فصلنامه شناخت پژوهی مطالعات سیاسی
سال ۲، شماره ۲(۵)،۸۳-۶۵
تاریخ دریافت: ۱۴۰۴/۰۲/۲۹
تاریخ پذیرش: ۱۴۰۴/۰۴/۱۳

Journal of cognitive research in political studies
Vol.2, No2 (5), 65-83
Reseived Date: 2025/05/19
Accept Date: 2025/07/04

DOI: 10.20241403/CRPS.2502.1034.2.4.3

استفاده از هوش مصنوعی در تعدیل محتوا

پژمان الهامی طالشمیکائیل[۱] | امیر کردکریمی[۲] | ولی شیرپور[۳] | حسام اوروجی[۴]

شماره ۲(۵)

سال ۲
تابستان ۱۴۰۴

مقاله پژوهشی

تاریخ دریافت:
۱۴۰۴/۰۲/۲۹
تاریخ پذیرش:
۱۴۰۴/۰۴/۱۳
صص: ۸۳-۶۵

چکیده

در این پژوهش استفاده از هوش مصنوعی در تعدیل محتوا برای مقابله با افراط گرایی‌های خشونت‌آمیز موجود در فضای مجازی، با دیدی انتقادی مورد بررسی قرار گرفت و در راستای تحقق این هدف از روش گردآوری داده‌ها به صورت کتابخانه‌ای بهره گرفته شد و ابزار گردآوری اطلاعات در این روش ترجمه، تلخیص، نقل قول مستقیم و غیر مستقیم است. در این راستا، تمرکز این پژوهش بر روی اندازه‌گیری دقت هوش مصنوعی در تعدیل محتوا، موارد وقوع مثبت و منفی کاذب و نقض آزادی بیان و دموکراسی استوار است و به این شکل استدلال شده است که استفاده از تکنولوژی حذف خودکار محتوا اثربخشی محدودی دارد و استفاده از هوش مصنوعی در تعدیل محتوا می‌تواند منجر به نقض اصولی چون: آزادی بیان و دموکراسی گردد. در این پژوهش تأکید شده است که استفاده از روش پلتفرم‌زدایی که در آن شناسایی محتوای خشونت‌آمیز و افراطی، توسط هوش مصنوعی انجام می‌گیرد ولی تصمیم نهایی حذف از پلتفرم معمولاً توسط مدیران یا مالکان پلتفرم اتخاذ می‌شود، نسبت به حذف محتوا که در آن شناسایی محتوای خشونت‌آمیز و افراطی، توسط هوش مصنوعی و به صورت خودکار و بدون تصمیم نهایی مدیران یا مالکان پلتفرم است، جهت مقابله با افراط‌گرایی خشونت آمیز آنلاین اثربخشی بیشتری دارد.

کلمات کلیدی: هوش مصنوعی، تعدیل محتوا، افراط گرایی خشونت آمیز، حقوق بشر، آزادی بیان

[۱] دانش آموخته ارشد علوم سیاسی، دانشگاه پیام نور تهران، ایران (نویسنده مسئول)
pejhmanelhami@gmail.com
[۲] دکتری تخصصی علوم سیاسی، گرایش اندیشه سیاسی، دانشگاه آزاد اسلامی واحد تبریز، تبریز، ایران.
am.kordkarimi@yahoo.com
[۳] مربی گروه حقوق دانشگاه پیام نور، تهران، ایران. shirpourvali@pnu.ac.it
[۴] دانشجوی دکتری علوم سیاسی جامعه شناسی سیاسی دانشگاه علامه طباطبایی تهران.
hessamurujii@gmail.com

استناد: الهامی طالشمیکائیل، پژمان ؛ کردکریمی، امیر؛ شیرپور، ولی و اوروجی، حسام. (۱۴۰۴). استفاده از هوش مصنوعی در تعدیل محتوا. شناخت پژوهی مطالعات سیاسی، ۲(۲)،۸۳-۶۵. doi: 10.20241403/CRPS.2505.1050.2.5.4

Elhami Taleshmikaeil, P., Kord Karimi, A., Shipour, V. and Oroji, H. (2025). The application of artificial intelligence in content moderation. Cognitive research of political studies, 2(2),65-83. doi: 10.20241403/CRPS.2505.1050.2.5.4

مقدمه

تغییرات مداوم اینترنت چالش‌های اساسی را برای دولت‌ها در سراسر جهان در تلاش برای مقابله با افراط‌گرایی خشونت‌آمیز ایجاد کرده است (Piazza & Guler, 2019: 31). در این میان گروه‌های افراطی خشن تمایل به بهره برداری از اینترنت، به‌ویژه شبکه‌های اجتماعی دارند تا از طریق بهره‌بردای از ویژگی‌های این شبکه‌ها روایت‌های افراطی خود را به شکلی سریع و در مقیاسی بزرگ انتشار دهند، هدف آن‌ها از انتشار این روایات جذب بازیگران افراطی، تأمین مالی جهت گسترش ایدئولوژی‌های مضر و برنامه‌ریزی برای حملات بالقوه؛ است (United Nations Office on Drugs and Crime, 2012: 54).

حضور فزاینده و رو به رشد گروه‌های افراطی دارای رفتار خشونت آمیز در شبکه‌های اجتماعی همچون ایکس (توئیتر) و فیسبوک بسیاری از دولت‌ها را تحت فشار قرار داده است تا پاسخ‌های لازم برای مقابله با افراط‌گرایی‌هایی خشونت‌آمیز در فضای مجازی را مجدداً مورد بررسی قرار دهند در این بین بسیاری از کشورها اقدامات سخت‌تری را برای حذف محتوا اتخاذ کرده‌اند (GuhL et al., 2020: 13). در این زمینه پژوهش‌هایی نیز صورت گرفته است که در جدول ذیل به عنوان پیشینه‌ی پژوهش ارائه می‌گردد:

ردیف	عنوان	سال	پژوهشگر	نتایج تحقیق
۱	فناوری‌های نوظهور و تأثیر آن‌ها بر روابط بین‌الملل و امنیت جهانی[1]	۲۰۱۸	ایوان دانیلین[2]	در سطح بین‌المللی گنجاندن هوش مصنوعی و یادگیری ماشین در امور دفاعی و امنیتی ممکن است منجر به ایجاد یک رقابت تسلیحاتی میان کشورها شود که این رقابت نه تنها منجر به توسعه‌ی خطرآفرین تسلیحات مختلف می‌گردد بلکه ثبات استراتژیک جهانی را نیز خدشه‌دار می‌نماید.

[1] Emerging Technologies and Their Impact on International Relations and Global Security
[2] Ivan Danilin

۲	پیش‌بینی هوش مصنوعی و مبارزه با تروریسم[1]	۲۰۱۹	مک کندریک[2]	اولین ریسک در بکارگیری هوش مصنوعی توسط دولت‌ها، ریسکهای مرتبط با سوگیری‌های مغرضانه است که ممکن است در نهایت منجر به ایجاد خطا در نتایج شوند.
۳	پیامدهای هوش مصنوعی در امنیت ملی: درک مسائل امنیتی و چالش های اخلاقی[3]	۲۰۲۱	یو و کرول[4]	سیستم‌های هوش مصنوعی توسط انسان‌هایی ایجاد می‌شوند که می‌توانند به عنوان مثال، به دلیل اعتقادات شخصی، مغرضانه رفتار نماید هنگامی که یک دانشمند داده که یک مدل هوش مصنوعی ایجاد می‌کند، به عنوان مثال، در نظر می‌گیرد که تمام محتوای تروریستی همیشه توسط افراد مسلمان منتشر می‌شود، این نوعی القا و تعصب انسانی است و ممکن است این تعصب و سوگیری در مدلی که آن فرد طراحی می‌کند اثرگذار باشد.

