب رسانه دستکاری

شده مورد استفاده‌اش، معیارهای استفاده از آنها و پیامدهای آنها را شرح داده است (تصاویر اعتراضات همراه با شعارهای طرفدار دوترته)، در حالی که اشاره می‌کند این رویکرد باید تکامل یابد.

همچنین، سیاست رسانه دستکاری شده به صورت عمومی جریمه‌هایی را که در صورت عدم افشای استفاده از هوش مصنوعی توسط خود کاربر «ممکن است» اعمال شود، به صورت عمومی توصیف نمی‌کند. Meta به «هیئت» توضیح داد که این جریمه ها فقط در صورت گزارش به رده‌های بالاتر و در پاسخ به عدم افشای مکرر اعمال می‌شوند و ممکن است توزیع محتوا را تحت تأثیر قرار دهند یا موقتا منجر به تعلیق برخی ویژگی های حساب شوند. Meta در این زمینه اختیار گسترده ای دارد و اطلاعات واضح تری باید به کاربران ارائه شود.

ارائه اطلاعات بیشتر درباره منشأ محتوا، برچسب‌گذاری و خودافشایی در استاندارد جامعه اطلاعات نادرست ممکن است با تلاش ها برای کاهش فریب از طریق اقدامات مثبت برای ارتقای بی‌نقصی اطلاعات، باعث سردرگمی شود. همه موارد استفاده محتوای تولید شده توسط هوش مصنوعی و همه کاربردهای برچسب ها، به تلاش برای فریب پاسخ نخواهند داد. تدوین این قوانین در یک استاندارد جامعه جداگانه می‌تواند رویکرد Meta را روشن تر کند و رفتار کاربران را بهبود بخشد.

II. هدف مشروع

هرگونه محدودیت بر آزادی بیان باید یک یا چند هدف مشروع مذکور در ICCPR را نیز دنبال کند که شامل حفاظت از امنیت و حقوق دیگران می‌شود.

در ویدئوی تغییر یافته از رئیس جمهور بایدن هیئت تأکید کرد که «جلوگیری از گمراه شدن افراد، به خودی خود دلیل مشروعی برای محدود کردن آزادی بیان نیست.» با این حال، استاندارد جامعه اطلاعات نادرست همچنین در نظر دارد خطر آسیب فیزیکی یا خشونت قریب الوقوع به افراد را کاهش دهد که هدفی مشروع در ارتباط با حقوق دیگران است (نظر عمومی 34، پاراگراف 28).

III. ضرورت و تناسب

بر اساس ماده ICCPR 19(3)، ضرورت و تناسب ایجاب می‌کند که محدودیت های بیان «باید برای تحقق کار حفاظت‌شان مناسب باشند؛ آنها باید ابزاری با کمترین مداخله در میان ابزارهایی باشند که ممکن است کار محافظت‌شان را انجام دهند؛ آنها باید متناسب با منافعِ مورد حفاظت باشند» (نظر عمومی‌شماره 34، پاراگراف 34).

گزارشگر ویژه سازمان ملل در امور آزادی بیان اعلام کرده است که «در طول درگیری های مسلحانه، انسان‌ها در آسیب پذیرترین وضعیت خود هستند و برای تضمین امنیت و رفاه‌شان بیشترین نیاز را به اطلاعات دقیق و قابل اعتماد دارند. با این حال، دقیقا در همین شرایط است که آزادی عقیده و بیان آنها [...] با شرایط جنگ و اقدامات طرف های درگیر و سایر بازیگران برای دستکاری و محدود کردن اطلاعات به منظور اهداف سیاسی، نظامی و راهبردی محدود می‌شود» (77/288/A، پاراگراف 1).

گزارشگر ویژه همچنین تأکید کرده است که «شرکت ها ابزارهایی برای برخورد با محتوا به شیوه هایی مطابق با حقوق بشر دارند، که در برخی جهات دامنه وسیع تری نسبت به ابزارهای مورد استفاده کشورها دارند. این دامنه گزینه ها، به آنها امکان می دهد پاسخ های خود را متناسب با محتوای مسئله‌دار خاص، بر اساس شدت آن و عوامل دیگر، تنظیم کنند» (74/486/A، پاراگراف ۵۱).

