Quarterly Journal of Virtual Space and Social Media
Vol. 1, Series. 1, No. 1, Spring 2024/ 47 - 84
https://doi.org/ 10.22083/cssms.2024.206850

Content Moderation Mechanisms in Video Platforms: A Case Study of YouTube

***Hossein Hassani**, Assistant Professor of Cyberspace Studies, RICAC. hassani@ricac.ac.ir*
Email: Hassani@ricac.ac.ir

Abstract

Introduction: Today, a significant portion of platformized content on the internet (Flew, 2021) is produced in video format. The monitoring and regulation of content across various user-generated content-sharing platforms has introduced a new landscape for media content governance. One of the darker aspects of this shift is the facilitation of producing and disseminating various forms of inappropriate content, ranging from misleading information to harmful and disruptive videos. YouTube has become one of the most popular content-sharing platforms globally, attracting a diverse, international user base. The rapid advancement of new technologies, such as artificial intelligence, in the distribution of disruptive video content has made filtering and blocking these types of content more challenging than ever before. Accordingly, the primary focus of this paper is to examine the mechanisms employed by YouTube, a leading global video platform, in moderating and filtering content.

Methods: This research adopts a qualitative approach, specifically utilizing a case study design. A case study is a research method that involves an in-depth and detailed examination of a specific subject, such as an individual, group, event, organization, or phenomenon (Crave et al., 2011). Case studies employ various data collection methods, such as interviews, observations, document analysis, and archival records, to gather comprehensive information on the topic of interest. The goal of a case study is to provide a contextualized, nuanced understanding of the subject, examine relationships among variables, identify patterns, and generate insights that may contribute to theoretical development or practical solutions.

This study, conducted during 2023-2024, aims to offer a detailed and holistic analysis of content moderation processes on YouTube, with a specific focus on the platform's handling of video content and the application of algorithmic moderation facilitated by artificial intelligence. The study includes an analysis of YouTube's documents and policies. Data collection relied primarily on YouTube's own documents and materials, such as community guidelines, various policies, published reports, and external studies and reports analyzing YouTube's content moderation. Data analysis was performed through qualitative content analysis, with main categories extracted after reviewing various documents and further substantiated through extracted evidence.

Conclusion: Content moderation has become an essential component of digital platforms, protecting users from harmful content while ensuring an inclusive and safe online environment. The challenge of managing vast volumes of user-gener-

ated content has led to the development of advanced moderation mechanisms that incorporate both artificial intelligence and human moderation efforts. AI-based moderation systems have demonstrated exceptional efficiency in identifying and flagging problematic content at scale. However, they are not without limitations, as they often struggle with contextual understanding and language nuances, which can sometimes lead to incorrect flagging and removal of lawful content.

Content moderation has both theoretical and practical dimensions. First, platforms must develop a comprehensive set of documents-such as terms of service, community guidelines, privacy and safety policies, and misinformation policies-based on overarching documents, legal requirements, and platform-specific approaches. Additionally, certain operational procedures should be established within the platform to clearly and unambiguously distinguish harmful from non-harmful content. Human moderators play a critical role in addressing these limitations by bringing judgment and discernment to the moderation process. They are better equipped to understand the complexities of language, culture, and context, ensuring a more accurate content assessment.

Content moderation is a complex mechanism that spans guidelines and practical content filtering procedures. As new forms of disruptive content are increasingly created and shared, these guidelines must be regularly updated and remain as clear and precise as possible, allowing both human and machine moderators to operate with ease and clarity.

To achieve this goal, Iranian platforms should invest in advanced AI technology while fostering a supportive and empathetic environment for human moderators. Additionally, Iranian platforms should enhance transparency, accountability, and collaboration with users, industry stakeholders, and regulatory authorities to ensure that content moderation practices align with social values and legal requirements. This approach will ultimately create a safe and inclusive online environment for Iranian users while mitigating risks associated with user-generated content, particularly regarding video content.

Keywords

Platform governance, Content moderation, Video platform, YouTube.

فصلنامه فضای مجازی و رسانه‌های اجتماعی
سال ۱ / شماره ۱/ شماره مسلسل ۱ / بهار ۱۴۰۳/ ۸۴ - ۴۷
Quarterly Journal of Virtual Space and Social Media
Vol. 1, Series. 1, No. 1, Spring 2024/ 47 - 84
https://doi.org/ 10.22083/cssms.2024.206850

سازوکارهای پالایش محتوا در پلتفرم‌های ویدئویی؛ مطالعه موردی یوتیوب

حسین حسنی[۱]

چکیده

هدف این مقاله، مطالعه نحوهٔ پالایش محتواهای زیان‌آور در پلتفرم کاربرمحور و ویدئویی یوتیوب است. امروزه موضوع حکمرانی و تنظیم‌گری محتوایی پلتفرم‌ها به یک موضوع چالش‌برانگیز تبدیل شده است. این مسئله به‌ویژه در پلتفرم‌هایی که مبتنی‌بر ویدئو فعالیت می‌کنند، چالش‌های بیشتری ایجاد کرده است. پخش زنده محتوای ویدئویی و نیز انواع محتواهای ویدئویی کذب همانند جعل عمیق که در پلتفرم‌های ویدئو، مبنا به اشتراک گذاشته می‌شوند، مقابله و جلوگیری از انتشار محتواهای ویدئویی زیان‌آور، غیرقانونی و غیراخلاقی را دشوارتر کرده است. این تحقیق با استفاده از روش مطالعهٔ موردی و تحلیل محتوای کیفی، خط‌مشی‌های یوتیوب، دستورالعمل‌های انجمن و کردارهای عملی تعدیل محتوا در یوتیوب را بررسی می‌کند تا مشخص شود این پلتفرم چگونه محتوای تولیدشده توسط کاربر را پالایش می‌کند. نتایج این مقاله در کل، بیانگر پیچیدگی نظارت برمحتوای ویدئویی، کاربرد روزافزون هوش مصنوعی برای تعدیل محتوای ویدئویی و مخاطرات ناشی از آن است. برخلاف تصور عموم و برخی سیاست‌گذاران در ایران، پلتفرم‌های جهانی سازوکارهای پیچیده و دائماً به‌روزشونده‌ای را برای حفاظت از اجتماع کاربران خود ایجاد کرده‌اند تا فضای امنی را برای تعامل کاربران و تداوم سودآوری و حفظ وجهٔ خود به وجود آورند. همچنین با توجه به اینکه پلتفرم‌های بومی در ایران در حال توسعه هستند، سازوکار تعدیل محتوا در پلتفرم یوتیوب جهانی می‌تواند الگویی برای پلتفرم‌های ایرانی باشد.

واژگان کلیدی

پلتفرم ویدئویی، تعدیل محتوا، حکمرانی پلتفرم، یوتیوب.

تاریخ دریافت: ۱۴۰۳/۰۲/۲۶ تاریخ پذیرش: ۱۴۰۳/۰۳/۲۰

۱. استادیار مطالعات فضای مجازی، پژوهشگاه فرهنگ، هنر و ارتباطات، تهران، ایران. (نویسنده مسئول).
hassani@ricac.ac.ir

مقدمه و طرح مسئله

امروزه بخش چشم‌گیری از محتوایی که در اینترنت پلتفرمی شده (Flew, 2021) تولید می‌شود، محتواهای ویدئویی هستند. انواع گوناگونی از محتواهای ویدئویی وجود دارد و محبوب‌ترین آن‌ها در پاییز ۲۰۲۳ به‌ترتیب نماهنگ (نزدیک به ۴۸ درصد)، کمدی، میم و ویدئوی منتشرشده به شکل ویروسی (۳۳/۵ درصد)، پخش زنده ویدئویی (۲/۷ درصد)، ویدئوی آموزشی یا نحوهٔ انجام کار (۲۵/۷ درصد)، ویدئوی درسی (۲۵/۶ درصد)، ویدئوی نقد محصولات (۲۵/۲ درصد)، کلیپ‌های ورزشی یا ویدئوی بخش‌های برگزیده (۲۴/۹ درصد)، ویدئوهای متنفذان و وبلاگ‌های ویدئویی (۲۳/۸ درصد) و ویدئوهای مرتبط با گیم (۲۳/۲ درصد). در رده‌بندی محبوب‌ترین شبکه‌های اجتماعی پس از فیس‌بوک، پلتفرم‌های یوتیوب (رده دوم با نزدیک به دو میلیارد و پانصدمیلیون کاربر)، اینستاگرام (رده چهارم با دومیلیارد کاربر) و تیک‌تاک (رده پنجم با نزدیک به یک میلیارد و ششصدمیلیون کاربر)، سه پلتفرم برتر جهانی محسوب می‌شوند (Dixon, 2024).

یک ویژگی اساسی پلتفرم‌های اخیر این است که آن‌ها مبتنی‌بر محتوای کاربرساخته هستند. به عبارت دیگر، آن‌ها زیرساخت رایانش را ایجاد کرده‌اند و کل فرایند تولید، اشتراک‌گذاری و مصرف محتوا و کردارها و تعاملات معنادار دیگر همچون اظهارنظر دربارهٔ محتواها، پسند کردن، انتشار ویروسی و داغ شدن آن‌ها توسط کاربران انجام می‌شود. درواقع با یک وضعیت بیش‌اتصال مواجه شده‌ایم که میلیاردها ویدئوی منتشرشده در پلتفرم‌های گوناگون ما را با نوعی وفور در بی‌انتهایی محتوای دیجیتال ویدئویی مواجه کرده است (Brubaker, 2023). با وجود اینکه یکی از نویدهای این وفور، دمکراتیزه شدن[1] و فرهنگ مشارکتی‌تر تولید و اشتراک‌گذاری محتوا توسط عموم کاربر بوده است، یکی از سویه‌های تاریک این تحول، تسهیل تولید و انتشار انواع محتواهای نابهنجار، از اطلاعات گمراه‌کننده تا ویدئوهای ناهنجار و زیان‌آور بوده است.

نظارت و کنترل انواع محتواهای پلتفرم‌های ویدئویی برخط، همانند خدمات اشتراک‌گذاری ویدئویی برخط، آی.پی.تی وی.ها، خدمات ویدئوهای درخواستی، نمایش زنده (لایو) جمعی در پلتفرم‌ها همانند لایو اینستاگرام و به‌طور خاص پلتفرم‌های اشتراک‌گذاری محتوای کاربرساخته، چشم‌انداز جدیدی را در عرصهٔ تنظیم

1. Democratization

محتوای رسانه‌ای ایجاد کرده است. همچنین باید به این نکتهٔ مهم نیز توجه کرد که تولیدکنندگان محتوا دیگر صرفاً عاملان انسانی نیستند و ربات‌ها و الگوریتم‌ها نیز در تولید و اشتراک‌گذاری محتوا نقش دارند؛ بنابراین اینکه مسئولیت تولید محتوا با چه کسی یا چیزی است و از طرف دیگر، چگونگی پالایش محتواهای نامناسب بیش از پیش پیچیده می‌شود. پالایش محتوا عموماً حکمرانی در پلتفرم (Gorwa, 2019) یا تعدیل محتوا نامیده می‌شود که منظور از آن «کردار سازمان‌یافته رصد محتواهای کاربرساخته منتشرشده در سایت‌های اینترنتی، رسانه‌های اجتماعی و مجراهای دیگر، به‌منظور تعیین تناسب آن محتوا با یک سایت خاص، موقعیت یا حوزهٔ قضایی است» (Roberts, 2017: 1).

امروزه یوتیوب یکی از محبوب‌ترین پلتفرم‌های اشتراک‌گذاری محتوا است که کاربرانی جهانی دارد. مدل کسب‌وکار و فعالیت آن مبتنی‌بر محتواهایی است که کاربران، خواه کاربران معمولی یا خالقان محتوا، به اشتراک می‌گذارند. این پلتفرم عمدتاً و در اصل یک پلتفرم ویدئو ـ مبنا است. پالایش محتواهای ویدئویی نسبت به قالب‌های دیگر محتوایی ازجمله متن دشوارتر و پیچیده‌تر است. توسعه فزاینده فناوری‌های جدید همانند هوش مصنوعی برای انتشار محتواهای نابهنجار ویدئویی، پالایش و مسدودسازی این نوع محتواها را دشوارتر از پیش می‌کند؛ بنابراین، مسئله اساسی این نوشتار این است که پلتفرم بزرگ جهانی ویدئو ـ محور یوتیوب چه سازوکاری را برای تعدیل و پالایش محتوا به‌کار گرفته است.