در پژوهش‌های فوق به صورت پراکنده به پیامدهای بهره‌گیری از هوش مصنوعی پرداخته شده است که می‌تواند راه گشا و مثمر ثمر واقع شود با این حال در زمینه‌ی مقابله با افراط‌گرایی خشونت‌آمیز آنلاین مطلبی ارائه نشده و از این منظر این پژوهش دارای نوآوری است. همچنین هدف این پژوهش که به روش گردآوری داده‌های کتابخانه‌ای انجام خواهد گرفت و ابزار گردآوری اطلاعات در آن ترجمه، تلخیص، نقل قول مستقیم و غیرمستقیم خواهد بود، «ارزیابی انتقادی استفاده از هوش مصنوعی در تعدیل محتوا برای مقابله با افراط‌گرایی خشونت‌آمیز آنلاین، با تمرکز بر سنجش دقت هوش مصنوعی در تعدیل محتوا، وقوع موارد مثبت و منفی کاذب و نقض آزادی بیان و دموکراسی، استفاده از معیارهای پلتفرم‌زدایی[5] (حذف و ممنوعیت یک کاربر در رسانه‌های ارتباط جمعی) و بررسی فرایند مهاجرت راست‌های افراطی[6] از رسانه‌های اجتماعی اصلی

[1] Artificial intelligence prediction and counterterrorism.
[2] McKendrick
[3] Implications of AI in National Security: Understanding the security issues and ethical challenges
[4] Yu, Carroll
[5] De-platforming
[6] Far-right

و متداول به پلتفرم‌های فناوری جایگزین[1]» است. در این راستا این پرسش قابل طرح است که: «اثر بخشی حذف خودکار محتوا برای مقابله با افراط‌گرایی خشونت‌آمیز آنلاین چگونه است؟» و به عنوان فرضیه می‌توان مطرح نمود: استفاده از حذف خودکار محتوا برای مقابله با افراط‌گرایی خشونت‌آمیز آنلاین از نظر اثربخشی محدود است، زیرا اطلاعات متنی را در نظر نمی‌گیرد که این مسأله باعث می‌شود بکارگیری این تکنولوژی در برخی موارد با خطاهایی مانند مثبت کاذب همراه باشد و نتواند به طور دقیق محتوای افراط‌گونه را مشخص کند. همچنین استفاده از هوش مصنوعی در تعدیل محتوا، اصول آزادی بیان و دموکراسی را نقض می‌کند و نقض این اصول ممکن است به دلیل فقدان شفافیت و مسئولیت پاسخگویی توسط شرکت‌های رسانه‌های اجتماعی تشدید شود و به نظر می‌رسد اقدامات پلتفرم‌زدایی ابزار مؤثرتری برای مقابله با افراط‌گرایی خشونت‌آمیز آنلاین است و بهتر آن است که استفاده از هوش مصنوعی در اقدامات پلتفرم‌زدایی برای افزایش شناسایی کاربران و شبکه‌های افراطی خشونت‌آمیز که در فضای مجازی فعایت می‌کنند مورد توجه قرار گیرد.

۱ـ تعاریف مفاهیم

در این بخش مواردی همچون افراط‌گرایی خشونت‌آمیز و مقابله با افراط‌گرایی خشونت‌آمیز بررسی خواهد شد:

۱ـ۱ـ افراط‌گرایی خشونت‌آمیز

در مورد تعریف افراط‌گرایی خشونت‌آمیز بین سیاستگذاران و دانشگاهیان اتفاق نظر وجود ندارد. دولت بریتانیا افراط‌گرایی را به‌عنوان «مخالفت صریح با ارزش‌های بنیادین بریتانیا از جمله دموکراسی، حاکمیت قانون، آزادی فردی و احترام متقابل و تحمل ادیان و عقاید مختلف» تعریف کرده است (Home Office, 2015: 14). با این حال، این تعریف از افراط‌گرایی با انتقادات قابل

به افراد، گروه‌ها و حزب‌های محافظه کار ارتجاعی، هوادار قدرت مطلقه، سنت گرا و فاشیست جناح راست گفته می‌شود که خواستار تثبیت وضع موجود و حفظ امتیازات طبقات ممتاز و مخالف اصلاحات اساسی هستند (jaafarifar, 2023: 7).

[1] رسانه‌های اجتماعی جایگزین، رسانه‌های اجتماعی هستند که در میان افرادی که عقاید افراطی دارند، مانند راست‌گرایان یا راست افراطی، محبوب شده‌اند. در حالی که اکوسیستم فناوری جایگزین منحصر به جناح راست نیست، اغلب بستری را برای کسانی فراهم می‌کند که به دلیل نقض شرایط خدمات، سخنرانی علیه یک قشر یا شخص، نفرت پراکنی یا تحریک خشونت، از رسانه‌های اجتماعی اصلی منع شده‌اند. رسانه‌های جایگزین در اواخر دهه ۲۰۱۰ و اوایل دهه ۲۰۲۰ محبوبیت زیادی پیدا کردند و شامل پلت‌فرم‌های اجتماعی می‌شوند که به‌طور خاص برای گروه‌های کاربران افراطی ایجاد شده‌اند و سخت‌گیری‌های کمتری را نسبت به محتوا اعمال می‌دارند. تلگرام به گونه‌ای جزء این شبکه‌های جایگزین قلمداد می‌شود (Stephens et al., 2018: 7).

توجهی مواجه شده است. یکی از انتقادات اصلی این است که تعریف دولت بریتانیا از ارزش‌های بنیادین مبهم است (Lowe, 2017: 4). که این مسأله می‌تواند منجر به بدنام سازی جوامع خاصی گردد (Vincent & Hunter-Henin, 2018: 3). همچنین به این دلیل که این تعریف گروه‌های راست افراطی را که نسبت به نژادها و قومیت‌های مختلف نگاه متعصبانه دارند، نادیده می‌گیرد (Allen, 2021: 7). به گفته‌ی کندنانی[1] و هایس[2] (۲۰۱۸) در تعریف افراط‌گرایی، عدم هماهنگی کلی وجود دارد، که این امر می‌تواند تأثیرات قابل توجهی بر ایجاد و اجرای سیاست های مقابله با افراط‌گرایی خشونت‌آمیز (CVE) داشته باشد (Kundnani, Hayes, 2018:1).

باید گفت، بین تعاریف افراط‌گرایی دو دیدگاه وجود دارد برخی بر اقدامات بازیگران برای رسیدن به یک هدف سیاسی تمرکز دارند و برخی بر مخالفت فعال با ارزش‌های اصلی جامعه تمرکز می‌کنند (Stephens et al., 2018: 10).

نئومن[3](۲۰۱۹) اشاره می‌کند که بیشتر تعاریف «افراط‌گرایی غیرخشونت‌آمیز» در مقایسه با تعاریف «افراط‌گرایی خشونت‌آمیز» که بر استفاده از روش‌های خشونت‌آمیز برای دستیابی به یک هدف سیاسی تمرکز دارند؛ دیدگاهی ایده‌آلیستی ارائه می‌کنند، با این حال حتی در این دیدگاه آرمانی و ایده‌آل بازهم بر ایدئولوژی افراطی تمرکز شده است.

در واقع ایجاد تمایز بین افراط‌گرایی خشونت‌آمیز و غیرخشونت‌آمیز بی‌اثر است، زیرا ممکن است فردی بتواند بدون مشارکت در اقدامات خشونت‌آمیز برای دستیابی به یک هدف سیاسی، دیدگاه‌های افراطی داشته باشد و تنها در صورت بروز فرصت دست به اعمال خشونت آمیز بزند (Schmid, 2014:7). با این حال و علی‌رغم موارد یاد شده، در این پژوهش بین افراط‌گرایی خشونت‌آمیز و غیرخشونت‌آمیز تمایز وجود خواهد داشت. زیرا ارزیابی انتقادی به طور خاص بر استفاده از هوش مصنوعی در تعدیل محتوا برای مقابله با افراط‌گرایی خشونت‌آمیز آنلاین متمرکز است و افراط‌گرایی خشونت‌آمیز مورد نظر در این پژوهش به عنوان «ترویج، چشم پوشی، توجیه یا حمایت از ارتکاب یک عمل خشونت‌آمیز جهت دستیابی به اهداف سیاسی، ایدئولوژیک، مذهبی، اجتماعی یا اقتصادی» تعریف می‌شود (FBI, n.d.: 1).