در ارزیابی ضرورت و تناسب اقدامات بالقوه، هیئت موارد زیر را در نظر گرفت: (آ) این که صفحه خود را به عنوان منبع خبری معتبر معرفی کرده باشد؛ (ب) این که محتوا مستقیما به یک درگیری مسلحانه جاری مربوط باشد؛ (پ) آسیب پذیری غیرنظامیانی که در وسط آن جنگ به دنبال اطلاعات تأییدشده هستند؛ (ت) گسترش سریع و مستند محتوای فریبنده تولید شده توسط هوش مصنوعی در طول این جنگ (نگاه کنید به 31528-PC مؤسسه Alan Turing)؛ (ث) توزیع محتوای مشابه یا تقریبا یکسان بین پلتفرم ها؛ و (ج) انگیزه های مشارکت و درآمدزایی برای تولید رسانه های دستکاری شده در طول جنگ ها.

«هیئت» معتقد است قرار دادن برچسب رسانه دستکاری شده با «هوش مصنوعی پرخطر» روی محتوا، الزامات ضرورت و تناسب را برآورده می‌کند و از عدم انجام این کار توسط Meta نگران است. این روش بسیار کمتر از حذف آن مداخله‌گر است، چون فریب احتمالا منجر به آسیب فوری نخواهد شد. در حال حاضر، این نوع برچسب‌گذاری، ماهیت اطلاع‌رسانی خواهد داشت و منجر به کاهش رتبه یا حذف از موارد توصیه شده نخواهد

شد. Meta به کاربرانی که سعی می‌کنند محتوا را با این برچسب به اشتراک بگذارند، یک پنجره پاپ آپ نشان می دهد. یک برچسب می‌تواند تأثیر فریب را بر کاربرانی که به دنبال اطلاعات دقیق آنلاین درباره این جنگ هستند، کاهش دهد.

موارد قبلی «هیئت» اثبات می‌کنند که پخش رسانه های دستکاری شده بدون برچسب ممکن است اعتماد به اصالت محتوای پلتفرم را به صورت گسترده تر تضعیف کند. این موضوع به ویژه در زمان درگیری های مسلحانه صادق است، که رسانه های دستکاری شده که نقض قوانین بین المللی انسانی را نشان می دهند، ممکن است اعتماد به این چارچوب های حقوقی و حمایت هایی را که از غیرنظامیان فراهم می‌کنند کاهش دهند. Meta برای انجام مسئولیت های حقوق بشری اش، باید گزارش محتوا به رده‌های بالاتر را بدون تکیه بر طرف‌های سوم خارجی انجام می‌داد، تا برچسب‌گذاری محتوا بدون تأخیر انجام می‌شد.

مکانیزم های فعلی برای چسباندن حتی برچسب استاندارد اطلاعات هوش مصنوعی به ویدیو (افشای خود کاربر یا ارجاع به تیم سیاست محتوا)، نه به اندازه کافی قوی و نه جامع هستند تا با مقیاس و سرعت محتوای تولیدشده توسط هوش مصنوعی، به ویژه در شرایط بحرانی یا جنگ که تعامل در پلتفرم افزایش می‌یابد، مقابله کنند. «هیئت» اشاره می‌کند که سیستمی که بیش از حد به افشای خود کاربر درباره استفاده از هوش مصنوعی و شکایت‌های برجسته‌شده (که به ندرت رخ می دهد) برای برچسب‌گذاری درست این محتوا وابسته است، نمی‌تواند با چالش های موجود در محیط کنونی را رویارویی کند.

برخی اعضا اشاره کردند که تعریف Meta از «آسیب فیزیکی یا خشونت قریب الوقوع» شامل راه های مختلف که محتوای فریبنده تولید شده توسط هوش مصنوعی می‌تواند تأثیرات اجتماعی کمتر مستقیم اما جدی در زمان درگیری مسلحانه داشته باشد، نمی‌شود. این امر مثلا ممکن است دسترسی به اطلاعات معتبر مورد نیاز برای پاسخگو کردن طرف‌های دولتی و سایر طرف ها را تضعیف کند، آسیب پذیری جمعیت ها را در برابر دستکاری افزایش دهد و اشکال دیگر نفوذ فریبنده را ممکن سازد. برای این اعضای هیئت، برچسب‌گذاری «رسانه های دستکاری شده» که همراه با کاهش رتبه یا حذف از موارد توصیه شده نباشد، برای رفع این نگرانی ها کافی نیست. محدود کردن دسترسی به این نوع محتوا در لحظات پرخطر مانند درگیری های مسلحانه، مداخله ای ضروری و متناسب خواهد بود. هیئت اذعان می‌کند که ارزیابی نادرست از سوی راستی‌آزمایان طرف سوم درباره این محتوا نیز منجر به چنین نتیجه ای می‌شد.