پیشینه پژوهش

مطالعات دربارهٔ حکمرانی و تنظیم‌گری پلتفرم‌ها در ایران یک حوزهٔ پژوهشی نوظهور است. هرچند در سطح جهان نیز مطالعاتی که دربارهٔ تنظیم‌گری محتوای پلتفرم‌ها، چه پلتفرم‌های ناشرمحور همانند سامانه‌های ویدئوی درخواستی یا کاربرمحور همانند یوتیوب و تیک‌تاک انجام شده‌اند، قدمتی کمتر از ده سال دارند؛ این موضوع به نوظهور بودن مطالعات پلتفرم برمی‌گردد. برخی مطالعات طی سالیان گذشته در ایران در این زمینه انجام شده‌اند که به آن‌ها اشاره می‌شود. ذکر این نکته الزامی است که مطرح شدن تنظیم‌گری محتوای سامانه‌های ویدئوی درخواستی و شکل‌گیری ساترا از عوامل اصلی توجه به محتواهای پلتفرم‌های ناشرمحور است، اما در مورد پلتفرم‌های کاربرمحور در ایران مطالعات چندانی انجام نشده است. سرحدی و طاهری (۱۳۹۹)

اعمال نظارت مطلوب بر انتشار صوت و تصویر در فضای مجازی از منظر حقوقی را مطالعه کرده‌اند. به اعتقاد آن‌ها نظارت بر فعالیت‌های مجازی باید به‌صورت پسینی اعمال شده و دولت صرفاً در صورت وقوع جرم یا تخلف، به موضوع ورود کرده و اعمال حاکمیت نماید. همچنین استفاده از ظرفیت هیئت‌منصفه و سازمان‌های مردم‌نهاد جهت نظارت مطلوب بر این عرصه می‌توانند مورد توجه قانون‌گذاران این عرصه قرار گیرد. خرم‌دل و همکارانش (۱۴۰۱) چالش‌های مربوط به مرجع تنظیم‌گر و ناظر بر تولیدات صوت و تصویر فراگیر در فضای مجازی را مطالعه کرده‌اند. آن‌ها پیشنهاد کرده‌اند که قانونی جامع در خصوص حقوق رسانه‌ها توسط مجلس وضع و مرجعی صالح معرفی شود که بتواند ضمن رعایت حق آزادی بیان در همه رسانه‌ها و رعایت اصل رقابتی بودن، به لحاظ فنی و حقوقی به امر تنظیم‌گری بپردازد. طحان نظیف و علی‌پور (۱۴۰۱) در مقاله‌ای جایگاه و آثار حقوقی خودتنظیم‌گری پلتفرم‌های دیجیتال را مطالعه کرده‌اند. به گفتهٔ آن دو، انواع روش‌های خودتنظیم‌گری را می‌شود توافقی میان کاربران با یکدیگر یا با صاحبان پلتفرم دانست که البته در شرایط گوناگون، بسته به نوع اعمال و زیرساخت‌های فنی آن، جایگاه متفاوتی در نظام حقوقی دارد. اخوان و همکارانش (۱۴۰۲) الگوی حکمرانی صوت‌وتصویر فراگیر در روسیه، ترکیه و کره جنوبی را به شکل تطبیقی مطالعه کرده‌اند. یافته‌های آن‌ها نشان می‌دهد که هر سه کشور الگوی حکمرانی رسانه هم‌گرا را پذیرفته‌اند و در صوت‌وتصویر فراگیر، بیشتر به حکمرانی از بالا به پایین گرایش دارند و رسانه‌ها در محدودهٔ توقعات دولت و بازار رفتار می‌کنند. قاسم‌زاده عراقی و همکاران (۱۴۰۲) تلاش کرده‌اند خدمات رسانه‌ای صوت‌وتصویر فراگیر در ایران (با تأکید بر تجربه اتحادیه اروپا) را از نظر مفهومی تبیین کنند. این مطالعه تطبیقی دربارهٔ تفسیر کشورهای عضو اتحادیه اروپا، نشان‌دهندهٔ تفاسیر واگرا درخصوص شاخص‌های سازنده خدمات رسانه‌ای صوت‌وتصویر است.

چارچوب مفهومی

در بخش بعدی مقاله که به چارچوب مفهومی مقاله اختصاص دارد، ابتدا مفهوم حکمرانی پلتفرم‌ها را مرور خواهیم کرد. سپس انواع تنظیم‌گری محتوا را مورد توجه قرار خواهیم داد و درنهایت به تعدیل محتوا خواهیم پرداخت و سپس مفهوم محتوای کاربرساخته ویدئویی را مرور خواهیم کرد.

حکمرانی و تنظیم‌گری پلتفرم‌ها و رسانه‌های اجتماعی

به‌طورکلی حکمرانی پلتفرم‌ها در تداوم حکمرانی اینترنت قرار دارد. پس از فراگیرتر شدن مفهوم پلتفرم به‌جای رسانه اجتماعی از حدود میانهٔ دهه ۲۰۱۰ و پلتفرمی شدن اینترنت (Flew, 2022) که با تمرکز بخش عمدهٔ ترافیک اینترنت در این زیرساخت‌های شبکه‌ای رایانش روی داد به حکمرانی پلتفرم‌ها توجه شد و اندیشمندان گوناگون اقدام به مفهوم‌پردازی و نظریه‌پردازی دربارهٔ آن کردند (برای نمونه، DeNardis & Hackl, 2015؛ Gillespie, 2018؛ Gorwa, 2019).

به‌طورکلی رویکردهای مرتبط با حکمرانی پلتفرم‌ها را به سه دسته می‌توان تقسیم‌بندی کرد: حکمرانی بر پلتفرم، حکمرانی به‌واسطهٔ پلتفرم‌ها و حکمرانی در پلتفرم. حکمرانی بر پلتفرم شامل اقدامات نهادهای قانون‌گذار و اجرایی برای تنظیم‌گری محتوای پلتفرم‌ها و روابط پلتفرم‌ها با یکدیگر است. حکمرانی به‌واسطهٔ پلتفرم‌ها شامل کاربرد زیرساخت‌های پلتفرمی برای ارائه خدمات بهتر به شهروندان در نظر گرفت که دولت پلتفرمی نیز درهمین‌راستا قرار دارد، اما حکمرانی در پلتفرم شامل تنظیم‌گری محتوا توسط خود پلتفرم‌ها است که تعدیل محتوا یا پالایش محتوا نیز نامیده می‌شود که این نوع از خودتنظیم به دلیل ماهیت پلتفرم‌ها که مبتنی‌بر محتوای کاربر ساخته است، رواج پیدا کرده است (کلانتری و حسنی، ۱۳۹۹). سرعت و حجم محتوای کاربرساخته و گونه‌گونی کاربران و مسائل نوپدید سبب شده تا در جامعه پلتفرمی کنونی بخش عمدهٔ نظارت محتوای به خود پلتفرم‌ها واگذار شود.

از منظری دیگر، از مفهوم تنظیم‌گری محتوا نیز سخن گفته شده است که شکل‌های مختلفی دارد. هرچند در اصل حکمرانی فراتر از تنظیم‌گری است (قاسم‌زاده عراقی و همکاران، ۱۴۰۲). تنظیم‌گری توسط حکومت یا دولت که شامل قوانین، سیاست‌ها و تنظیم‌گری‌هایی است که توسط دولت‌های ملی انجام می‌شود. خودتنظیم‌گری شامل دستورالعمل‌ها و اصول رفتارهای است که توسط خود پلتفرم‌ها یا صنایع رسانه‌ای ایجاد و اعمال می‌شود. تنظیم‌گری مشترک شامل همکاری بین تنظیم‌گران دولتی و ذی‌نفعان صنعت برای توسعه و اعمال مقررات مرتبط با محتوا است. تنظیم‌گری فناورانه که استفاده از ابزارها و سامانه‌ها برای کنترل دسترسی به محتوا یا اعمال مقررات محتوایی است و بالاخره تنظیم‌گری مصرف‌کننده یا کاربر است که طی آن یکایک کاربران مواجهٔ خود با محتوا از روش‌های گوناگون همانند استفاده از نرم‌افزارهای مسدودسازی محتوا، تطبیق تنظیمات حریم خصوصی براساس ترجیحات خود یا برگزیدن انواع خاص محتوا را کنترل می‌کنند.

علامت‌دهی[1] یا گزارش کاربران یک از دیگر انواع مشارکت کاربران در تنظیم‌گری محتوا است که با رویکرد فوق اندکی تفاوت دارد. معمولاً این نوع تنظیم‌گری به‌مثابهٔ بخشی از انواع تنظیم‌گری محتوا تلقی می‌شود و یک رویکرد مستقل به تنظیم‌گری به شمار نمی‌رود. تعدیل محتوا که در ادامه مورد بحث قرار می‌گیرد، شکلی از خودتنظیم‌گری است که در ادامه در مورد آن بحث می‌شود.

تعدیل محتوا

حکمرانی توسط پلتفرم‌ها، تنظیم خصوصی و تعدیل محتوا در معنای کلی، به معنای اعمال مقررات ناظر بر رفتار و گفتار کاربران در پلتفرم‌های رسانه‌های اجتماعی که مدل فعالیت آن‌ها مبتنی‌بر محتوای کاربر ساخته است. این نوع پلتفرم‌ها در اصل زیرساخت رایانشی برای میانجی‌گری ارتباط میان کاربران و اشتراک‌گذاری محتوا را فراهم می‌کنند، اما عمدتاً موظف‌اند پالایش محتوا را نیز انجام دهند که در قالب اصطلاحات مختلف، ازجمله تعدیل محتوا (Gillespie, 2018)، تنظیم‌گری محتوا (Tan, 2018)، سانسور محتوا (Gazethoni, 2023) و تنظیم‌گری پلتفرم (Gorwa, 2024) نامیده شده است. در این نوشتار از اصطلاح تعدیل محتوا استفاده می‌کنیم که شامل فرایند نظارت، بررسی و مدیریت محتوای کاربرساخته در پلتفرم‌های برخط بر اطمینان از تبعیت محتوا از دستورالعمل‌های اجتماع پلتفرم، شرایط خدمات و قوانین و مقررات اجرایی است.

تعدیل محتوا امروزه یکی از مباحث اصلی سیاست‌گذاری پلتفرم‌ها محسوب می‌شود و توجه دانشگاهیان را به مسائل وضعیت کنونی حکمرانی اینترنت جلب کرده است (Rickstein and Tronen, 2020: 3). این مسئله که چه محتواهایی باید در رسانه‌های اجتماعی نمایش داده شوند و چه محتواهایی خیر، موضوعی چالش‌برانگیز است که رسانه‌های اجتماعی از پلتفرم‌های جهانی گرفته تا پلتفرم‌های ملّی و محلی با آن دست‌به‌گریبان هستند. درواقع، تصمیم‌هایی که باید توسط فرد یا افراد ناظر محتوا گرفته شود، تابع عوامل تأثیرگذار بسیاری است که طیفی از ارزش‌های محلی، زیبایی‌شناختی، اخلاقی ـ دینی، سیاسی، فرهنگی ـ اجتماعی و هنری در آن دخیل هستند. به‌ویژه این موضوع در رسانه‌های اجتماعی که اساساً تولیدمحتوا در آن‌ها توسط کاربران انجام می‌شود و تولید اخبار جعلی و اطلاعات گمراه‌کننده، محتواهای

1. flagging

نفرت‌پراکنانـه و هرزه‌نگارانـه یکـی از دغدغه‌هـای اساسـی سیاسـت‌گذران، مقامـات و خانواده‌هـا اسـت، اهمیـت مرکـزی دارد.

امـروزه شـرکت‌ها، طیفـی از رویکردهـای گوناگـون بـرای تعدیـل محتـوا و نیـز ابزارهـای مختلـف بـرای اعمـال سیاسـت‌های محتـوا و حـذف محتواهـا و حسـاب‌های کاربـری قابـل اعتـراض اسـتفاده می‌کننـد. به‌گفتـهٔ سـینگ (۲۰۱۶) سـه رویکـرد عمـده بـرای تعدیـل محتـوا وجـود دارد کـه عبارت‌انـد از:

۱. تعدیـل محتـوای دسـتی: ایـن رویکـرد کـه نوعـاً بـه اسـتخدام و آمـوزش و به‌کارگیـری گردانندگان انسـانی بـرای بازبینـی و اتخـاذ تصمیـم دربـارهٔ مـوارد محتـوا تکیـه دارد، شـکل‌های مختلفـی می‌توانـد داشـته باشـد. پلتفرم‌هـای بـزرگ تمایـل دارنـد در اصـل بـر کارکنـان قـراردادی برون‌مبتنـی بـر برون‌سـپاری بـرای انجـام ایـن کار تکیـه کننـد. پلتفـرم کوچـک تـا متوسـط از گردانندگان تمام‌وقـت و خانگـی را اسـتخدام می‌کننـد یـا اینکـه از گردانندگان کاربـری اسـتفاده می‌کننـد کـه داوطلبانـه بـه بازبینـی محتـوا می‌پردازنـد.