---

[1] Kundnani
[2] Hayes
[3] Neumann

۱ـ۲ـ مقابله با افراط‌گرایی خشونت‌آمیز

مقابله با افراط‌گرایی خشونت‌آمیز به عنوان طیفی از واکنش‌های خفیف و خودخواسته برای مقابله با افراط‌گرایی، برای نمونه، رادیکال‌سازی، تعریف می‌شود (Selim, 2016: 25). تعاریف مرتبط با مقابله با افراط‌گرایی خشونت‌آمیز عمدتاً حول این مفهوم متمرکز می‌شوند که «اصلاح، اعاده اعتبار، باز اجتماعی کردن و پیشگیری» افراط‌گرایی خشونت‌آمیز را کاهش می‌دهد (LaFree & Freilich,2019: 132).

۱ـ۳ـ پلتفرم‌زدایی

پلتفرم‌زدایی یا حذف از پلتفرم، اصطلاحی است که در عصر دیجیتال توجه زیادی را به خود جلب کرده است. این واژه به جلوگیری از به اشتراک گذاشتن نظرات یا محتوای افراد، گروه‌ها یا نهادها در یک پلتفرم دیجیتال اشاره دارد و معمولاً توسط مدیران یا مالکان پلتفرم در پاسخ به نقض قوانین یا دستورالعمل‌های پلتفرم توسط فرد یا گروه انجام می‌شود. پلتفرم‌زدایی یک مسئله چالش برانگیز است و چالش‌ها اغلب در مورد ایجاد تعادل بین آزادی بیان و لزوم حفظ یک محیط آنلاین امن و محترمانه ایجاد می‌شود.

پلتفرم‌زدایی ریشه در دهه‌های ۱۹۷۰ و ۱۹۸۰ دارد، در آن سال‌ها پلتفرم‌زدایی به‌عنوان تاکتیکی برای جلوگیری از سخنرانی افراد یا گروه‌هایی که دیدگاه‌های تبعیض‌آمیز را در رویدادهای عمومی ترویج می‌کردند، به کار گرفته می‌شد. این امر اغلب از طریق اعتراضات یا اقدامات مستقیم، مانند اشغال مکانی که قرار بود رویداد در آن برگزار شود، محقق می‌شد. با ظهور اینترنت و ظهور پلتفرم‌های دیجیتال، مفهوم عدم استفاده از پلتفرم به پلتفرم‌زدایی تبدیل شده است. تفاوت اساسی این است که حذف پلتفرم به صراحت به حذف افراد یا گروه‌ها از پلتفرم‌های دیجیتال به جای پلتفرم‌های فیزیکی اشاره دارد. این امر، پلتفرم‌زدایی را به ابزاری بسیار قدرتمندتر تبدیل کرده است، زیرا پلتفرم‌های دیجیتال اغلب دسترسی و نفوذ بیشتری نسبت به هر مکان فیزیکی دارند (Llansó et al., 2020: 6).

هفت دلیل مهم برای پلتفرم‌زدایی عبارت‌اند از (Wakefield, 2021: 10):

- نقض دستورالعمل‌ها: پلتفرم‌های دیجیتال اغلب افراد، گروه‌ها یا نهادها را به دلیل نقض دستورالعمل‌های تعیین‌شده، از پلتفرم خود حذف می‌کنند. این دستورالعمل‌ها معمولاً بر محتوا و رفتار قابل قبول تأکید دارند و نقض آن‌ها می‌تواند منجر به حذف از پلتفرم گردد.

– انتشار اطلاعات نادرست یا سخنان نفرت‌پراکن: یکی از دلایل اصلی حذف از پلتفرم، انتشار اطلاعات نادرست، مضر یا سخنان نفرت‌پراکن است. پلتفرم‌ها این اقدام را برای جلوگیری از انتشار محتوایی که می‌تواند خشونت، تبعیض یا آسیب عمومی را تحریک کند، انجام می‌دهند.

– مشارکت در آزار و اذیت یا قلدری: مشارکت مداوم در فعالیت‌های آزار و اذیت یا قلدری، زمینه مشترکی برای حذف پلتفرم است. این اقدام برای محافظت از ایمنی و رفاه کاربران انجام می‌شود. ترویج فعالیت‌های غیرقانونی: حمایت یا تسهیل فعالیت‌های غیرقانونی می‌تواند منجر به حذف از پلتفرم شود. این فعالیت‌ها شامل ترویج خشونت، مصرف مواد مخدر غیرقانونی یا سایر اقدامات مجرمانه می‌شود.

– نقض حقوق مالکیت معنوی: پلتفرم‌ها ممکن است کاربرانی را که به طور مداوم مطالبی را به اشتراک می‌گذارند و در طی آن قوانین مالکیت معنوی را نقض می‌کنند، حذف کنند.

– نقض‌های امنیتی و رفتارهای متقلبانه: کاربرانی که درگیر فعالیت‌های متقلبانه یا ایجاد خطرات امنیتی مانند هک یا فیشینگ هستند، از پلتفرم حذف می‌شوند.

– جعل هویت یا ارائه اطلاعات نادرست: حذف از پلتفرم زمانی اتفاق می‌افتد که کاربری در حال جعل هویت دیگران یا ارائه اطلاعات نادرست از هویت خود برای فریب یا گمراه کردن سایر کاربران باشد.

۲ـ بررسی مهم‌ترین اقدامات دولت‌ها در زمینه‌ی تعدیل محتوا

دولت آلمان، قانون (NetzDG)[1] را در سال ۲۰۱۷ به تصویب رساند. این قانون پس از آن تصویب شد که پست‌هایی با محتوای جعلی، نژادپرستانه و ضداسلام در شبکه‌های اجتماعی در سال‌های ۲۰۱۶ و ۲۰۱۷ به‌صورت گسترده در آلمان منتشر شد. بر اساس این قانون چنانچه شبکه‌های اجتماعی ظرف مدت بیست‌وچهار ساعت[2] در ارتباط با پست (پیام)های حاوی اقدام علیه نظم عمومی و فراخوان به ارتکاب جرم و تأیید آن (تحریک عمومی به اعمال مجرمانه)، نفرت و انزجار، جرم و جنایت، تهمت و افتراء، اخبار کذب، جرائم تروریستی شامل تشکیل گروه‌های تروریستی، توهین

[1] Network Enforcement Act (Netzwerkdurchsetzungsgesetz)

[2] تا پیش از تصویب این قانون، این تنها شرکتها بودند که درباره حذف محتویات مجرمانه تصمیم می گرفتند. بموجب این قانون، محتویاتی که در ماهیت مجرمانه آنها تردیدی نیست، باید ظرف بیست و چهار ساعت پس از دریافت شکایت پاک شوند اما در مواردی که درباره مجرمانه‌بودن آنها تردید وجود دارد، یک هفته فرصت برای حذف این گونه مطالب از سوی قانونگذار در نظر گرفته شده است.