در سراسر این صنعت، چالش های فنی عمده برای تضمین برچسب‌گذاری دقیق و مداوم محتوای تولید شده توسط هوش مصنوعی وجود دارد. Meta به «هیئت» توضیح داد که از شاخص های استاندارد صنعت – فراداده ای که ابزارهای هوش مصنوعی مولد اغلب در محتوا جاسازی می‌کنند – برای برچسب‌گذاری خودکار تصاویر ثابت

استفاده می‌کند. با این حال، همه ابزارهای هوش مصنوعی مولد در حال حاضر فراداده لازم برای برچسب‌گذاری را ضمیمه نمی‌کنند. حتی اگر ابزاری فراداده را ضمیمه کند، کاربران می‌توانند به راحتی آن را از محتوا حذف کنند و سپس در شبکه های اجتماعی به اشتراک بگذارند. مکانیزم های محدود Meta ممکن است بازتاب دهنده این چالش های کنونی باشد؛ با این حال، مسئولیت شرکت است که فعالانه به فن‌آوری های در حال تحول سریع پاسخ دهد. این مسئله فراتر از محتوای فریبنده است، چون این محدودیت ها توانایی کاربران را برای تأیید اصالت همه اطلاعات محدود می‌کند. با پیشی گرفتن حجم و کیفیت ویدئو و صدای تولید شده توسط هوش مصنوعی از ابزارهای شناسایی و برچسب‌گذاری موجود، این موانع بیشتر خواهند شد. «هیئت» اکیدا به Meta توصیه می‌کند بهبود مکانیزم های شناسایی و برچسب‌گذاری خود را در اولویت بگذارید تا همه اشکال محتوای تولید شده توسط هوش مصنوعی را در پلتفرم هایش بهتر ثبت کند، کاربران را به درستی مطلع کند که ممکن است با رسانه های دستکاری شده تعامل داشته باشند و تمرکز را بر محتواهایی که بیشترین ریسک را دارند معطوف کند.

هیئت حضور Meta در کمیته راهبری C2PA را به رسمیت می‌شناسد. C2PA بیان می‌کند که با تحول سریع اشتراک گذاری اطلاعات، ردیابی منشأ رسانه ها حیاتی است. این پروتکل یک استاندارد فنی باز برای تعیین منشأ و ویرایش محتوای دیجیتال ارائه می دهد. گزارش هایی مبنی بر اینکه Meta به صورت مداوم استانداردهای C2PA را – حتی روی محتوای تولید شده توسط هوش مصنوعی از ابزارهای خودش – اجرا نمی‌کند نگران کننده هستند. یک تحقیق جدید نشان داد که هنگام بررسی توسط ابزارهای C2PA، تنها بخشی از تصاویر و ویدئوهای تولید شده توسط ابزارهای هوش مصنوعی Meta اعتبارنامه محتوا را ارائه داده و برچسب‌گذاری مناسبی دریافت کرده اند.

محتوای مورد ظاهرا ابتدا از پلتفرم Meta در TikTok منتشر شده، سپس به سرعت در پلتفرم های مختلف به اشتراک گذاشته شده است، طوری که با وجود گزارش AFP درباره نادرستی محتوای مشابه، پست های مشابه در Facebook و Instagram و X منتشر شده است. در نظرات عمومی، نیاز به همکاری قوی بین پلتفرم ها در زمان درگیری مسلحانه برای کاهش سرعت گسترش محتوای فریبنده تولید شده توسط هوش مصنوعی تأکید شده بود.

همان طور که پیش تر توضیح داده شد، موارد تعیین CPP و رویدادهای پرطرفدار باید به Meta اجازه می‌داد تا حمایت مؤثرتر از راستی‌آزمایان طرف سوم را در طول بحران تضمین کند. به ویژه، پخش امتیازها به دسته وسیع تری از ویدیوهای بسیار مشابه می‌توانست آسیب های بالقوه را به صورت قابل توجه کاهش دهد، از جمله با تنزل درجه آن. این مورد، ناکارآمدی های رویکرد فعلی Meta را در طول درگیری های مسلحانه برجسته می‌کند و

نگرانی هایی را که هیئت در موارد قبلی و در زمینه های مختلف ابراز کرده بود، تشدید می‌کند (تصمیم مربوط به تصاویر اعتراضات همراه با شعارهای طرفدار دوترته).