۲. تعدیـل محتـوای خـودکار: ایـن رویکـرد شـامل اسـتفاده از تصمیم‌گیـری، پالایـش و ابزارهـای تعدیـل خـودکار بـرای پرچم‌زنـی، جداسـازی و حـذف قطعه‌هـای خاصـی از محتـوا یـا حسـاب کاربـری اسـت. کردارهـای تعدیـل و کشـف محتـوای کامـل خـودکار به‌طـور گسـترده در همـه مقوله‌هـای محتـوای ناخوشـایند بـه کار نمـی‌رود؛ زیـرا مشـاهده شـده اسـت کـه آن‌هـا فاقـد دقـت و تأثیرگـذاری بـرای انـواع معینـی از گفتـار کاربـر هسـتند. گرچـه ایـن ابزارهـا بـرای انـواع معینـی از محتـوای ناخوشـایند / قابـل ایـراد، هماننـد محتـوای سوءاسـتفادهٔ جنسـی از کـودکان (سـی.اس.ای.ام) به‌طـور گسـترده بـه کار می‌رونـد.

۳. تعدیـل محتـوای پیونـدی: در ایـن رویکـرد، عناصـری از رویکردهـای دسـتی و خودکار باهـم ترکیـب می‌شـوند. معمـولاً ایـن روش شـامل اسـتفاده از ابزارهـای خـودکار بـرای پرچـم زنـی و اولویـت دادن بـه مـوارد محتوایـی خـاص بـرای مرورگـران انسـانی اسـت کـه آن‌هـا سـپس دربـارهٔ آن مـورد داوری نهایـی را انجـام می‌دهنـد. ایـن رویکـرد بـه شـکل وسـیع‌تری توسـط پلتفرم‌هـای کوچـک و بـزرگ اقتبـاس شـده اسـت؛ به‌طـوری کـه بـه کاهـش بـار کاری اولیـهٔ مرورگـران انسـانی کمـک می‌کنـد.

خودکار شدن و کاربرد هوش مصنوعی برای تعدیل محتوا

طی چند سال گذشته، پلتفرم‌های بزرگ و کوچک که میزبان محتوای کاربرساخته هستند، به‌دلیل فراگیری فزایندهٔ محتوای ناخوشایند، بیش از پیش ازسوی دولت‌هایشان تحت فشار قرار گرفتند تا این نوع محتواها را حذف کنند. در واکنش به این مسئله، بسیاری از شرکت‌ها ابزارهای خودکاری را برای ارتقای کردارهای تعدیل محتوای خود توسعه داده یا اقتباس کرده‌اند که بسیاری از آن‌ها با هوش مصنوعی و یادگیری ماشین تغذیه می‌شوند (Singh, 2019). هرچند باید خاطرنشان شود که تصمیم‌گیری‌های ظریف دربارهٔ اینکه کدام محتواهای کاربرساخته پذیرفتنی و کدام پذیرفتنی نیست، فراتر از توانایی فرایندهای ماشینی است و کاربرد فیلترهای الگوریتمی نیز هنوز در سطح پایینی از پیچیدگی قرار دارد؛ بنابراین ضرورت داوری دربارهٔ محتواهای کاربرساخته و به‌ویژه محتوای ویدئویی و تصویری نیازمند کنشگران انسانی است تا با تکیه بر توانایی‌های زبانی و آگاهی فرهنگی، دربارهٔ متناسب بودن محتوای کاربرساخته با اصول و هنجارهای حاکم بر یک سایت یا پلتفرم مشخص تصمیم بگیرند (Roberts, 2017).

یکی از مزیت‌های اصلی کاربرد روش خودکار، امکان انجام حجم عظیمی از بازبینی محتوا در زمان کوتاه و نیز سرعت دادن به فعالیتی است که در صورت انجام آن‌ها توسط انسان‌ها به مدت‌زمان زیادی نیاز است. ماشین‌ها به‌طور خستگی‌ناپذیر وظایف تعدیل خود را براساس برنامه‌ریزی زمانی و با موشکافی انجام می‌دهند و از قواعدی که به آن‌ها آموزش داده شده است، بدون هیچ استثنایی تبعیت می‌کنند. ریکشتین و ترونن (۲۰۲۰) ضعف و قوت رویکرد انسانی و ماشینی را باهم مقایسه کرده‌اند که این موارد در جدول ۱ ذکرشدند.

جدول ۱. تفاوت‌های تعدیل محتوای ماشینی و انسانی؛ منبع: ریکشتین و ترونن (۲۰۲۰)

	رفتار انسان		ماشین
نقاط قوت	ـ سرعت انطباق ـ همدلی ـ حساسیت نسبت به اطلاعات زمینه‌ای	ـ انعکاس تصمیماتی که قبلاً اتخاذ شده است و انطباق صرفاً به شکل تدریجی صورت می‌گیرد ـ ظرفیت محدود برای درک شوخی، کنایه و ریشخند	نقاط ضعف
نقاط ضعف	ـ محدودیت در سرعت و تعداد پیام‌های پردازش شده ـ عدم انسجام ـ آسیب‌پذیری به هنگام مواجه با محتوای غیرانسانی	ـ همزمان، ۷/۲۷ ـ کارآمد در پویش پایگاه‌های داده بزرگ ـ منسجم در قواعد مستخرج از داده‌های آموزشی ـ ازنظر روانی آسیب نمی‌بینند	نقاط قوت

تعدیل محتوای الگوریتمی شامل طیفی از تکنیک‌های گوناگون از علم آمار و علم رایانه است و از منظر پیچیدگی و مؤثر بودن متغیر هستند. هدف همه آن‌ها شناسایی، تطبیق، پیش‌بینی یا طبقه‌بندی برخی اجزای محتوا (متن، صوت، تصویر یا ویدئو) براساس ویژگی‌های دقیق یا ویژگی‌های عمومی آن‌ها است. البته براساس نوع تطبیق یا طبقه‌بندی مورد نیاز و نیز نوع محتوایی که بررسی می‌شود، تفاوت‌های عمده‌ای در تکنیک‌های به‌کاررفته وجود دارد (Rickstein and Tronen, 2020).

سینریش (۲۰۲۰) برخی از چالش‌های تعدیل خودکار یا ماشینی را ذکر کرده است. به نظر او، آشکارترین کاستی تعدیل محتوای خودکار افزایش ریسک موارد مثبت و منفی کاذب است. برای نمونه امکان دارد ویدئوی آموزشی دربارهٔ تغذیه با شیر مادر به‌عنوان تصویر هرزه‌نگاری تلقی شود؛ اما اطلاعات گمراه‌کننده تسلیحاتی ممکن است به‌عنوان گزارش خبری از منبعی معتبر در نظر گرفته شود. موضوع دیگر گسترش مقیاس است. قوانین و ارزش‌های فرهنگی که مقررات و رویهٔ قضایی را شکل می‌دهند از منظر مقیاس در سطوح محلی، منطقه‌ای و ملّی قرار دارند، اما پلتفرم‌ها در سطح جهانی عمل می‌کنند. برخلاف حکومت‌های قانونی، شرکت‌ها هیچ وظیفه‌ای برای حفظ ارزش‌های دمکراتیک ندارند و وظایف آن‌ها اغلب در تقابل با این ارزش‌ها قرار دارد؛ بنابراین امکان دارد بین رویه‌های تعدیل خودکار شرکتی و دولت‌ها تضادهایی ایجاد شود که ماشینی شدن تعدیل این فرایند را تشدید می‌کند.

محتواهای کاربرساختهٔ ویدئویی

محتواها ممکن است در قالب‌ها و به شکل‌های گوناگون همانند متن، تصویر، ویدئوها و اطلاعات موقعیت مبنا و فراداده‌های مرتبط باشند (Moons, Li and Shua, 2014). با اوج‌گیری رسانه‌های اجتماعی تصویرمحور، همانند یوتیوب و اینستاگرام، تولید، اشتراک‌گذاری و مصرف محتوای تصویری ــ عکس و ویدئو، بیش از پیش توسعه یافته است. به‌ویژه در دوران فراگیری بیماری کووید ۱۹، شاهد ازدیاد مصرف ویدئو بودیم که «طی آن پلتفرم‌های ارائه‌دهندهٔ خدمات ویدئویی، رسانه‌های اجتماعی تصویری (به‌ویژه پخش زنده اینستاگرام) و همین‌طور پیام‌رسان‌های مختص تماس ویدئویی در کانون آن قرار داشتند (حسنی، ۱۳۹۹: ۲۰۴).

براساس سند منتشرشده توسط آفکام (۲۰۲۰) یا ادارهٔ ارتباطات انگلستان، پلتفرم‌های اشتراک‌گذاری ویدئو (وی.اس.پی) نوعی خدمت ویدئویی برخط هستند که یکی از

ویژگی‌های اصلی این نوع خدمت این است که به کاربران اجازه می‌دهد ویدئوهای خود را بارگذاری کنند و آن را با کاربران و اعضای اجتماع به اشتراک بگذارند. همچنین آن‌ها به کاربران اجازه می‌دهند با طیف گسترده‌ای از محتوا و جنبه‌های اجتماعی درگیر شوند.

نیکولچف (۲۰۱۸) در گزارشی که برای نهاد ناظر رسانه‌های صوتی تصویری اتحادیهٔ اروپا نگاشته است، می‌نویسد اشتراک‌گذاری ویدئو در حال تطور است. مهم‌ترین ویژگی پلتفرم‌های اشتراک ویدئو دسترسی آزاد به این نوع محتوا برای همه، عدم درگیری پلتفرم در انتخاب محتوای منتشرشده، گزینش یا تعدیل الگوریتمی یا انسانی محتوا، کسب درآمد ازطریق آگهی و وارسی یا تعدیل پس از شکایت صاحبان حقوق یا توسط خود پلتفرم بوده است.

تولید و انتشار ویدئوها یکی از روندهای روبه‌رشد حوزهٔ محتوایی است که در قالب‌های گوناگون همچون پخش برخط (استریم)، بازی‌های ویدئویی و پخش زنده عرضه می‌شود. شیوع کرونا و افزایش فاصله‌گذاری اجتماعی و قرنطینه اجباری براساس سیاست‌های دولت‌ها سبب شده است، حوزهٔ صوت‌وتصویر و به‌ویژه پخش آنلاین از این تحولات متأثر شود و مصرف ویدئو با افزایش چشم‌گیری روبه‌رو شده است. تولید محتوای کاربرساخته از روندهای روبه‌رشد در پلتفرم‌های آنلاین و تصویری خواهد بود. برخی از انواع این پلتفرم‌ها عبارت‌اند از:

- محتوای زودگذر: استوری‌ها نقش مهمی در به‌روزرسانی رسانه‌های اجتماعی برعهده دارند و بخش عمده و مهمی از مصرف محتوا را تشکیل می‌دهند. این نوع محتوا به رایج‌ترین روش اشتراک محتوا در فیس‌بوک در سال ۲۰۱۸ تبدیل شده است.
- پخش محتوای زنده: یکی از روندهای روبه‌گسترش در حوزهٔ محتوا، پخش زنده و محتوای زنده است که یکی از روندهای مهم در بستر پلتفرم‌های بزرگ همانند یوتیوب است. امروزه هرکدام از کاربران در رسانه‌های اجتماعی که این خدمت را ارائه می‌دهند، از قابلیت پخش زنده برخوردار شده‌اند. پخش زنده یکی از چالش‌های مهم پیش روی گردانندگان محتواست (اکبری نوری، ۱۳۹۹).

تحولات اشتراک‌گذاری ویدئوی برخط و ازجمله توسعهٔ انتشار زنده در رسانه‌های اجتماعی گوناگون، به‌عنوان جایگزین برای رسانه‌های ارتباط‌جمعی، جریان اصلی مخاطراتی را از حیث نظارت و تنظیم مقررات ایجاد کرده است. این امکان وجود دارد که به شکل بالقوه بی‌نهایت «خرده تلویزیون برخط» به انتشار محتوا بپردازند. این موضوع به‌ویژه برای دولت‌هایی که نوعی سیاست رسانه‌ای متمرکز و اعمال مقررات سخت‌گیرانه را برای انتشار محتوای ویدئویی پیگیری می‌کنند، مخاطرات جدی ایجاد می‌کند.