به پیروان مذاهب، بددهنی و تشویق به خشونت و مزاحمت جنسی اقدام نکنند و مصادیق مجرمانه را از شبکه‌های خود حذف نکنند باید خود را برای پرداخت تا مبلغ پنجاه میلیون یورو (پنجاه‌وهشت میلیون دلار) جریمه آماده کنند. طبق این قانون، لازم است رویه‌ای مؤثر و شفاف برای رسیدگی به شکایت‌های برداشتن محتوای غیرقانونی طبق تعریف قوانین اجرایی شبکه‌های اجتماعی ارائه شود و شبکه‌های اجتماعی ملزم هستند هر شش ماه یک‌بار گزارش شفاف‌سازی منتشر کنند. با این‌حال، به‌دلیل اولویت داشتن «قانون خدمات دیجیتالی اتحادیه اروپا»[1]، قوانین اجرایی شبکه‌های اجتماعی از ۲۵ اوت ۲۰۲۳ به بعد بر YouTube اعمال نشده است (Miller, 2017: 3).

علاوه بر این، دولت بریتانیا سند آسیب‌های آنلاین[2] را در سال ۲۰۱۹ منتشر کرد که در آن در مورد لزوم معرفی یک چارچوب قانونی جایگزین برای مقابله با افراط‌گرایی‌های فضای آنلاین تأکید شده بود. در این سند دولت پیشنهادهایی را در راستای مقابله با آسیب‌های آنلاین ارائه می‌کرد. در بیانیه این سند که توسط وزیر کشور و وزیر فرهنگ، رسانه، ورزش و امور دیجیتال ارائه شده بود؛ محتوای آسیب‌زا به‌عنوان عامل تضعیف‌کننده مزایای انقلاب دیجیتال قلمداد می‌شد (HM Government, 2019: 1).

در سال ۲۰۲۱، دولت بریتانیا پیش‌نویس یک لایحه ایمنی آنلاین[3] را تهیه کرد که برای عدم حذف محتوای مضر، از جمله تصاویر افراطی و تبلیغات، جریمه‌هایی تا سقف ۱۸ میلیون پوند یا ۱۰ درصد از گردش مالی سالانه شرکت‌های رسانه‌های اجتماعی، در نظر گرفته بود تا بدین طریق از امنیت کاربران فضای مجازی محافظت کند (Wakefield, 2021: 2). همچنین در این زمینه در ۲۶ اکتبر ۲۰۲۳ قانون ایمنی آنلاین[4] در انگلستان تصویب شد؛ قانونی که اختیارات جدیدی به آفکام[5]، نهاد تنظیم‌گر ارتباطات در انگلستان، می‌دهد. در این قانون، محتوای غیرقانونی در بیش از ۱۰ مورد از جمله جرائم تروریستی، جرائم استثمار و سوءاستفاده جنسی از کودکان، جرائم نفرت‌پراکنی، جرائم شدید پورنوگرافی، تقلب و جرائم خدمات مالی، جرائم مداخله خارجی و... تعریف می‌شود و وظایفی برای سرویس‌های اینترنتی و پلتفرم‌های آنلاین مشمول این قانون در راستای مقابله با جرائم

---

[1] Digital Services Act (DSA)
مقرراتی در قانون اتحادیه اروپا برای به روز رسانی دستورالعمل تجارت الکترونیکی ۲۰۰۰ در مورد محتوای غیرقانونی، تبلیغات شفاف و اطلاعات نادرست است.

[2] Online Harms White Paper
[3] Online Safety Bill
[4] Online Safety Act
[5] Ofcom

مذکور مقرر شده است که می‌توان به مواردی از جمله ارزیابی ریسک آسیب‌های ناشی از محتوای غیرقانونی، ارزیابی ریسک آسیب‌های ناشی از محتوای مضر برای کودکان، شفاف‌سازی نحوه محافظت از امنیت کاربران و مکتوب کردن آن جهت مشاهده عموم، تسهیل گزارش‌دهی محتوای غیرقانونی و محتوای مضر توسط کاربران، اقدام مؤثر در راستای مدیریت و کاهش ریسک‌های شناسایی‌شده از جمله حذف محتوای غیرقانونی و اقدام مناسب جهت جلوگیری از مواجهه کاربران با آن و مواردی از این قبیل اشاره کرد.

باید گفت، با توجه چارچوب‌های نظارتی سخت‌گیرانه‌تری که توسط برخی دولت‌ها در مورد حذف محتوا اتخاذ شده است، شرکت‌های رسانه‌های اجتماعی ملزم به شناسایی و حذف محتوا در یک بازه زمانی بسیار کم هستند (Gorwa et al, 2020: 5) و این الزام شرکت‌ها را به بهره‌گیری از هوش مصنوعی[1] جهت تعدیل محتوای افراطی خشونت‌آمیز موجود در شبکه‌های اجتماعی وامی‌دارد و به آن‌ها یاری می‌رساند تا این مطالب افراط گونه که هر روز در شبکه‌های اجتماعی بارگزاری می‌گردند، شناسایی و حذف شوند (Llansó et al., 2020: 4). در این زمینه طیف وسیعی از فرایندهای خودکار تعدیل محتوا که توسط هوش مصنوعی انجام می‌گیرد، می‌توانند مورد اشاره قرار گیرند؛ از جمله ابزار یادگیری ماشینی[2] و هشینگ[3] (Gorwa et al., 2020: 1).

۳ـ بررسی میزان اثرگذاری هوش مصنوعی در تعدیل محتوا و چالش‌های مرتبط با آن

در این مبحث مواردی همچون اندازه‌گیری دقت هوش مصنوعی در تعدیل محتوا و نقض آزادی بیان و دموکراسی، بررسی خواهد شد:

۳ـ۱ـ اندازه‌گیری دقت هوش مصنوعی در تعدیل محتوا

ابزارهای خودکار پیش‌بینی که از آن با عنوان «یادگیری ماشین (ML)[4]» یاد می‌شود، وظیفه‌ی شناسایی و تمایز بین انواع محتوا، بر اساس مجموعه داده‌های خاص اطلاعاتی بارگزاری شده در آن‌ها را بر عهده دارند (Llansó et al., 2020: 3).

---

[1] Artificial intelligence

[2] ML

[3] "هش" بخش اصلی و اساسی در رمزنگاری است و نقش بزرگی در رمزنگاری ارزهای دیجیتال ایفا می‌کند.

[4] Machine learning

در سال ۲۰۱۸، دولت بریتانیا اعلام کرد که یک ابزار کاملاً جدید «یادگیری ماشین» توسط علوم داده ابر هوش مصنوعی[1] و وزارت خانه[2] توسعه داده شده است. این ابزار توانست صداها و تصاویر ویدئوهای تبلیغاتی دولت اسلامی (داعش) را شناسایی و تجزیه‌وتحلیل کند (Home Office, 2018: 7). تحقیقاتی که در مورد آزمایش میزان عملکرد این ابزار جدید انجام گرفت، نشان دهنده‌ی این موضوع بود که ابزار جدید یادگیری ماشین می‌تواند به‌طور خودکار ۹۴ درصد از محتوای دولت اسلامی (داعش) را با دقت بیش از ۹۹.۹ درصد شناسایی کند (Home Office, 2018).

در همین راستا، بین ژوئن و دسامبر ۲۰۱۷، یوتیوب آمار مشابهی از موفقیت ۹۸ درصدی در حذف محتوای افراطی خشونت‌آمیز که به‌صورت خودکار شناسایی شده بودند منتشر کرد و اشاره کرد که تقریباً نیمی از آن‌ها ظرف ۲ ساعت پس از پست حذف شدند (Wojcicki, 2017: 1). هرچند که این آمار بیانگر عملکرد عالی این ابزارها در تعدیل محتواست با این حال آمار یاد شده ممکن است در مواردی گمراه کننده باشند (Gillespie, 2020: 5).

بخش اعظم محتوایی که توسط این ابزارها شناسایی و حذف می‌شود، حاوی محتوای تکراری است که قبلاً توسط عوامل انسانی ارزیابی و گزارش شده‌اند، این مسأله نشان دهنده‌ی این موضوع است که این ابزارها به‌صورت خودکار نمی‌توانند محتوای مضری را که در مورد آن گزارشی دریافت نکرده‌اند و حاوی داده‌های تکراری نیستند را شناسایی نمایند (Gillespie, 2020: 3).