6. تصمیم هیئت نظارت

«هیئت» تصمیم Meta برای باقی گذاشتن محتوا بدون برچسب هوش مصنوعی پرخطر را لغو می‌کند.

7. توصیه‌ها

A. سیاست محتوا

اطلاعات نادرست

1. برای اطمینان از بررسی سریع اطلاعات نادرست که منجر به خطر آسیب فیزیکی یا خشونت قریب الوقوع در بحران ها می‌شود، Meta باید استاندارد جامعه اطلاعات نادرست را اصلاح کند تا اطمینان حاصل شود که اجرای این قانون وابسته به سیگنال های شرکای خارجی نیست. باید اهرمی‌تحت پروتکل سیاست بحران وجود داشته باشد تا منابع را برای شناسایی به موقع و فعالانه این نوع محتوای نقض کننده که با تخصص داخلی پشتیبانی می‌شود، تخصیص دهد تا محتوا تحت این سیاست شناسایی، بازبینی و رسیدگی شود (از جمله چسباندن برچسب ها تحت سیاست رسانه دستکاری شده و بررسی انتشار حساب ها و صفحاتی که نشانه هایی از سوءاستفاده از مشارکت را نشان می دهند).

زمانی که Meta سیاست اطلاعات نادرست خود را به روزرسانی کند تا این الزامات را برای دسته آسیب فیزیکی یا خشونت منعکس کند، «هیئت» این موضوع را در نظر خواهد گرفت.

محتوای تولید شده توسط هوش مصنوعی

2. Meta برای کمک به افزایش اعتماد به اطلاعات در پلتفرم های Meta، باید یک استاندارد جامعه برای محتوای تولیدشده توسط هوش مصنوعی ایجاد کند که جدا از استاندارد جامعه اطلاعات نادرست باشد. استاندارد جدید جامعه باید اطلاعات جامع درباره حفظ منشأ ارائه کند (یعنی ثبت اطلاعات دقیق درباره تاریخچه یک محتوای دیجیتال)، پروتکل های برچسب‌گذاری هوش مصنوعی و قوانین خودافشایی.

هنگام انتشار استاندارد جدید جامعه از سوی Meta به صورت خاص درباره محتوای تولیدشده توسط هوش مصنوعی، «هیئت» این موضوع را در نظر خواهد گرفت.

3. برای بهبود وضوح قوانین، Meta باید توضیح روشنی درباره جریمه های عدم افشای محتوای دیجیتال یا تغییر یافته منتشر کند. باید معیارهایی برای جریمه ها ارائه دهد و مشخص کند کدام ویژگی های حساب به این دلیل محدود شده اند و برای چه مدت.

زمانی که Meta استاندارد جامعه را به روزرسانی کند تا جزئیات جریمه را درج کند و راهنمای اصلاح شده را در مرکز شفافیت عمومی خود در دسترس قرار دهد، «هیئت» این موضوع را در نظر خواهد گرفت.

B. اجرای قانون

منشأ

4. برای اطمینان از اینکه کاربران می‌توانند محتوای تولید شده توسط هوش مصنوعی را به صورت قابل اعتماد شناسایی کنند، Meta باید اعتبارنامه های محتوا (ارائه شده توسط ائتلاف منشأ و اصالت محتوا) را در مقیاس وسیع اجرا کند و اطمینان حاصل کند که هر زمان که جزئیات منشأ در دسترس باشد، به وضوح و به صورت مداوم برای کاربران قابل مشاهده و قابل دسترسی باشند. منشأ نباید فقط به صورت داخلی قابل شناسایی باقی بماند یا محدود به سیستم های بک اِند باشد.

زمانی که Meta گزارشی ارائه می دهد که تغییرات ایجاد شده در رابط ها و محصولاتش را توضیح می دهد تا تضمین کند که اعتبارنامه های محتوا به صورت مداوم و واضح در صورت امکان به کاربران نشان داده می‌شوند، «هیئت» این امر را اجرا شده در نظر خواهد گرفت.

شناسایی و برچسب‌گذاری

5. برای بهبود دقت تشخیص و برچسب‌گذاری، Meta باید در ابزارهای شناسایی قوی تری برای محتوای چندفرمتی تولیدشده توسط هوش مصنوعی سرمایه گذاری کند (صدا، صدا-تصویر و تصویر) و محتوا. ابزارها باید به تیم های رده بالای بررسی گزارش کمک کنند تا ترندهای محتوای هوش مصنوعی مولد، از جمله آسیب های احتمالی محتوای فریبنده در شرایط بحرانی را بهتر شناسایی کنند.