چالش تعدیل محتواهای ویدئویی جدید

تعدیل یا نظارت بر محتوای ویدئویی که به‌صورت زنده پخش می‌شود، ضروری است. شبکه‌های اجتماعی، پلتفرم‌های اشتراک‌گذاری محتوا و اجتماعات بازی‌های برخط سبب محبوبیت پخش زنده شده‌اند. برای اطمینان از محافظت از بینندگان محتوایی که به‌صورت زنده منتشر می‌شود، باید ابزارهای تعدیل ویدئوی زنده توسعه یابند؛ گرچه نظارت و تعدیل این نوع محتوا نسبت به سایر انواع محتوای غیرزنده چالش‌برانگیزتر است. سطح زیان‌آور بودن ممکن است به‌سرعت اوج بگیرد یا اینکه عناصر پیشین و کنونی محتوا برای بررسی در دسترس قرار نداشته باشد. برای دستیابی به تعدیل در زمان واقعی به سیستم‌های بسیار بهینه‌ای نیاز داریم که هم تصاویر را به شکل قاب‌به‌قاب و همین‌طور صدای همراه آن را تحت نظارت قرار دهند.

وینچکامب (۲۰۱۹) برخی انواع این نوع محتواها را ذکر کرده است که از چت زنده تا دیپ فیک یا جعل عمیق را در می‌گیرد. این موارد نشان می‌دهد که فرایند نظارت بر محتواهای زنده و نیز انواع جدید محتوا تا چه اندازه دشوار است و پیچیدگی آن نیز افزوده می‌شود. در جدول ۲ برخی از این قالب‌های جدید محتوا ذکر شده است.

جدول ۲. چالش‌های جدید تعدیل محتوا (منبع: وینچکامب، ۲۰۱۹)

قالب	توصیف	مثال	انواع رسانه تشکیل‌دهنده			
			متن	تصویر	ویدئو	صوت
چت زنده	متن برخطی که به‌طور همزمان به اشتراک گذشته می‌شود	خدمات پیام‌رسان آنی و چت‌روم‌های برخط	*	–	–	–
ویدئوی زنده	ویدئویی که به‌طور همزمان بارگذاری و منتشر می‌شود	«داستان‌های» رسانه‌های اجتماعی	–	–	*	*
گیف	تصویری با قاب‌های گوناگون که در قالب یک فایل تصویر رمزگذاری می‌شود	تصویر پویانمایی‌شده که یک صحنه فیلم را نشان می‌دهد	*	*	–	–
میم	تصویر، گیف یا ویدئویی که با شرح عکسی همراه می‌شود که غالباً توسط کاربران اینترنت به اشتراک گذاشته می‌شود	تصویری که به برچسبی خورده است و به یک ترانه عامه‌پسند یا سخن مشهور اشاره دارد	*	*	*	–
دیپ فیک	تصویر، صوت و ویدئو و به‌طور بالقوه متنی که به کمک هوش مصنوعی سنتز (ترکیب) شده است	ویدئوهای جعلی از سیاست‌مداران، کنشگران و سلبریتی‌ها که هرگز در واقعیت اتفاق نمی‌افتد	*	*	*	*

برای نمونه، دیپ فیک‌ها (جعل عمیق) به‌طور بالقوه می‌توانند بسیار زیان‌آور باشند. کشف آن‌ها دشوار است. جعل عمیق از تکنیک‌های یادگیری ماشینی برای ایجاد محتوای جعلی استفاده می‌کند که می‌تواند به‌صورت برخط منتشر شود. امکان دارد جعل‌های عمیق تصویر، ویدئو، صوت یا متن تولیدشده از مجموعهٔ داده‌های فعلی را باهم ترکیب کند. از این تکنیک می‌توان برای ایجاد نسخه‌های رایانه‌ای ساخته از سیاست‌مداران، کنشگران و سلبریتی‌ها، برای نمونه، برای شبیه‌سازی رویدادهایی استفاده کرد که هرگز در واقعیت روی نداده‌اند. جعل‌های عمیق ابزاری نیرومند و زیان‌آور هستند که از آن‌ها می‌توان برای گمراه کردن مخاطبان استفاده کرد تا به‌وسیلهٔ محتوای برخط تغییریافته و گمراه‌کننده، آنچه را می‌بینند، باور کنند. این محتواها می‌توانند برای گمراه کردن مخاطبان، ایجاد دستورکارهای سیاسی و ایجاد محتوای زیان‌آور در فضای برخط به‌کار روند. با بهبود تکنیک‌های یادگیری ماشینی و دسترسی به آموزش و یادگیری با داده‌ها، سیاست‌های تعدیل محتوا باید ابزارهایی برای کشف این نوع محتواهای پیشرفته توسعه دهند.

روش‌شناسی

روش انجام این پژوهش کیفی و از نوع مطالعهٔ موردی است. مطالعه موردی روشی برای پژوهش است که شامل بررسی عمیق و دقیق موضوع خاصی مانند فرد، گروه، رویداد، سازمان یا پدیده است (Crave et al. 2011). مطالعات موردی از انواع روش‌های جمع‌آوری داده‌ها مانند مصاحبه، مشاهدات، تحلیل اسناد و سوابق بایگانی برای جمع‌آوری اطلاعات جامع دربارهٔ موضوع مورد علاقه استفاده می‌کنند. هدف از مطالعهٔ موردی ارائهٔ درک متنی و غنی از موضوع، بررسی روابط میان متغیرها، شناسایی الگوها و ایجاد بینش‌هایی است که می‌تواند به توسعهٔ نظریه یا راه‌حل‌های عملی کمک کند.

هدف این مطالعه که در سال ۱۴۰۲ تا ۱۴۰۳ انجام شده، این است که با مطالعهٔ دقیق و همه‌جانبهٔ چگونگی انجام فرایند تعدیل محتوا که با تمرکز بر یوتیوب انجام می‌شود، بتوان سازوکار این کردار را به‌ویژه درمورد محتواهای ویدئویی و چگونگی کاربرد تعدیل محتوای الگوریتمی نشان داد که به‌کمک هوش مصنوعی انجام می‌شود، نشان دهد. این مطالعه شامل تحلیل اسناد و سیاست‌های پلتفرم یوتیوب است. برای گردآوری داده‌ها، منبع اصلی اسناد و مدارک خود یوتیوب است که شامل اصول

راهنمـای اجتماعـی، خط‌مشـی‌های مختلـف، گزارش‌هـای منتشرشـده و نیـز مطالعـات و گزارش‌هایـی اسـت که هـدف آن تحلیـل تعدیـل محتـوا در یوتیـوب بوده‌انـد. تحلیـل داده‌هـا به‌صـورت تحلیـل محتـوای کیفـی انجـام شـده اسـت و مقوله‌هـای اصلـی پـس از مـرور اسـناد مختلـف اسـتخراج و بـا نشـان دادن شـواهد اسـتخراج و تحلیـل شـده‌اند.

تحلیل یافته‌ها

مطالعـه دربـارهٔ خط‌مشـی تعدیـل محتـوا در یوتیـوب از ایـن جهـت اهمیـت دارد کـه ایـن رسـانهٔ اجتماعـی را در اصـل به‌عنـوان پلتفـرم ویدئویـی می‌شناسـیم. برخـلاف متـن نوشـتاری کـه شـامل مجموعـه‌ای از واژگان اسـت و شناسـایی مطالـب غیرقانونـی و غیراخلاقـی در آن، هـم بـرای عوامـل انسـانی و هـم ماشـینی سـاده اسـت، تصویـر ماهیتـی بسـیار پیچیـده دارد. انـواع قالب‌هـای جدیـدی کـه بـه محتـوای تصویـری افـزوده می‌شـوند، هماننـد گیـف و میـم و...، نظـارت بـر ایـن نـوع محتـوای کاربرسـاخته را پیچیده‌تـر و حسـاس‌تر می‌کنـد، به‌ویـژه بـه ایـن دلیـل کـه محتـوای ویدئویـی بیـش از پیـش محبوبیـت پیـدا می‌کنـد و اصلی‌تریـن رسـانه تعامـل و ارتبـاط در رسـانه‌های اجتماعـی تبدیـل می‌شـود. به‌همین‌دلیـل نظـارت بـر آن نیـز دشـوارتر و پیچیده‌تـر اسـت و بیـش از پیـش نیـز پیچیده‌تـر می‌شـود.

پلتفرم یوتیوب

یوتیـوب در کنـار فیس‌بـوک و توییتـر را می‌تـوان در زمـرهٔ نسـل جدیـد و مثالـی شـبکه‌های اجتماعـی در نظـر گرفـت کـه علاوه‌بـر اینکـه توانسـتند جایـگاه خـود را تثبیـت کننـد و مدلـی بـرای درآمدزایـی بیابنـد کـه زمینـهٔ مانـدگاری آن‌هـا را فراهـم کـرد، رفته‌رفتـه بـه پلتفرم‌هـای بزرگـی تبدیـل شـدند (Hash, 2021). یوتیـوب در اصـل بـا ایـدهٔ سـاده‌ای آغـاز شـد؛ امـا رفته‌رفتـه جایـگاه خـود را به‌عنـوان پلتفـرم جهانـی تثبیـت کـرد. هـدف اصلـی یوتیـوب «حـذف موانـع فنـی پیـش روی کاربـران غیرمتخصصـی بـود کـه می‌خواسـتند ویدئوهـای خـود را در وب بارگـذاری کننـد. ایـن وب‌سـایت واسـطی بسـیار سـاده و یکپارچـه را فراهـم کـرده بـود کـه بـه افـراد اجـازه می‌داد بـدون نیـاز بـه دانـش فنـی زیـادی و بـا اسـتفاده از مرورگرهـای وب اسـتاندارد و سـرعت متوسـط اینترنـت، ویدئوهـا را بارگـذاری و منتشـر و نیـز ویدئوهـا را بـه برخـط تماشـا کننـد» (Burgess, 2018: 13).

فـن دایـک (۱۳۹۶) بـرای معرفـی ایـن پلتفـرم از «یوتیـوب: اتصـال صمیمانـه بیـن تلویزیـون و بـه اشـتراک‌گذاری ویدئـو» اسـتفاده می‌کنـد و می‌نویسـد: «ایـن سـایت به‌عنـوان

پلتفرم برای به اشتراک‌گذاری ویدئوهای آماتور خودساخته و نیز به‌عنوان «جایگزینی» برای تماشای تلویزیون بزرگ شد. یوتیوب در همهٔ سطوح جایگزین بود: فناوری متفاوت، تغییر جهتی در روال‌های روزمرهٔ کاربران، نوع جدیدی از محتوا و اصلاح ریشه‌ای صنعت سنتی پخش رادیوتلویزیونی که شامل مدل‌های تجاری آن نیز هست» (فن دایک، ۱۳۹۶: ۲۰۷).

پلتفرم‌هایی همانند یوتیوب و تیک‌تاک به‌دلیل آنکه تصویرمحور هستند، چالش‌های جدیدی را برای حکمرانی پلتفرم‌ها ایجاد می‌کنند. ازمنظر حکمرانی بر پلتفرم‌ها، ازآنجاکه یوتیوب در یک زمینه‌ای ـ سیاسی فعالیت می‌کند که مدل تنظیم‌گری آن بازار ـ محور است (Bradford, 2023)، عموماً حکمرانی بر این پلتفرم اندک و احتمالاً محدود به چالش‌های امنیت ملی است، اما از منظر حکمرانی در پلتفرم تا حد زیادی تابع شرکت گوگل است؛ هرچند رویه‌های حکمرانی ویژه خود را نیز توسعه داده است. از طرف دیگر، یوتیوب یک پلتفرم مبتنی‌بر محتوای کاربرساخته است. کاربران گوگل اجتماعی جهانی از کاربران با ویژگی‌های جمعیت‌شناختی، فرهنگی و زبانی گوناگون را تشکیل می‌دهند و انواع ژانرهای مختلف ویدئو و با فرمت‌های گوناگون را در آن با اهداف گوناگون به اشتراک می‌گذارند؛ بنابراین، حکمرانی مؤثر برای اطمینان از اینکه پلتفرم برای کاربران عادی، تولیدکنندگان محتوا و جامعه به‌طورکلی ایمن، دربرگیرنده و ارزشمند باشد، دشوار و مستلزم کاربرد سازوکارهای تنظیم‌گری محتوایی پیچیده است.