همچنین استدلال شده است که ابزارهای تخصصی یادگیری ماشین که برای شناسایی و حذف موارد تکراری ایجاد شده‌اند، محتوای یکسانی را شناسایی نمی‌کنند. به عبارت دیگر ممکن است آن‌ها یک پست را حاوی محتوای خشونت‌آمیز تشخیص دهند و درصدد حذف آن برآیند؛ ولی ممکن است در اخبار و یا در سخنرانی، از بخش‌هایی کوتاه و تدوین شده از محتوای پستی که پیش از این توسط این ابزارها حذف شده بود استفاده شود و در این شرایط این ابزارها قادر به شناسایی این موارد نیستند؛ چراکه اطلاعات بارگزاری شده در این ابزارها که بر اساس آن قادرند عملکرد صحیحی داشته باشند محدود است (Llansó, 2020: 25). بنابراین، می‌توان استدلال کرد که استفاده از اقدامات حذف خودکار محتوا برای مقابله با افراط‌گرایی خشونت‌آمیز آنلاین، از نظر اثربخشی محدود است، زیرا فقدان آگاهی زمینه‌ای، می‌تواند کاربردهای نادرست و گسترده ایجاد کند

[1] Artificial super intelligence Data Science
[2] Home Office

(Engstrom & Feamster, 2017:11). درواقع هنگام اندازه‌گیری دقت هوش مصنوعی در تعدیل محتوا، مهم است که اتفاقات مثبت و منفی کاذب[1] و نحوه برخورد شرکت‌های رسانه‌های اجتماعی با آن‌ها را در نظر بگیریم (POST, 2020: 2). کاهش هم‌زمان احتمال وقوع مثبت کاذب و منفی کاذب در تعدیل محتوا غیرممکن است و در نتیجه، شرکت‌های رسانه‌های اجتماعی مجبور به اولویت‌بندی هستند (United Nations Office of Counter-Terrorism (UNCCT) & United Nations Interregional Crime and Justice (UNICRI), 2021).

باید گفت، اگرچه هر دو پیامد عواقب قابل توجهی دارند، اما افزایش احتمال مثبت کاذب، خطر نادیده گرفته شدن و عدم تشخیص محتوای افراطی خشونت‌آمیز زیان بار را به حداقل می‌رساند (UNCCT & UNICRI, 2021).

باید گفت، دستورالعمل‌ها و استانداردهای مختلفی وجود دارد که شرکت‌های رسانه‌های اجتماعی باید هنگام حذف محتوای افراطی از پلتفرم‌های خود آن‌ها را در نظر بگیرند. پایبندی به این دستورالعمل‌ها حتی ممکن است سبب شود که محتوای عاری از زیان حذف شود. همچنین ممکن است افراط‌گرایی در قانون هر کشوری تعریف خاص خود را داشته باشد و محتوایی که در یک کشور جزء موارد افراط‌گرایی به حساب می‌آید در کشور دیگر عاری از ویژگی‌های افراط‌گرایی باشد (van der Vegt et al., 2019). درواقع یکی از دلایل ایجاد این چالش و ایجاد مثبت و منفی‌های کاذب به دلیل عدم ارائه‌ی یک تعریف استاندارد در مورد «افراط‌گرایی» است که آن تعریف مورد قبول دول مختلف باشد (Duarte et al., 2018).

چالش دیگری نیز در این زمینه وجود دارد، ممکن است هوش مصنوعی به‌گونه‌ای برنامه‌ریزی شده باشد که پست‌ها یا کلیدواژه‌هایی همچون «حمایت»[2] یا «تجلیل»[3] را حذف کند. درواقع این دو کلید واژه به‌دلیل داده‌های محدود بارگزاری شده در هوش مصنوعی باعث تفسیر نادرست خواهند شد. ممکن است این کلمات تنها برای ابراز همدردی به کار برده شده باشند و نه حمایت از یک حرکت و یا گروه افراطی ولی به دلیل محدودیت داده تفسیر غلطی از این دو کلیدواژه ایجاد می‌شود و پست عاری از افراطی‌گری به همین دلیل حذف می‌گردد (Díaz & Hecht-Felella, 2021).

---

[1] مثبت کاذب به حذف خودکار محتوای واقعی از پلت فرم رسانه های اجتماعی و منفی کاذب به سیستمی اشاره دارد که محتوای مضر را به عنوان واقعی تشخیص می دهد(Ofcom, 2019).

[2] support
[3] glorification

آفکام[1] (۲۰۱۹)، بیان می‌دارد زمانی که در مقیاس جهانی عملی انجام شود، برای نمونه تعدیل محتوا، برای مؤثر واقع شدن آن بایستی به تفاوت‌های فرهنگی و قانونی توجه شود و این عمل بدون اجماع بین دولت‌ها، سازمان‌های مجری قانون و شرکت‌های رسانه‌های اجتماعی در مورد رسیدن به یک تعریف واحد از محتوای «افراطی»، میسر نیست (van der Vegt et al., 2019: 5).

نگرانی‌های دیگری نیز در این زمینه وجود دارد برای نمونه: بیشتر محتوای راست افراطی، به‌طور مستقیم با یک گروه مرتبط نیستند، که این پراکندگی باعث می‌شود که از همان ابتدا هنگام شناسایی محتوا برای حذف مشکلاتی ایجاد شود (Conway, 2020: 4). درواقع امروز شاهد هستیم که گفتمان راست افراطی در بسیاری کشورهای غربی؛ به‌ویژه در میان برخی رهبران سیاسی غربی جایگاه مخصوصی پیدا نموده است که این مسأله باعث ایجاد چالش به‌ویژه در زمینه‌ی تمایز قائل شدن بین افراط‌گرایی خشونت‌آمیز راست افراطی در محتوا۲، شده است (Ganesh & Bright, 2019).

۳ـ۲ـ نقض آزادی بیان و دموکراسی

استفاده از هوش مصنوعی در تعدیل محتوا باعث ایجاد نگرانی در مورد نقض اصول اساسی آزادی بیان و دموکراسی شده است (Llansó et al, 2020: 7).

به‌عنوان مثال، قانون NetzDG آلمان به‌خصوص به دلیل نقض آزادی بیان با انتقادات قابل توجهی مواجه شده است (Tworek & Leerssen, 2019: 2). زیرا این قانون، «تهدیدی برای گفتمان آزاد یا دموکراتیک است» و در مواردی باعث ترویج حذف محتوای قانونی می‌گردد که این مسأله حق آزادی بیان کاربران را خدشه‌دار می‌کند (Global Network Initiative (GNI), 2017). برخی همچنین ابراز نگرانی کرده‌اند که این قانون ممکن است باعث شود که شبکه‌های اجتماعی با توجه به این‌که مالکیت خصوصی دارند در هنگام ارزیابی کردن قانونی بودن محتوا، به‌صورت سلیقه‌ای عمل نمایند (Federal Cabinet, 2017: 3).

با این حال، برخی ابراز داشته‌اند که تاریخ آلمان نازی و استفاده از دموکراسی ستیزه‌جو[3] نشان داده است که آزادی بیان می‌تواند برای محافظت از هنجارهای یک دولت دموکراتیک محدود شود

---

[1] Ofcom

[۲] برای نمونه استفاده از میم (meme) که تصویر یا ویدئویی است که همراه با یک پیام متنی در پیام‌رسان‌ها و شبکه‌های اجتماعی دست‌به‌دست می‌شود و در اغلب مواقع لحن طنز دارند، توسط جوامع راست افراطی به دلیل مفهوم و پیام غیر مستقیمی که ارائه می‌دهند؛ رایج است که این مسأله سبب می‌شود سیستم‌های حذف خودکار نتوانند آن‌ها راشناسایی و حذف کنند(Lee, 2020: 3).