زمانی که شرکت تأیید کند ابزارهای قوی تری تصویب شده اند و داده های شفافیت درباره عملکرد این ابزارها را به اشتراک گذاشت، «هیئت» این توصیه را اجراشده در نظر خواهد گرفت. این یافته ها باید بر اساس زبان و کشور و اینکه پروتکل سیاست بحران فعال شده است یا نه، تفکیک شوند. این داده ها باید بازتاب دهنده دوره های زمانی مشابه قبل و بعد از معرفی این تغییرات باشند.

6. برای اطمینان از برچسب‌گذاری دقیق تر، Meta باید اطلاعات منشأ و واترمارک های نامرئی را به محتوای تولید شده توسط ابزارهای هوش مصنوعی Meta متصل کند تا به صورت مداوم در سراسر پلتفرم ها قابل شناسایی و برچسب‌گذاری باشند. این باید شامل اعمال اعتبارنامه های محتوا در نقطه ایجاد و نیز استفاده از نشانگرهای استاندارد صنعتی برای نسبت دادن به تمام محتوای تولید شده توسط Meta AI باشد.

زمانی که شرکت گزارشی به «هیئت» داد درباره این که هوش مصنوعی Meta چگونه داده‌های منشاء و واترمارک‌های نامرئی را به محتوای به اشتراک گذاشته شده در پلتفرم اضافه می‌کند، «هیئت» این توصیه را اجرا شده تلقی خواهد کرد.

7. برای اینکه استفاده از برچسب های هوش مصنوعی ریسک بالا و پرخطر در محتوای فریبنده را منسجم تر کند، Meta باید مسیرهایی برای افزودن این برچسب ها به محتوا به صورت بسیار مکررتر، با کمک کانال های واضح تر برای انتقال شکایت از سیستم های خودکار و بازبینی با حجم بالا ایجاد کند، تا این نوع برچسب‌گذاری بتواند با حجم بسیار بالاتر انجام شوند.

زمانی که مسیرهای جدید برای گزارش به رده‌های بالاتر به منظور افزودن برچسب‌های هوش مصنوعی ریسک بالا و پرخطر به محتوا وجود داشته باشد، و Meta گزارشی درباره حجم این برچسب های افزوده شده در سال 2026، به تفکیک سه ماهه به «هیئت» داد، «هیئت» این توصیه را اجراشده تلقی خواهد کرد. نبود مبنا (یعنی حجم کل محتوای بدون برچسب که این آستانه را ندارد) نباید مانعی برای ارائه این اطلاعات به «هیئت» باشد.

***یادداشت فرآیند اجرایی:**

- تصمیم‌های «هیئت نظارت» توسط گروهی مشتمل بر پنج «عضو» اخذ می‌شوند و بر اساس رأی اکثریت کل «هیئت» تصویب می‌گردند. تصمیم‌های «هیئت» الزاماً بیان‌کننده نظرات همه «اعضا» نیست.

- مطابق با منشور «هیئت نظارت»، این «هیئت» می‌تواند موارد زیر را بازبینی کند: درخواست تجدیدنظر کاربرانی که محتوای «متای» آن‌ها برداشته شده است، درخواست تجدیدنظر کاربرانی که محتوایی را

گزارش کرده‌اند و «متا» آن‌ها را برنداشته است و تصمیم‌هایی که «متا» به آن ارجاع می‌دهد (ماده ۲ منشور، بخش ۱). «هیئت» اختیار الزام آور برای تأیید یا لغو تصمیمات محتوای Meta داد (منشور) ماده 3، بخش 5؛ ماده 4 منشور). «هیئت» می‌تواند توصیه‌های غیرالزام‌آوری ارائه کند که «متا» باید به آن‌ها پاسخ دهد (ماده ۳ منشور، بخش ۴؛ ماده ۴). وقتی «متا» برای اقدامی برپایه توصیه‌ها متعهد شود، «هیئت» بر پیاده‌سازی آن نظارت می‌کند.

- درخصوص تصمیم این پرونده، تحقیق مستقلی ازطرف «هیئت» انجام گرفت. Duco Advisors، شرکتی مشاوره‌ای که بر تعاملات ژئوپولیتیک، اعتماد و امنیت و فناوری تمرکز دارد، در این زمینه به «هیئت» کمک کرد.