براساس مقوله‌بندی کلی اسناد و رویه‌های تعدیل محتوا در یوتیوب برخی مقوله‌های کلی شناسایی شدند که در ادامه ذکر می‌شوند و براساس زیرمقوله‌های مرتبط تحلیل می‌شوند.

اسناد هدایتگر و بالادستی تعدیل محتوا

به‌طورکلی برخی اسناد و دستورالعمل‌ها به‌مثابهٔ نوعی قانون اساسی پلتفرمی عمل می‌کنند و برخی دیگر را می‌توان آیین‌نامه یا دستورالعمل عمل تعدیل محتوا قلمداد کرد.

شرایط ارائه خدمت

پیش‌از بحث دربارهٔ شیوه‌هایی که یوتیوب برای تعدیل محتواهای کاربرساخته و به‌طور خاص محتواهای ویدئویی به‌کار می‌برد، مراجعه و بحث دربارهٔ اسناد و دستورالعمل‌های خود این پلتفرم‌ها از این نظر اهمیت دارد که مبنای عمل آن‌ها است.

در سایت یوتیوب بخشی به شرایط ارائه خدمت اختصاص دارد که گفته شده در ۵ ژانویهٔ ۲۰۲۲ به‌روزرسانی شده است. نسخهٔ پیشین شرایط ارائه خدمت در ۱۷ مارس ۲۰۲۱ به اجرا گذاشته شده است. این نسخه مبنای این مقاله است که مضامین کلی آن را به شکل موضوعی براساس هدف پژوهشی مرور می‌کنیم.

شرایط ارائه خدمت یوتیوب در اصل سندی طولانی نیست؛ اما با مطالعهٔ آن می‌توان تصویری کلی از رابطهٔ متقابل روابط تعریف‌شده میان یوتیوب و کاربران و نیز انتظارات این پلتفرم از کاربران را دریافت کرد. همچنین الزامات استفاده از این خدمت و نیز مقوله‌های مختلف کاربران در آن مشخص شده است. همچنین حقوق کاربران به هنگام استفاده از این رسانهٔ اجتماعی و نیز شرایطی که به این استفاده اعمال می‌شود، نیز بیان شده است. علاوه‌بر این، شرایط بارگذاری محتوا و نیز اصول راهنمای رفتار اجتماعی[1] نیز ذکر شده است (YouTube, 2021).

به‌طورکلی یوتیوب که خود را به‌مثابهٔ خدمت تعریف کرده است، به کاربران امکان می‌دهد تا «ویدئوها و محتواهای دیگر را کشف کنند، تماشا کنند و به اشتراک بگذارند و همین‌طور محلی را فراهم کرده است تا افراد باهم متصل شوند، آگاه شوند و الهام‌بخش دیگران در سطح جهان شوند؛ همچنین یوتیوب به‌عنوان پلتفرم توزیع برای خالقان محتوای اصلی و آگهی‌دهندگان بزرگ و کوچک عمل می‌کند».

به‌طورکلی براساس سند شرایط ارائه خدمت یوتیوب هر نوع استفاده کاربران از این پلتفرم یا خدمت مشمول چند نظام‌نامه یا سند است. این مقررات عبارت‌اند از: اصول راهنمای اجتماع یوتیوب[2]، خط‌مشی[3]، خط‌مشی‌های ایمنی و حق تألیف[4] که به شکل مستمر به‌روزرسانی می‌شوند و البته چنانچه کاربران آگهی منتشر کنند یا اینکه برای محتوای منتشرشده در یوتیوب از پشتیبان مالی استفاده نمایند، خط‌مشی‌های آگهی دهی[5] در یوتیوب نیز به آن‌ها اعمال می‌شود. علاوه‌بر این، باید به خط‌مشی حریم خصوصی، هشدار حریم خصوصی بخش کودکان یوتیوب[6]، شرایط پردازش داده در یوتیوب[7]، اصول راهنما و خط‌مشی‌های محتوای یوتیوب[8] نیز اشاره کرد.

1. community guidelines
2. YouTube Community Guidelines
3. Policy
4. Safety and Copyright Policies
5. Advertising on YouTube Policies
6. YouTube Kids Privacy Notice
7. YouTube Processing Terms
8. YouTube's content Policies and Guidelines

یکی از دلایل پیچیدگی و تنوع مقررات تنظیم نحوهٔ اشتراک‌گذاری محتوا در یوتیوب، تحولات سریعی است که در این حوزه در حال وقوع است. قانون جرائم رایانه‌ای ایران مصوب سال ۱۳۸۸ است. از آن زمان تاکنون دوازده سال گذشته است. طی این دوره‌ای که بیش از ده سال به طول انجامیده است، تحولات و تغییرات بسیاری در حوزهٔ رسانه‌های جدید، ازجمله فراگیری رسانه‌های اجتماعی، پلتفرم‌های برخط و کارافزارهای تلفن همراه هوشمند رخ داده است که به‌طور طبیعی ماهیت، کارایی‌ها و ویژگی‌های آن‌ها، زمینهٔ وقوع جرائم، اعمال غیراخلاقی و ناهنجاری‌های جدیدی را فراهم کرده است که به هیچ وجه با یک دههٔ گذشته مقایسه نیست. با توجه به‌کندی دولت‌ها و نهاد قانون‌گذاری در تدوین و تصویب قوانین جدید، متولی بخش عمده‌ای از مقابله با محتواهای زیان‌آور پلتفرم‌ها هستند که ازطریق تدوین مقررات و اعمال تعدیل محتوا، کارکرد حکمرانی خود را انجام می‌دهند.

یوتیوب نیز منطق بر این فرایند و انتظار طبیعی پلتفرم‌ها برای مشارکت در حکمرانی پلتفرم‌ها، برخی محدودیت‌ها را برای فعالیت کاربران در این خدمت وضع کرده است که در همان سند شرایط خدمت می‌توان آن‌ها را مشاهده کرد. یوتیوب اعلام کرده است که امکان دسترسی و استفاده از این خدمت تا زمانی میسر است که افراد از این مقررات تبعیت کنند. کاربران یوتیوب از انجام اعمال زیر منع شده‌اند:

۱. دسترسی، بازتولید، بارگیری، توزیع، انتقال، پخش، نمایش، فروش، سوءاستفاده، تغییردادن، تغییر شکل یا هر نوع استفاده از خدمات یا هر نوع محتوا؛ مگر اینکه خدمت آشکار مجوز آن را داده باشد یا مجوز کتبی پیشین از یوتیوب و در صورت لزوم کسب اجازه از صاحبان حقوق؛

۲. درمعرض خطر قراردادن، غیرفعال کردن، کاربرد متقلبانه یا ایجاد هرگونه اختلال در هریک از بخش‌های خدمات (و هر نوع تلاش به‌منظور انجام این اعمال)، ازجمله دربارهٔ جنبه‌های امنیتی یا ویژگی‌هایی که الف) نسخه‌برداری یا سایر کاربردهای محتوا را محدود می‌کند یا از آن جلوگیری می‌کند یا ب) استفاده از خدمت یا محتوا را محدود می‌کند.

۳. گردآوری یا انباشت اطلاعاتی که هویت یک شخص را برملا می‌کند، مثلاً نام کاربری یا چهره، مگر اینکه آن شخص اجازه داده باشد یا باکسب اجازه از یوتیوب.

این موارد به‌طورکلی شامل ملاحظاتی دربارهٔ سوءاستفاده از این خدمات و آسیب رساندن به آن است و نیز شامل نقض حریم خصوصی کاربران یوتیوب که البته مقررات

دیگری نیز وجود دارند که جزئیات این محدودیت‌ها را بیان می‌کنند. ازآنجاکه هدف این مطالعه بیشتر مناقشه مرتبط با محتواهای زیان‌آور و نحوه مقابله با آن‌ها است، بیشتر بر مقررات یوتیوب از این منظر تمرکز می‌کنیم و بخش‌های دیگر این سند ازجمله مواردی که یوتیوب اعلام کرده است در صورت آسیب رسیدن به کاربر، این پلتفرم در برابر آن هیچ مسئولیتی ندارد، هدف این نوشتار نیست.

دستورالعمل‌های اجتماعی یوتیوب

یکی از اسناد مهمی که معمولاً در تعیین مداخله یا عدم مداخله پلتفرم در روال جریان محتوای منتشر شده نقش دارد، اصول راهنما یا دستورالعمل اجتماع است که معمولاً پلتفرم‌های بزرگ ازجمله فیس‌بوک نسخه‌های به‌روزشده آن را منتشر می‌کنند؛ گرچه معمولاً تصور می‌شود عموم کاربران از این دستورالعمل‌ها به‌اصطلاح پرش می‌کنند و آن‌ها را نادیده می‌گیرند.

در بخش مرور کلی دستورالعمل اجتماعی یوتیوب (YouTube Community Guidelines, 2021) هدف از این دستورالعمل تعیین «محتوای غیرمجاز» ــ که البته با محتوای مجرمانه در مقررات حقوقی ناظر بر فعالیت اینترنت در ایران متفاوت است ــ در یوتیوب ذکر شده است. این خط‌مشی بر تمامی انواع محتوا در این پلتفرم شامل «ویدئوها، کامنت‌ها، لینک‌ها و علامت تأیید یا لایک» اعمال می‌شود. این دستورالعمل‌ها بخشی از مجموعه گسترده‌تری از خط‌مشی‌ها (YouTube, 2024) هستند که مواردی ازجمله همین دستورالعمل اجتماع، حق تألیف، خط‌مشی‌های درآمدزایی و حذف قانونی محتوا را دربرمی‌گیرند یوتیوب این اصول اجتماعی را به کمک ترکیبی از ابزارهای گوناگون ازجمله ناظران انسانی و یادگیری ماشینی، بدون ملاحظه پس‌زمینهٔ سیاسی و وابستگی افراد اعمال می‌کند و هدف از آن تبدیل یوتیوب به یک «اجتماع امن‌تر» است.

مجموعه محتواهای غیرمجاز در یوتیوب

اصول راهنمای اجتماعی یوتیوب شامل چندین دسته است که عبارت‌اند از:

اسپم و کردارهای فریبکارانه، محتوای حساس، محتوای خشونت‌آمیز یا خطرناک، کالاها و اجناس مشمول مقررات و اطلاعات نادرست.

هرکدام از این دسته‌ها شامل زیرمجموعه‌های مختلفی است و موارد مختلفی در ذیل آن قرار دارد که به‌طور اختصار برخی از آن‌ها را مرور می‌کنیم. دربارهٔ هرکدام از این موارد تعریف، نمونه‌ها و نیز شیوهٔ تعدیل محتوا توسط یوتیوب مرور می‌شود.

اسپم و کردارهای فریبکارانه[1]

فعالیت جعلی[2]

دخیل کردن تصنعی کاربران یوتیوب از طریق تعداد بازدیدها، لایک‌ها و کامنت‌ها یا روش‌های دیگر با استفاده از سیستم‌های خودکار یا نمایش دادن ویدئوها برای کاربران بی‌اطلاع و نیز محتواهایی که هدف آن‌ها صرفاً برانگیختن بازدید باشد، ممنوع است.

مثال‌ها: یک ویدئوی گواهی‌دهنده که طی آن یک تولیدکننده خود را در وضعیتی نشان می‌دهد که به شکلی موفقیت‌آمیز ترافیک صفحه تصنعی را از یک طرف سوم خریداری می‌کند.

روش برخورد یوتیوب در صورت وقوع این تخلف (تعدیل محتوا):

در صورت وقوع این تخلف، یوتیوب محتوا را حذف و ضمن ارسال یک ایمیل فرد از این اقدام مطلع می‌شود. چنانچه بار نخست باشد، فرد فقط هشداری را دریافت می‌کند و مجازاتی متوجه وی نیست.

موارد دیگر در قالب جدول ۳ فقط تعریف می‌شوند و روش تعدیل محتوای یوتیوب ذکر می‌شود:

جدول ۳. انواع اسناد تعدیل محتوا در یوتیوب.