[3] Militant Democracy

(Tworek & Leerssen, 2019: 10) و در این حالت به‌سختی می‌توان از نقض آزادی بیان در موارد حذف خودکار محتوا پیشگیری کرد (Ganesh & Bright, 2019: 3). هرچند که ممکن است در برخی شرایط، تعدیل محتوا و حذف خودکار آن‌ها برای حفاظت از کاربران در جهت آسیب‌های آنلاین لازم باشد با این حال این حقیقت را نمی‌توان نادیده گرفت که قوانینی چون قانون (NetzDG) در کنار سایر چارچوب‌های نظارتی سخت‌گیرانه در مورد تعدیل محتوا، احتمالاً اصول اساسی دموکراسی را تضعیف می‌کند، زیرا با توجه به این قانون شرکت‌های رسانه‌های اجتماعی شخصاً باید در مورد محتوای افراطی تصمیم‌گیری کنند و این مسأله باعث می‌شود که تصمیمات گرفته شده در این شرکت‌ها با اصول یک دولت دموکراتیک متفاوت باشد (West, 2021: 5). برای حل این مشکل، پیشنهاد شده است که ساختارهای نظارتی جدید، مانند «دادگاه‌های الکترونیک»، جهت نظارت بر تصمیم‌گیری‌ها در مورد نحوه تعدیل انواع خاصی از محتوا توسعه یابند، این مسأله باعث کاهش قدرت شرکت‌های رسانه‌های اجتماعی در مقابل کاربرانشان خواهد شد و بدین ترتیب دموکراسی و آزادی بیان ترویج می‌یابد (Centre for Data Ethics and Innovation, 2020: 15).

در این میان باید گفت فقدان شفافیت و پاسخگویی سبب می‌شود که نقض آزادی بیان در طی فرایند تعدیل و حذف محتوا توسط سیستم‌های حذف خودکار شدت بیشتری بگیرد (West, 2021: 2)

بر اساس تحقیقات انجام شده توسط مرکز اخلاق و نوآوری داده (The Centre for Data Ethics and Innovation, 2021: 6)، رسانه‌های اجتماعی باید با انتشار اطلاعات، در مورد دقت فرآیند حذف خودکار محتوای خود و نحوه تصمیم‌گیری در مورد محتوای ممنوعه، شفاف‌تر باشند.

لازم به ذکر است که انواع خاصی از ابزارهای یادگیری ماشین، فرآیندهای مربوط به هنگام تصمیم‌گیری در مورد حذف محتوای افراطی خشونت‌آمیز بالقوه را ثبت نمی‌کنند (Coeckelbergh, 2019: 2). در نتیجه، در برخی موارد که در آن یک فرد به‌دنبال درخواست تجدیدنظر برای حذف مثبت کاذب است، دلیل حذف مشخص نخواهد شد؛ زیرا فرآیند تصمیم‌گیری خودکار غیرقابل کشف است (Henschke & Reed, 2021: 5). این مسأله به خودی خود بسیار چالش‌ساز خواهد بود؛ چراکه فرآیند تصمیم‌گیری در پس حذف خودکار محتوا باید هم توسط شرکت‌های رسانه‌های اجتماعی و هم توسط افرادی که درخواست تجدیدنظر کرده‌اند قابل بررسی باشد (Henschke & Reed, 2021: 6)؛ بنابراین استفاده از اقدامات حذف خودکار محتوا در رسانه‌های اجتماعی متداول و اصلی برای مقابله با افراط‌گرایی‌های آنلاین راست‌های افراطی بی‌اثر است، زیرا هم اصول اساسی آزادی بیان و هم دموکراسی را تضعیف می‌کند.

۴ـ استفاده از اقدامات پلتفرم‌زدایی

پلتفرم‌زدایی به مسدود کردن موقت یا حذف دائمی کاربران یا سازمان‌ها و گروه‌های خطرناک از پلتفرم‌های رسانه‌های اجتماعی به دلیل نقض دستورالعمل‌ها و استانداردهای شرکت‌ها اشاره دارد (Rogers, 2020: 3).

استفاده از پلتفرم‌زدایی به دلیل تشویق افراط‌گرایان، به‌ویژه راست‌گرایان افراطی، به مهاجرت از رسانه‌های اجتماعی رایج به رسانه‌های اجتماعی جایگزین مورد انتقاد قرار گرفته است، زیرا این شبکه‌های جایگزین فضای محافظت شده‌ای را برای ترویج ایدئولوژی راست افراطی فراهم می‌کنند (Donovan et al., 2018: 1) و در این شرایط این گروه‌ها نیازی به رعایت چارچوب‌های بیان شده در شبکه‌های اجتماعی متداول ندارند و شبکه‌های اجتماعی جایگزین نیز نظارت سخت‌گیرانه‌ای در حذف محتوا ندارند (Guhl et al., 2020: 3).

برخی نیز در جواب این گروه استدلال می‌کنند که پلتفرم‌زدایی از گروه‌های راست افراطی، هرچند ممکن است سبب شود که برخی از کاربران دنبال کننده‌ی این گروه‌ها به شبکه‌های جایگزین مهاجرت کنند، ولی در واقعیت تعداد کاربران مهاجرت کرده هرگز به گستردگی کاربران شبکه‌های اجتماعی اصلی نخواهد بود و تعداد کاربران کمتری عضو شبکه‌های جایگزین می‌شوند و این پلتفرم‌زدایی باعث محدود شدن تعداد کاربران بالقوه خواهد شد (Nouri et al, 2021: 16). در نمونه‌ای متشکل از ۲۵ گروه راست افراطی، کمی بیش از ۱۰ درصد از دنبال‌کنندگان از رسانه‌های اجتماعی اصلی به رسانه‌ها جایگزین مهاجرت کرده بودند، که نشان می‌دهد پلتفرم‌زدایی از گروه‌های راست افراطی، هم دسترسی این گروه‌ها را کاهش می‌دهد و هم این‌که مستقیماً کاربران را به سمت جایگزین سوق نمی‌دهد (Guhl et al, 2020: 5).

در مارس ۲۰۱۸، گروه راست افراطی «بریتین فرست»[1] به دلیل نقض «استانداردهای جامعه»[2] از فیسبوک حذف شد، اما به پلتفرم فناوری جایگزین «گب»[3] مهاجرت کردند (Nouri et al., 2019: 8). تحقیقاتی که به‌دنبال این مهاجرت شکل گرفت نشان داد با وجود مهاجرت این گروه به یک پلتفرم کوچک‌تر و با کاربران کمتر، به دلیل فقدان سانسور، تعامل با محتوای افراطی منتشر شده توسط اعضای «بریتین فرست» افزایش یافته است (Nouri et al, 2019: 9) که این مسأله نشان

[1] Britain First
[2] Community Standards
[3] Gab

دهنده‌ی آن است که بیان دیدگاه‌های افراطی قبلاً در بسترهای رسانه‌های اجتماعی اصلی محدود شده بودند. با این حال، تحقیقات دیگری نیز در مورد این شبکه نشان داد علی‌رغم این که فعالیت‌های گروه‌های راست افراطی در این شبکه ثابت مانده است و حجم مخاطبان کاهش یافته است با این حال گفتمان موجود در این شبکه به سمت گفتمان کمتر افراطی سوق پیدا کرده است که این مسأله ناکارآمدی ماهیت انتخابی قانون NetzDG و سایر تلاش‌های قانونی برای تنظیم محتوای افراطی در رسانه‌های اجتماعی اصلی را برجسته می‌کند (Rogers, 2020: 7). با این حال برخی تأکید می‌کنند که تحقیقات تجربی بیشتری برای تعیین میزان تأثیر اقدامات غیرپلتفرم‌زدایی مورد نیاز است (Guhl et al, 2020: 5).