عمل	مفهوم
جعل هویت[1]	محتوایی که هدف آن جعل هویت یک شخص یا کانال است در یوتیوب مجاز نیست. همچنین از حقوق دارندهٔ علامت تجاری نیز محافظت می‌شود.
لینک خارجی[2]	لینک‌هایی که کاربران یوتیوب را به‌سوی وب‌سایت‌های دارای محتوای غیرمجاز هدایت می‌کنند که ناقض اصول راهنمای اجتماعی این پلتفرم است غیرمجاز هستند.
هرزنامه‌ها، کردارهای فریبکارانه و کلاهبرداری‌ها[3]	هرزنامه‌ها، کردارهای فریبکارانه و کلاهبرداری‌هایی که هدف آن سوءاستفاده از اجتماع کاربران یوتیوب هستند غیرمجاز هستند.
فهرست پخش[4]	فهرست پخش روش مناسبی برای ترکیب ویدئوهایی است که کاربران معمولاً تمایل دارند به شکلی مجموعه آن‌ها را تماشا کنند اما درصورتی‌که این فهرست پخش حاوی محتوای غیرمجاز باشد یا به اجتماع کاربران یوتیوب آسیب بزند جلوی آن گرفته می‌شود.

1. Spam & deceptive practices
2. Fake engagement
1. Impersonation
2. External link
3. Spam, deceptive practices & scams
4. Playlist policy

عمل	مفهوم
خط‌مشی‌های دیگر	علاوه‌بر این‌ها برخی خط‌مشی‌های دیگر نیز مشخص شده‌اند که در صورت نقض آن‌ها با تعدیل محتوا از طرف یوتیوب مواجه می‌شوند که عبارت‌اند از: ـ حساب‌های کاربری غیرفعال ـ تشویق به نقض شرایط ارائه خدمت ـ انتشار مجدد محتواهایی که پیشتر حذف شده بودند یا محتواهای تولیدشده توسط تولیدکنندگان مسدود شده یا افرادی که به فعالیت آن‌ها خاتمه داده شده است.

محتوای حساس[1]

ایمنی کودکان[2]

یوتیوب اجازه انتشار محتواهایی را که بهزیستی عاطفی و فیزیکی افراد صغیر ـ افراد زیر ۱۸ سال ـ را به مخاطره می‌اندازد، نمی‌دهد. این دسته شامل این موارد است: جنسی‌سازی کودکان؛ یعنی محتوایی که شامل رابطه جنسی کودکان و سوءاستفاده جنسی از کودکان است؛ کنش‌های زیان‌آور و خطرناک که شامل افراد صغیر است؛ ایجاد فشار عاطفی به افراد صغیر؛ محتواهایی با مضمون جنسی و خشونت‌آمیز که خانواده را هدف گرفته است و نیز آزار و اذیت سایبری که کودکان را هدف گرفته است.

مثال: ویدئویی که افراد زیر سن قانونی را در وضعیتی به تصویر می‌کشد که فعالیت‌های شهوت‌انگیز، جنسی یا شامل پیشنهاد جنسی را نمایش می‌دهد و نیز شامل اعمال جنسی همانند بوسه و فعالیت جنسی گروهی است.

در صورت وقوع این تخلف، یوتیوب محتوا را حذف و ضمن ارسال یک ایمیل فرد از این اقدام مطلع می‌شود. چنانچه بار نخست باشد، فرد فقط هشداری را دریافت می‌کند و مجازاتی متوجه وی نیست، اما در صورت تکرار اخطار شدیدی به فرد مرتکب داده می‌شود و در صورت اینکه فرد سه بار اخطار شدید و طی نود روز دریافت کند، به فعالیت کانال وی خاتمه داده می‌شود.

موارد دیگر که در ذیل محتوای حساس قرار گرفته‌اند که در جدول ۴ ذکر شده‌اند و عبارت‌اند از:

1. Sensitive content
2. Child safety

جدول ۴. انواع محتواهای حساس در سایت یوتیوب

عمل	مفهوم
عکس‌های کوچک[1]	منظور از عکس‌های کوچک شامل مواردی است که دربردارندهٔ تصاویر هرزه‌نگارانه است و اعمال جنسی، برهنگی و سایر تصاویر جنسی است. همین‌طور تصاویر خشونت‌آمیز و همین‌طوری محتواهایی که دربردارندهٔ فحاشی است هم شامل عکس‌های کوچک است.
عریانی و محتواهای جنسی[2]	محتوایی که آشکارا منظور از آن ایجاد برانگیختگی جنسی اجازه نمایش در یوتیوب را ندارد. انتشار تصاویر هرزه‌نگارانه می‌تواند به حذف محتوا یا خاتمه فعالیت کانال منجر شود.
خودکشی و خودآزاری[3]	یوتیوب اجازهٔ اشتراک‌گذاری محتواهایی را که خودکشی و خودآزاری ترویج می‌شود، نمی‌دهد که منظور ویدئوهایی است که سبب ایجاد شوک روانی و انزجار می‌شوند یا بینندگان را در معرض خطر قرار می‌دهند.
الفاظ رکیک[4]	محتواهایی که حاوی الفاظ رکیک هستند و برای افراد زیر ۱۸ سال مناسب نیستند و شامل مواردی است که طی آن آشکارا از الفاظ یک گفتار جنسی استفاده شده است. در ویدئویی که از کفرگویی افراطی استفاده شده است و از الفاظ کفرآمیز سنگین در عنوان ویدئو یا تصویر استفاده شده باشد.

محتوای خشن یا خطرناک[5]

آزار و قلدری سایبری[6]

یوتیوب اجازه نمی‌دهد تا محتواهایی که افراد در آن‌ها تهدید می‌شوند در یوتیوب به نمایش درآیند. همچنین محتواهایی که در آن‌ها افراد با حمله‌های طولانی‌مدت و کینه‌توزانه و بر اساس ویژگی‌های غریزی هدف گرفته می‌شوند اجازه انتشار در یوتیوب را ندارند. همچنین محتواهایی که هدف آن شرمسار کردن، فریب دادن یا حمله به افراد صغیر است نیز نباید در یوتیوب منتشر شوند.

سایر مواردی که در این دسته قرار می‌گیرند شامل مواردی هستند که در جدول ۵ ذکر شده‌اند:

1. Thumbnails
2. Nudity & sexual content policies
3. Suicide & self-harm
4. Vulgar language
5. Violent or dangerous content
6. Harassment & cyberbullying

جدول ۵. انواع محتواهای خشن تعریف شده در سایت یوتیوب.

عمل	مفهوم
محتوای زیان‌آور یا خطرناک[1]	محتواهایی که در آن‌ها تشویق به انجام فعالیت‌های خطرناک یا غیرقانونی انجام می‌شود که خطر ایجاد آسیب فیزیکی جدی یا مرگ را در پی دارد.
گفتار نفرت‌پراکنانه[2]	گفتار نفرت‌پراکنانه در یوتیوب ممنوع است و هر نوع محتوایی که خشونت یا نفرت علیه یک فرد یا گروه را بر اساس سن، کاست، معلولیت، قومیت، هویت جنسیتی، ملیت، نژاد، وضعیت مهاجرت، دین، جنس/جنسیت، گرایش جنسی، وضعیت خدمت سربازی و قربانی بودن در یک رویداد بزرگ را تشویق نمایند توسط این پلتفرم حذف می‌شود.
سازمان‌های جنایی خشن[3]	محتواهایی که هدف آن تقدیس، ترویج یا کمک به سازمان‌های جنایی است نباید در یوتیوب منتشر شود. محتواهایی که توسط سازمان‌های تروریستی یا جنایی خشن نیز تولید می‌شوند اجازه انتشار در یوتیوب را ندارند.
محتوای خشن یا خشونت‌آمیز[4]	محتوای خشن یا دارای خونریزی که هدف آن ایجاد شوک یا انزجار در بینندگان است و یا محتوایی که دیگران را تشویق می‌کند تا اعمال خشونت‌آمیز مرتکب شوند مجاز نیستند در یوتیوب منتشر شوند.

یک موضوع مهم دربارهٔ محتواهای غیرمجاز در یوتیوب نحوهٔ اعمال اصول راهنمای اجتماعی[5] است. یوتیوب در قالب یک ویدئو تلاش کرده است تا نحوهٔ این عمل را توضیح دهد. همان‌طور که یوتیوب در این بخش توضیح داده است، در هر دقیقه در حدود پانصد ساعت ویدئو در یوتیوب بارگذاری می‌شود. به‌همین‌دلیل یوتیوب توان سیستم‌های یادگیری ماشینی و اعضای اجتماع کاربران یوتیوب را برای علامت‌گذاری و هشدار نسبت به محتواهایی که به شکل بالقوه مسئله‌ساز هستند با هم ترکیب کرده است. سپس سیستم‌های یادگیری ماشینی و بازبینان متخصص محتوای علامت‌گذاری را که برخلاف اصول راهنمای اجتماعی هستند، حذف می‌کنند (YouTube, 2021).

همان‌طور که مشاهده می‌شود، عمل تعدیل محتوا در یوتیوب دو مرحله است. مرحله اول شامل شناسایی یا نشان‌دار کردن[6] محتوای غیرمجاز و مرحله دوم شامل حذف است. مرحله به کمک فناوری یادگیری ماشینی و عموم کاربران یوتیوب رخ می‌دهد. مرحله دوم حذف این نوع محتواها است که بدین منظور متخصصان عضو یوتیوب و نیز فناوری ماشینی این عمل را برعهده دارند.

1. Harmful or dangerous content
2. Hate speech
3. Violent criminal organizations
4. Violent or graphic content
5. Enforcement of Community Guidelines
6. flagging

در ویدئویی که یوتیوب با عنوان «یوتیوب چگونه خط‌مشی‌ها را اعمال می‌کند»[1]، منتشر کرده است، مراحل مختلف این عمل با جزئیات بیشتری شرح داده شده است (YouTube, 2021). همان‌طور جنیفر فلانوری اوکانر[2] مدیر اعتماد و ایمنی مدیریت محصول در یوتیوب و مت هلپرین[3] معاون و مسئول جهانی بخش اعتماد و ایمنی توضیح می‌دهند، بخش فراوانی از محتوای بسیار زیادی که در تمام اوقات شبانه‌روز در طول هفته از تمام کشورهای جهان در این پلتفرم بارگذاری می‌شوند، مبتنی‌بر خط‌مشی اصول راهنمای اجتماعی یوتیوب هستند. فقط بخشی اندکی از محتواها مسئله‌ساز هستند و اصل حفاظت از اجتماع کاربران یوتیوب را که هدف دستورالعمل اجتماعی است، نقض می‌کنند.

کردارهای عملی مقابله با محتواهای زیان‌آور

اگر اسناد هدایتگر را بتوان پردازش نظری تعدیل محتوا در یوتیوب قلمداد کرد، مواجه و مقابله عملی با محتواهای زیان‌آور شامل مجموعه اقدامات عملی است که با هدایت اسناد فوق‌الذکر روی می‌دهد.

ـ نحوهٔ یافتن محتوای زیان‌آور

در ابتدا همه کاربران واردشده به یوتیوب می‌توانند گزارش کنند یا نظر بدهند که یک ویدئو زیان‌آور است. این عمل «نشان‌گذاری محتوا» نامیده می‌شود، اما همه بینندگان (برای مثال) قادر به شناسایی نمادها، لوگوها یا زبان رمزگذاری‌شده‌ای نیستند که سازمان‌های تروریستی از آن استفاده می‌کنند. به‌همین‌دلیل آن‌ها از یک برنامه نشان‌گذار قابل اعتماد از محتوا استفاده می‌کنند که یوتیوب برای نهادهای دولتی و سازمان‌های مردم‌نهاد متخصص در حوزه‌های خاص ایجاد کرده است؛ بنابراین هر نوع محتوای نشان‌گذاری‌شده، لزوماً حذف نمی‌شود و به نشان‌گذار مورد اعتماد یوتیوب نیز ارجاع داده می‌شوند. در صورت مهم‌ترین ابزار شناسایی تیوتیوب یادگیری ماشین است.

نحوهٔ عملکرد یادگیری ماشین

یادگیری ماشین روشی برای این است تا به رایانه‌ها آموزش بدهیم ما دنبال چه چیزی

1. How YouTube Enforces Policies
2. Jennifer Flannery O'Connor
3. Matt Helprin

هستیم و این طریق نشان دادن نمونه‌های از آن چیز به ماشین است. ماشین براساس این نمونه‌ها الگوهایی ایجاد می‌کند که این الگوها موضوعات را برای آن‌ها تبیین می‌کنند و از این الگوها برای انجام پیش‌بینی برای یافتن نمونه‌های جدیدی استفاده می‌کنند که با آن الگو تطابق دارد.