با توجه به اثربخشی اقدامات پلتفرم‌زدایی در مقابله با افراط‌گرایی خشونت‌آمیز، مهم است که استفاده از هوش مصنوعی در اقدامات پلتفرم‌زدایی در نظر گرفته شود؛ زیرا ممکن است این روند را تقویت کند. ابزارهای یادگیری ماشین اغلب برای شناسایی و حذف محتوای خشونت‌آمیز و تبلیغات افراطی از پلتفرم‌های رسانه‌های اجتماعی اصلی استفاده می‌شوند، اما برای عملکرد بهتر لازم است که این سیستم‌های خودکار در فرآیندهای پلت‌فرم‌زدایی ترکیب شوند (Chaabene et al., 2021: 2). البته در حال حاضر نیز روش‌هایی وجود دارد که به‌صورت خودکار قابلیت شناسایی کاربرانی که رفتار افراطی دارند را فراهم نموده است با این حال این روش‌ها، تکنیک‌های محدودی برای شناسایی گروه‌های افراطی دارند. بدین منظور و در راستای رفع این محدودیت‌ها برخی از محققین یک مدل یادگیری ماشین را پیشنهاد نموده‌اند که ابزارهای لازم را برای شناسایی و پیش‌بینی رفتار خشونت‌آمیز افراطی را بر اساس نمودار توئیتر ارائه می‌دهد (Chaabene et al., 2021: 3).

نتیجه‌گیری

با ارزیابی انتقادی استفاده از هوش مصنوعی در تعدیل محتوا، می‌توان نتیجه گرفت که استفاده از حذف خودکار محتوا برای مقابله با افراط‌گرایی خشونت‌آمیز آنلاین از نظر اثربخشی محدود است؛ زیرا اطلاعات متنی را در نظر نمی‌گیرد و باعث ایجاد برنامه‌های نادرست، مانند موارد مثبت کاذب می‌شود. هنگام استفاده از ابزارهای حذف خودکار محتوا، رسانه‌های اجتماعی باید از اهمیت تفاوت‌های فرهنگی و حقوقی بین کشورهایی که در آن‌ها فعالیت می‌کنند آگاه باشند. با این حال، تأکید شد که پرداختن به موضوع مثبت کاذب و منفی کاذب در صورت عدم توافق جهانی در مورد تعریف واحد از کلمه «افراطی» ممکن نخواهد بود. همچنین مشخص شد که استفاده از هوش

مصنوعی در تعدیل محتوا از نظر اثربخشی محدود است؛ زیرا نمی‌تواند محتواهای افراطی خاص را به‌طور دقیق شناسایی کند، که این مسأله منجر به عادی‌سازی محتوای افراطی در میان رسانه‌های اجتماعی می‌شود. همچنین بیان شد که به‌کارگیری چارچوب‌های نظارتی سخت‌گیرانه در مورد تعدیل محتوا، اصول اساسی دموکراسی را تضعیف می‌کند، زیرا در این مسیر شرکت‌های رسانه‌های اجتماعی برای تصمیم‌گیری در مورد محتوای افراطی به حال خود رها شده‌اند. علاوه بر این، نگرانی‌هایی در مورد نقض آزادی بیان وجود دارد، زیرا با اصرار به حذف محتوای قانونی، حق کاربران برای آزادی بیان محدود می‌شود و این امر با فقدان شفافیت و پاسخگویی برای شرکت‌ها در طول فرآیند تصمیم‌گیری حذف خودکار محتوا تشدید می‌شود. با این حال، باید اذعان کرد که برخی از این انتقادات اجتناب‌ناپذیر هستند و بسیاری از آن‌ها با به‌کارگیری «اقدامات اضافی» یا اقدامات «مقابله با افراط‌گرایی خشونت‌آمیز» آنلاین قابل پیشگیری می‌باشند.

باید گفت، در این راستا توسعه ساختارهای نظارتی جدید، مانند «دادگاه‌های الکترونیک»، پیشنهاد شده است. توسعه‌ی این دادگاه‌ها سبب می‌شود، قدرت رسانه‌های اجتماعی در برابر کاربرانشان کاهش یابد که در نتیجه‌ی آن از دموکراسی حمایت می‌شود. همچنین در این پژوهش بیان شد که اقدامات پلتفرم‌زدایی ابزار مؤثرتری برای مقابله با افراط‌گرایی خشونت‌آمیز آنلاین، به‌ویژه در برابر گروه‌های راست افراطی است؛ زیرا نشان داده شده است که به‌طور مؤثری دامنه کلی روایت‌های راست افراطی را کاهش می‌دهد. با این حال برای تعیین میزان تأثیر اقدامات پلتفرم‌زدایی جهت تأمین امنیت کاربران رسانه‌های اجتماعی در مقابل گروه‌های افراطی تحقیقات تجربی بیشتری نیاز است. در این میان لازم است استفاده از هوش مصنوعی در اقدامات پلتفرم زدایی در دستور کار قرار گیرد، زیرا می‌تواند ظرفیت و توانایی حذف گروه‌های افراطی را افزایش دهد و ترویج ایدئولوژی‌های آن‌ها را محدود کند.

تعارض منافع

بنا بر اظهار نویسندگان، مقاله پیش‌رو فاقد هر گونه تعارض منافع بوده است.

Translated References to English

Allen, C. (2021). Extremism in the UK: New definitions threaten human and civil rights. The Conversation. https://theconversation.com/extremism-in-the-uk-new-definitions-threatenhuman-and-civil-rights-157086

Centre for Data Ethics and Innovation. (2020). online targeting: Final report and recommendations.

https://www.gov.uk/government/publications/cdei-review-of-online-targeting/onlinetargeting-final-report-and-recommendations#fn:189

Centre for Data Ethics and Innovation. (2021). the role of AI in addressing misinformation on social media platforms. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1008700/Misinformation_forum_write_up__August_2021-web_accessible.pdf

Chaabene, N.E.H.B., Bouzeghoub, A., Guetari, R., Ghezala, H.H.B. (2021). Applying machine learning models for detecting and predicting militant terrorists behavior in Twitter. In 2021 IEEE international conference on systems, man, and cybernetics (SMC) (pp. 309–314). IEEE. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9659253

Coeckelbergh, M. (2019). Artificial intelligence: Some ethical issues and regulatory challenges. Technology and Regulation, 2019, 31–34. https://techreg.org/article/view/10999/11973

Conway, M. (2020). Routing the extreme right: Challenges for social media platforms. The RUSI Journal, 165(1), 108–113. https://www.tandfonline.com/doi/pdf/10.1080/03071847.2020.1727157?needAccess=true

Díaz, Á., Hecht-Felella, L. (2021). Double standards in social media content moderation. https://www.brennancenter.org/sites/default/files/2021-08/Double_Standards_Content_Moderation.pdf

Donovan, J., Lewis, B., Friedberg, B. (2018). Parallel ports. Sociotechnical change from the Alt-Right to Alt-Tech. In M. Fielitz & N. Thurston (Eds.), Post-digital cultures of the far right (pp. 49–66). Transcript Verlag. https://www.degruyter.com/document/doi/10.14361/9783839446706-004/html

Duarte, N., Llanso, E., Loup, A.C. (2018). Mixed messages? The limits of automated social media content analysis. https://cdt.org/wp-content/uploads/2017/12/FAT-conferencedraft-2018.pdf

Engstrom, E., Feamster, N. (2017). The limits of filtering: A look at the functionality and shortcomings of content detection tools. https://static1.squarespace.com/static/571681753c44d835a440c8b5/t/58d058712994ca536bbfa47a/1490049138881/FilteringPaperWebsite.pdf

Federal Bureau of Investigation. (n.d.). What is violent extremism? https://www.fbi.gov/cve508/teen-website/what-is-violent-extremism

Federal Cabinet. (2017). Declaration on freedom of expression. https://deklaration-fuermeinungsfreiheit.de/en/.