یادگیری ماشین چگونه به یافتن محتوای زیان‌آور کمک می‌کند؟

سیستم‌های یادگیری ماشینی به‌طور دائم در حال رصد محتواهای به اشتراک گذاشته‌شده در یوتیوب هستند. یوتیوب از سال ۲۰۱۷ سرمایه‌گذاری بر یادگیری ماشین برای کشف محتوای زیان‌آور را آغاز کرده است. کاربرد این فناوری برای شناسایی محتوای زیان‌آور به‌طور خاص در حوزه‌های گفتار نفرت‌پراکنانه یا نفرت‌پراکنی کلامی و ایمنی کودکان آغاز شده است. امروزه بیش از نود درصد از محتوایی که از یوتیوب حذف می‌شود در ابتدا توسط سیستم یادگیری ماشین شناسایی شده است. همچنین بیشتر محتوایی که از یوتیوب حذف می‌شوند کمتر از ده بازدید (در مقایسه با ده هزار بازدید) داشته‌اند. به این معنا که آن‌ها به‌سرعت شناسایی و حذف شده‌اند. کاربرد این فناوری سبب کاهش فراوان انتشار محتواهای زیان‌آوری شده است که ناقض اصول راهنمای اجتماعی یوتیوب بوده‌اند و کاربران آن‌ها را مشاهده کرده‌اند.

به‌دلیل اینکه فناوری یادگیری ماشین کامل نیست و برخی تلاش می‌کنند تا این فناوری را دور بزنند و نیز اینکه یادگیری ماشین نمی‌تواند جایگزین داوری انسانی شود، به‌ویژه در موقعیت‌های پیچیده‌ای که زمینه اهمیت زیادی دارد، برای مثال گفتار نفرت‌پراکنانه در یک راهپیمایی و نیز گزارش خبری از آنکه رویداد واحدی را دربرمی‌گیرد. به‌دلیل همین محدودیت‌هاست که یوتیوب به گروهی از بازبینان متخصص انسانی تکیه می‌کند که با سیستم یادگیری ماشینی یوتیوب همکاری کنند تا محتواها را مرور و آن‌ها را نشان‌گذاری کنند.

برای محتوای نشان‌دار شده چه اتفاقی می‌افتد؟

سیستم یادگیری ماشینی برخی از محتواها را به‌سرعت حذف می‌کند. یکی از دلایل آن این است که این سیستم اطمینان زیادی دارد که آن نوع محتوا دستورالعمل اجتماعی را نقض کرده است و اینکه این سیستم محتوای مشابه همانند هرزنامه را پیشتر حذف کرده است، اما در موارد دیگر در صورت شناسایی محتوا توسط سیستم یا نشان‌گذاری

توسط کاربران و سیستم در حذف آن اطمینان کافی نداشته باشد، این وظیفه بازبینان آموزش دیده یوتیوب است که تصمیم بگیرند آن نوع محتوا ناقض اصول راهنمای اجتماعی یوتیوب هست یا خیر.

افراد مختلفی در ناحیه‌ای زمانی مختلف در سرتاسر جهان برای یوتیوب کار می‌کنند که به زبان‌های گوناگون تسلط دارند که می‌توانند به نشان‌گذاری‌های انجام‌شده توسط اعضای اجتماعی کاربران یوتیوب، نشان‌گذاران مورد اعتماد و سیستم یوتیوب در تمام مدت شبانه‌روز واکنش نشان دهند. آن‌ها محتواهایی را که آشکارا ناقض خط‌مشی یوتیوب باشند، حذف می‌کنند؛ برای مثال ویدئوهایی که در آن‌ها فعالیت‌های سازمان‌های تروریستی تبلیغ می‌شود.

اما اگر بازبینان یوتیوب تصمیم بگیرند که آن نوع محتوا از اصول راهنمای اجتماعی یوتیوب تخطی نکرده است، آن محتوا باقی می‌ماند.

موارد استثناء در خط‌مشی یوتیوب

برخی محتواها که در نگاه اول به نظر می‌رسد برخلاف خط‌مشی یوتیوب باشند، براساس تصمیم بازبین‌ها همچنان در یوتیوب باقی می‌مانند. به این محتواها ادسا[1] گفته می‌شود که حروف اول مطالب آموزشی، مستند، علمی یا هنری است. برای مثال، فیلم مستندی که در آن روش درمانی برای همجنس‌خواهی ارائه می‌شود، همچنان می‌تواند در یوتیوب باقی بماند یا محتواهایی که هدف آن آموزش نحوهٔ مقابله با قلدری مجازی باشد، هرچند این موارد را به تصویر بکشد، همچنان در یوتیوب می‌تواند نمایش داده شود. برای اینکه یک ویدئو جزو استثنائات ادسا هست یا خیر براساس عنوان، توضیحات و زمینه آن ویدئو تصمیم‌گیری می‌شود. این تصمیمات ظریف فقط توسط عوامل انسانی انجام می‌شود. هرچند سیستم‌ها نیز برای یادگیری آن‌ها آموزش داده می‌شوند. به‌علت پیچیدگی تصمیم‌گیری در این موارد عوام انسانی نیز به‌طور مرتب و عمیق با اصول راهنمای اجتماعی آشنا شوند.

براساس آمار ارائه‌شده در این ویدئو بیش از ۹۸ درصد از تولیدکنندگان محتوا هرگز مقررات یوتیوب را نقض نمی‌کنند. با این همه دربارهٔ اینکه در صورت نقض دستورالعمل‌های یوتیوب به چه شیوه‌ای عمل می‌شود، پیشتر اشاره شده است. به‌طورکلی اگر افراد در طول سه ماه سه بار هشدار دریافت کنند، کانال آن‌ها بسته

1. EDSA

می‌شود اما آمار نشان می‌دهد ۹۴ درصد افرادی که برای نخستین‌بار اخطار دریافت می‌کنند، دیگر مرتکب خطا نمی‌شوند؛ بنابراین مواردی که در کل منجر به مسدودسازی کانال در یوتیوب می‌شود، بسیار اندک است. با این همه موارد استثنایی هم وجود دارد. چنانچه کانال متعلق به یک سازمان تروریستی باشد یا در آن سوءاستفاده جنسی از کودکان به نمایش درآید، از این فرصت سه مرحله‌ای که یوتیوب معمولاً برای متخلفان در نظر گرفته است و کانال بلافاصله مسدود می‌شود.

برای اینکه نشان دهیم تیم بررسی محتوا چه وظیفه‌ای عظیمی دارد، باید اشاره شود که در ربع اول سال ۲۰۲۱ در حدود ۹٫۵ میلیون ویدئو، یک میلیارد کامنت و ۲٫۲ کانال از یوتیوب حذف شده است.

همچنین درصورتی‌که خالق یک ویدئو مدعی باشد که در حذف ویدئو اشتباهی صورت گرفته است، می‌تواند درخواست تجدیدنظر کند. در این صورت محتوای مربوطه مجدداً بررسی می‌شود و در صورت اثبات اشتباه آن ویدئو مجدداً بازگردانده می‌شود.

آنچه در این بخش گفته شد، روند کلی تعدیل محتوا است که فرایند آن شرح داده شد. تعدیل محتوا در یوتیوب به‌طورکلی تابع قواعدی است که در اصول راهنمای اجتماع یوتیوب ذکر شده است. مقرراتی که بر نحوهٔ انتشار محتوا در این پلتفرم حاکم است، تعیین می‌کند که تعدیل محتوا چگونه روی بدهد؛ اما اسناد دیگری هم وجود دارند که ناظر بر اداره و تنظیم مقررات در یوتیوب است که آن‌ها را نیز مرور خواهیم کرد.

به‌طورکلی براساس سند شرایط ارائه خدمت یوتیوب هر نوع استفاده کاربران از این پلتفرم یا خدمت مشمول چند نظام‌نامه یا سند است. این مقررات عبارت‌اند از: اصول راهنمای اجتماع یوتیوب[1]، خط‌مشی[2]، خط‌مشی‌های ایمنی و حق تألیف[3] که به شکل مستمر به‌روزرسانی می‌شوند و البته چنانچه کاربران آگهی منتشر کنند یا اینکه برای محتوای منتشرشده در یوتیوب از پشتیبان مالی استفاده نمایند، خط‌مشی‌های آگهی‌دهی[4] در یوتیوب نیز به آن‌ها اعمال می‌شود. علاوه‌بر این‌ها باید به خط‌مشی حریم خصوصی، هشدار حریم خصوصی بخش کودکان یوتیوب[5]، شرایط پردازش داده

1. YouTube Community Guidelines
2. Policy
3. Safety and Copyright Policies
4. Advertising on YouTube Policies
5. YouTube Kids Privacy Notice

در یوتیوب[1]، اصول راهنما و خط‌مشی‌های محتوای یوتیوب[2] نیز اشاره کرد.

ازآنجاکه هدف این پژوهش صرفاً مطالعه کردارهای تعدیل محتوا در پلتفرم‌ها است، دو موضوع برجسته می‌شود. نخست اسنادی که برای تعدیل محتوا بر آن‌ها تکیه می‌شود که مهم‌ترین آن‌ها اصول راهنمای اجتماعی یوتیوب است دوم نحوهٔ انجام تعدیل محتوا است؛ بنابراین برخی از موارد و اسناد فوق، گرچه بخشی از فرایند کلی حکمرانی بر یوتیوب را شامل می‌شوند، اما صرفاً به شکل اختصار بررسی می‌شوند.

قسمتی از بخش پشتیبانی و کمک گوگل به کمک به کاربران در مورد یوتیوب[3] اختصاص دارد. در این بخش ایمنی، خط‌مشی یا سیاست‌گذاری و حق تألیف قرار دارد. بخش عمدهٔ بحث‌های فرعی در قسمت خط‌مشی‌های یوتیوب شامل همان مباحثی است که در اصول راهنمای اجتماعی نیز دربارهٔ آن‌ها بحث شده است. آن‌ها عبارت‌اند از: هرزنامه و کردارهای فریب‌آمیز، محتوای حساس، محتوا خشونت‌آمیز یا خطرناک، کالاهای مشمول تنظیم مقررات و نیز اطلاعات نادرست.

یک بخش دیگر عبارت است از بهترین کردارها برای تولیدکنندگان و نیز خط‌مشی‌های قانونی. در این قسمت بخشی با عنوان «مسئولیت تولیدکننده»[4] وجود دارد. بر طبق نظر یوتیوب «تولیدکنندگان در قلب یوتیوب قرار دارند. بخشی از تولیدکنندگان بودن به این معنا است که افراد عضو یک اجتماع جهانی بانفوذ و بزرگ هستند»؛ بنابراین این افراد برای حفاظت و حراست از این اجتماع باید موافقت خود را با اصول راهنمای اجتماع و شرایط ارائه خدمت اعلام کنند.

حفاظت و حراست از اجتماع یوتیوب مورد تأکید یوتیوب است و این پلتفرم به کاربران خود خاطرنشان می‌کند:

به خاطر داشته باشید شما به‌عنوان خالقان (تولیدکنندگان) در یوتیوب با هم در و هم در خارج از آن مسئولانه رفتار کنید. اگر مشاهده کنیم که رفتار یک خالق در و خارج از پلتفرم به کاربران، اجتماع، کارکنان یا اکوسیستم ما آسیب می‌زند، اقداماتی را برای حفاظت از این اجتماع انجام خواهیم داد (YouTube, 2024).

گزارش کردن محتوای زیان‌آور توسط کاربران

همان‌طور که در قسمت پیشین اشاره شد، یکی از عناصر سازنده و اصلی تعدیل محتوا در

1. YouTube Processing Terms
2. YouTube's content Policies and Guidelines
3. You Tube Help
4. Creator Responsibility

یوتیوب اعضای اجتماع کاربران هستند. بخشی از شروع فرایند تعدیل محتوا با گزارش‌هایی آغاز می‌شود که توسط کاربران به اطلاع مسئولان نظارت بر محتوا در یوتیوب رسانده می‌شود؛ بنابراین ضرورت دارد این بخش با جزئیات بیشتری واکاوی شود.