Ganesh, B., Bright, J. (2019). Extreme digital speech: Contexts, responses, and solutions. https://www.voxpol.eu/download/vox-pol_publication/DCUJ770-VOX-Extreme-Digital-Speech.pdf

Gillespie, T. (2020). Content moderation, AI, and the question of scale. Big Data & Society, 7(2),1–5. https://journals.sagepub.com/doi/pdf/10.1177/2053951720943234

Global Network Initiative. (2017). Proposed German legislation threatens free expression around the world. https://globalnetworkinitiative.org/proposed-german-legislation-threatensfree-expression-around-the-world/

Gorwa, R., Binns, R., Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. Big Data & Society, 7(1), 1–15. https://journals.sagepub.com/doi/full/10.1177/2053951719897945

Guhl, J., Ebner, J., Rau, J. (2020). The online ecosystem of the German farright. https://www.isdglobal.org/wp-content/uploads/2020/02/ISD-The-Online-Ecosystem-ofthe-German-Far-Right-English-Draft-11.pdf

Henschke, A., Reed, A. (2021). Toward an ethical framework for countering extremist propaganda online. Studies in Conflict & Terrorism, 1–18. https://www.tandfonline.com/doi/pdf/10.1080/1057610X.2020.1866744?needAccess=true

HM Government. (2019). Online harms white paper. https://dera.ioe.ac.uk/33220/1/Online_Harms_White_Paper.pdf
Home Office. (2015). Revised prevent duty guidance: For England and Wales. https://www.gov.uk/government/publications/prevent-duty-guidance/revised-prevent-duty-guidance-for-englandand-wales
Home Office. (2018). New technology revealed to help fight terrorist content online. https://www.gov.uk/government/news/new-technology-revealed-to-help-fight-terrorist-content-online
Jaafarifar. E. (2023). The rise of the far-right and the challenges of European politics, Quarterly Journal of International Political Studies, Volume 2, Issue 2
Kundnani, A., Hayes, B. (2018). The globalization of countering violent extremism policies: Undermining human rights, instrumentalising civil society. https://www.tni.org/files/publication-downloads/the_globalisation_of_countering_violent_extremism_policies.pdf
LaFree, G., Freilich, J.D. (2019). Government policies for counteracting violent extremism. Annual Review of Criminology, 2, 13.1–13.22.
Lee, B. (2020). Neo-Nazis have stolen our memes: Making sense of extreme memes. In M. Littler&B. Lee (Eds.), Digital extremisms: Readings in violence, radicalization and extremism in the online space (pp. 91–108). Palgrave Studies in Cybercrime and Cybersecurity.
Llansó, E.J. (2020). No amount of "AI" in content moderation will solve filtering's priorrestraint problem. Big Data & Society, 7(1), 1–6.
Llansó, E., Van Hoboken, J., Leerssen, P., Harambam, J. (2020). Artificial intelligence, content moderation, and freedom of expression. https://www.ivir.nl/publicaties/download/AI-Llanso-Van-Hoboken-Feb-2020.pdf
Lowe, D. (2017). Prevent strategies: The problems associated in defining extremism: The case of the United Kingdom. Studies in Conflict & Terrorism, 40(11), 917–933. https://www.tandfonline.com/doi/pdf/10.1080/1057610X.2016.1253941?needAccess=true
Miller, J. (2017). Germany votes for 50m euro social media fines. British Broadcasting Corporation News. https://www.bbc.co.uk/news/technology-40444354
Neumann, P.R. (2013). The trouble with radicalization. International Affairs, 89(4), 873–893. https://www.jstor.org/stable/pdf/23479398.pdf?ab_segments=0%2Fbasic_search_solr_cloud%2Fcontrol&refreqid=fastly-default%3Aa01b19ab612357c46a149138dd388925
Nouri, L., Lorenzo-Dus, N., Watkin, A.L. (2019). Following the whack-a-mole: Britain First's visual strategy from Facebook to Gab. https://static.rusi.org/20190704_grntt_paper_4.pdf
Nouri, L., Lorenzo-Dus, N., Watkin, A.L. (2021). Impacts of radical right groups' movements across social media platforms–a case study of changes to Britain First's visual strategy in its removal from Facebook to Gab. Studies in Conflict & Terrorism, 1–27. https://doi.org/10.1080/1057610X.2020.1866737
Ofcom. (2019). Use of AI in online content moderation. https://www.ofcom.org.uk/_data/assets/pdf_file/0028/157249/cambridge-consultants-ai-content-moderation.pdf
Parliamentary Office of Science and Technology. (2020). Online extremism. https://researchbriefings.files.parliament.uk/documents/POST-PN-0622/POST-PN-0622.pdf
Piazza, J.A., Guler, A. (2019). The online caliphate: Internet usage and ISIS support in the Arab world. https://www.tandfonline.com/doi/pdf/10.1080/09546553.2019.1606801?needAccess=true
Rogers, R. (2020). Deplatforming: Following extreme Internet celebrities to Telegram and alternative social media. European Journal of Communication, 35(3), 213–229. https://journals.sagepub.com/doi/full/10.1177/0267323120922066

Schmid, A.P. (2014). Violent and non-violent extremism: Two sides of the same coin. https://opev.org/wp-content/uploads/2019/10/Violent-and-Non-Violent-Extremism-Alex-P.-Schmid.pdf

Selim, G. (2016). Approaches for countering violent extremism at home and abroad. The Annals of the American Academy of Political and Social Science, 668(1), 94–101. https://journals.sagepub.com/doi/pdf/10.1177/0002716216672866

Stephens, W., Sieckelinck, S., Boutellier, H. (2018). Preventing violent extremism: A review of the literature. Studies in Conflict & Terrorism, 1–16. https://www.tandfonline.com/doi/full/10.1080/1057610X.2018.1543144

Tworek, H., Leerssen, P. (2019). An analysis of Germany's NetzDG law. https://cdn.annenbergpublicpolicycenter.org/wp content/uploads/2020/06/NetzDG_TWG_Tworek_April_2019.pdf

United Nations Office of Counter-Terrorism, & United Nations Interregional Crime and Justice. (2021). Countering terrorism online with artificial intelligence. https://www.un.org/counterterrorism/sites/www.un.org.counterterrorism/files/countering-terrorism-online-with-aiuncct-unicri-report-web.pdf

United Nations Office on Drugs and Crime. (2012). The use of the Internet for terrorist purposes. https://www.unodc.org/documents/frontpage/Use_of_Internet_for_Terrorist_Purposes.pdf

van der Vegt, I., Gill, P., Macdonald, S., Kleinberg, B. (2019). Shedding light on terrorist and extremist content removal.https://rusi.org/explore-our-research/publications/special-resources/shedding-light-on-terrorist-and-extremist-content-removal

Vincent, C., Hunter-Henin, M. (2018, February 10). The trouble with teaching 'British values' in school. Independent. https://www.independent.co.uk/news/education/british-values-educationwhat-schools-teach-extremism-culture-how-to-teachers-lessons-a8200351.html

Wakefield, J. (2021). Government lays out plans to protect users online. British Broadcasting Corporation News. https://www.bbc.co.uk/news/technology-57071977

West, L.J. (2021). Counter-terrorism, social media and the regulation of extremist content. In S. Miller, A. Henschke., Feltes, J. (Eds.), Counter-terrorism: The ethical issues (pp. 116–128). Edward Elgar Publishing. https://www.elgaronline.com/view/edcoll/9781800373068/9781800373068.00016.xml

Wojcicki, S. (2017). Expanding our work against abuse of our platform. Youtube Official Blog. https://blog.youtube/news-and-events/expanding-our-work-against-abuse-of-our/