گزارش کردن و اجرا بخشی از خط‌مشی یوتیوب برای اجرای تعدیل محتوا است (YouTube, 2024a). یوتیوب به‌طورکلی برای گزارش کردن و نشان‌دار کردن محتوایی که به نظر آن‌ها نامناسب است بر اعضای اجتماع کاربران خود تکیه می‌کند. گزارش محتوا به شکل ناشناس صورت می‌گیرد و مشخص نمی‌شود که کدام کاربر محتوای را گزارش کرده است. گزارش یک محتوا به معنای حذف خودکار آن محتوا نیست؛ بلکه وارد فرایند بازبینی می‌شود. در این مرحله محتواهایی که برخلاف اصول راهنمای اجتماعی یوتیوب باشند از یوتیوب حذف می‌شوند، اما محتواهایی که ممکن است برای مخاطبان نوجوان و زیر سن قانونی یا به اصطلاح صغیر نامناسب باشند، برچسب محدودیت سنی[1] می‌خورند[2].

نشان‌دار کردن یا پرچم زدن یک محتوا دارای مراحل و شرایط و عناصر گوناگونی است. گزارش محتوا می‌تواند شامل گزارش یک ویدئو، گزارش فهرست پخش، گزارش تصویر ویدئویی کوچک، گزارش یک لینک، گزارش یک کامنت، گزارش پیام مبادله‌شده در یک چت همزمان و حتی گزارش یک آگهی تبلیغاتی باشد. در کنار هر محتوای منتشرشده در یوتیوب گزینه‌ای برای گزارش‌دهی وجود دارد. عموماً دلیل اینکه از نظر کاربر محتوای یک ویدئو ناقض اصول راهنمای اجتماعی است از میان دلایل موجود باید بیان شود و در صورت لزوم دلایل اضافی نیز که به گروه بازبینی کمک می‌کند نیز می‌تواند ارائه شود.

علاوه‌بر موارد فوق، گزینه‌های دیگری نیز برای انجام گزارش وجود دارد. یکی از آن‌ها گزارش یک کانال است. این زمانی است که ضرورت دارد، بیش از یک بخش از محتوا گزارش شود. گزارش می‌تواند شامل گزارش نقض حریم خصوصی و گزارش موارد قانونی نیز باشد.

---

1. age-restricted

۲. گاهی اقوات محتواهایی که از خط‌مشی یوتیوب تخطی نمی‌کنند، اما برای افراد کمتر از هجده سال مناسب نیستند، در ردهٔ محتوای دارای محدودیت سنی قرار می‌گیرند. این خط‌مشی علاوه‌بر ویدئوها به شرح ویدئوها، تصاویر، پخش برخط ویدئو و سایر محصولات و ویژگی‌های یوتیوب اعمال می‌شود. محتواهایی که ایمنی کودکان را مخاطره می‌اندازند. ازجمله ویدئویی که شامل شرکت کودکان در فعالیت‌های خطرناکی است و کودکان به راحتی می‌توانند آن‌ها را تقلید کنند، محتواهای زیان‌آور و خطرناک، محتواهای دارای تصاویر برهنه و جنسی، محتوای خشونت‌آمیز و ویدئوهایی که زبان اهانت‌آمیز دارند در این رده قرار می‌گیرند. هرکدام از این موارد شامل ویژگی‌ها و شروط مختلفی است که سبب می‌شود تا در این دسته قرار گیرند.

برنامهٔ نشان‌دارکنندهٔ مورد اعتماد یوتیوب[1] را یوتیوب برای کمک به ارائه ابزارهای نیرومند به افراد، نهادهای دولتی و سازمان‌های غیردولتی توسعه داده است. این برنامه امکان گزارش کردن چندین ویدئو در آن واحد را فراهم می‌کند. تصمیم‌های گرفته‌شده در محتوای نشان‌دارشده را شفاف می‌کند و زمینه ادامه بحث درباره محتواهای گوناگون یوتیوب را فراهم می‌کند.

سینگ (۲۰۱۹) یک سال پس از انتشار اصول سانتاکلارا میزان توجه به اصول پیشنهادشده را دربارهٔ پلتفرم‌های یوتیوب، فیس‌بوک و توئیتر بررسی کرده است. در آوریل ۲۰۱۸ گوگل ازطریق یوتیوب نخست گزارش شفافیت جامع خود درباره تعدیل محتوا را منتشر کرد که اطلاعات سودمندی را دربارهٔ حذف محتوا در این پلتفرم براساس تخطی از اصول راهنمای اجتماعی آن منتشر کرد که اولین باری است که یک پلتفرم اینترنتی داده‌های مرتبط با کردارهای تعدیل محتوای خود را منتشر کرده بود.

هم اکنون گزارشی باعنوان اجرای اصول راهنمای اجتماعی یوتیوب در سایت گوگل در دسترس قرار دارد که آخرین دورهٔ آن به ژوئیه تا سپتامبر ۲۰۲۱ تعلق دارد (Google, 2021). اولین گزارش مربوط به ژوئیه تا سپتامبر ۲۰۱۸ اختصاص داشته است. این گزارش‌ها را می‌توان در راستای شفاف‌سازی مدنظر اصول سانتاکلارا دربارهٔ نحوه تعدیل محتواهای کابرساخته قلمداد کرد[2].

براساس این گزارش یوتیوب هم به افراد و هم فناوری برای نشان‌دار کردن محتوای نامتناسب و اعمال اصول راهنما تکیه می‌کند. نشان‌دار می‌تواند به‌وسیلهٔ سیستم‌های نشان‌دار کردن خودکار، اعضای برنامه نشان‌دارکنندگان مورد اعتماد (سازمان‌های مردم‌نهاد، نهادهای حکومتی و افراد) یا توسط اعضای عادی اجتماعی کاربران یوتیوب اتفاق بیفتد. در این گزارش خلاصه‌ای از این فرایند در یوتیوب ذکر شده است.

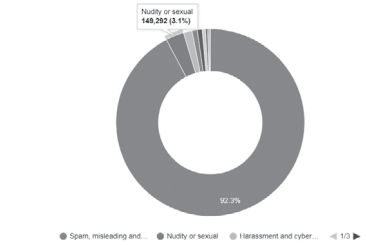
کانال‌های حذف‌شده

این نمودار تعداد کانال‌هایی را که یوتیوب براساس دلیل حذف کرده است، نشان می‌دهد. بخش عمدهٔ دلیل خاتمه فعالیت کانال، ناشی از اختصاص کانال به هرزنامه (اسپم) یا دارا بودن محتوای جنسی مربوط به بزرگسالان بوده است.

براساس این گزارش در فاصله زمانی ژوئیه تا سپتامبر ۲۰۲۱ در مجموع بیش از چهار میلیون و هشتصد هزار کانال از یوتیوب حذف شده است. پس از حذف یک کانال تمام ویدئوهای آن نیز حذف می‌شوند که به این ترتیب بیش از ۷۵ میلیون و ۲۵۰ هزار ویدئو تعلیق شده‌اند.

1. You Tube Trusted Flagger Program

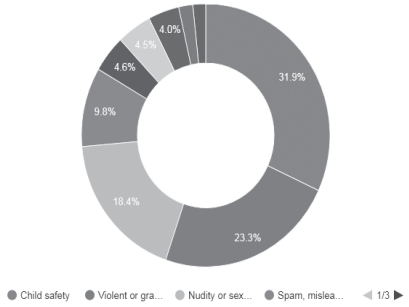
۲. اکنون گزارش آوریل تا ژوئن ۲۰۲۴ در دسترس قرار دارد.



شکل ۲. میزان ویدئوهای حذف‌شده از یوتیوب

همان‌طور که در شکل ۲ مشخص است، بیش از ۹۲ درصد از ویدئوهای حذف‌شده (در حدود ۴ میلیون ۴۳۰ ویدئو) شامل اسپم، مطالب گمراه‌کننده و اسکم بوده است. بیش از ۳/۱ درصد در حدود ۱۴۹ هزار ویدئو مرتبط با برهنگی و تصاویر جنسی، در حدود ۱/۵ درصد یا ۷۰ هزار آزار و قلدری سایبری و حدود ۹/. درصد نیز شامل محتوای نفرت پراکنانه یا مبتنی بر سوءاستفاده بوده است.

یوتیوب برای بازبینی ویدئوهای نشان‌گذاری‌شده و حذف محتواهای ناقض اصول راهنما اجتماع از تیم‌هایی در سرتاسر جهان استفاده می‌کند. در فاصله زمانی ژوئیه ۲۰۲۱ تا سپتامبر ۲۰۲۱ تعداد ۶۲۲۹۸۸۲ حذف‌شده که شامل نشان‌دار کردن خودکار نیز هست. این تعداد در یک طول یک فصل از یوتیوب حذف شده است.



شکل ۳. دلایل حذف ویدئو در یوتیوب

شکل ۳ نشانگر حجم و ویدئوهای حذف‌شده، براساس دلیل حذف ویدئو است. حذف بر مبنای اصول راهنمای اجتماعی انجام می‌شود. بازبین‌ها ویدئوهای نشان‌دارشده را بر مبنای همه اصول راهنمای اجتماعی و خط‌مشی‌های یوتیوب ارزیابی می‌کنند و دلیل اولیه علت نشان‌دار شدن آن ویدئو نقشی در این مرحله از تصمیم‌گیری ندارد. همان‌طور که در این شکل مشخص است، بیش از یک میلیون و ۹۸۶ هزار ویدئو به‌دلیل امنیت کودکان

حذف شدند. محتوای خشونت‌آمیز (۳/۲۳ درصد) و نیز محتوای جنسی (بیش از ۱۸ درصد) در رده‌های بعدی علت حذف محتوا توسط بازبین‌ها قرار دارد.

سوزان وجیستکی[1] در توضیح نحوهٔ فعالیت یوتیوب برای حذف محتواهای زیان‌آور گفته است که آن‌ها برای حفاظت از آزادی بیان برخلاف فیس‌بوک به‌جای مسدود کردن فعالیت افراد یا سازمان‌هایی که به‌دلیل انتشار محتوای زیان‌آور شناخته‌شده هستند، محتواهایی را که از استانداردهای این پلتفرم تخطی کنند، حذف می‌کند. آن‌ها به‌جای تمرکز بر اشخاص بر آنچه گفته می‌شود، تمرکز می‌کنند (Adams and Huffs, 2021).

چالش‌های تعدیل محتوا در یوتیوب

اخیراً افرادی که وظیفهٔ بازبینی محتوا را در پلتفرم‌های مختلف ازجمله یوتیوب برعهده دارند، اقدام به طرح شکایت علیه این شرکت‌ها به‌دلیل آسیب‌های روانی ناشی از مواجه مداوم با تصاویر خشونت‌آمیز و غیراخلاقی کرده‌اند. آن‌ها این شرکت‌ها را متهم کرده‌اند که به‌اندازهٔ کافی از متولیان تعدیل محتوا حمایت نمی‌کنند. برای مثال، اخیراً یکی از تعدیل‌گران سابق یوتیوب شکایتی را علیه یوتیوب مطرح کرده است. وی معتقد است که این شرکت از افرادی که موظف هستند، ویدئوهای خشونت‌آمیز را ردیابی و حذف کنند، حفاظت نمی‌کند (Elias, 2020).

موضوع دیگر به چالش‌های کاربرد یادگیری ماشینی برای تعدیل محتوا برمی‌گردد. مشخص شده است که سیستم یادگیری ماشینی همچنان نیازمند یادگیری بیشتر است و گاهی اتکاء بیش از حد به آن مشکلاتی را ایجاد می‌کند. برای نمونه، یوتیوب برای مقابله با انتشار اطلاعات نادرست دربارهٔ ویروس کرونا در این رسانه اجتماعی به ماشین‌ها تکیه کرده بود، اما بررسی‌ها نشان داده است که یادگیری ماشینی بیش از حد نسبت به سانسور محتواهایی که روی مرز قرار داشته‌اند، اقدام کرده است؛ بنابراین یوتیوب مجبور شده است برای مقابله با محتوای زیان‌آور از بازبینان انسانی بیشتری استفاده کند. پس از توقف فعالیت حدود ده هزار تیم قدرتمند پالایش محتوا توسط یوتیوب به‌دلیل بیماری همه‌گیر، یوتیوب به سیستم‌های یادگیری ماشینی خود خودمختاری بیشتری برای جلوگیری از تماشای محتواهای گفتار نفرت‌پراکنانه، خشن و یا شکل‌های دیگر محتوای زیان‌آور یا اطلاعات نادرست داد، اما یکی از نتایج کاهش نظارت انسانی افزایش تعداد ویدئوهای حذف‌شده بود که بخش زیادی از آن‌ها فاقد

1. Susan Wojcick

نقض مقررات بودند. در این مدت تعداد ویدئوهای حذف‌شده نسبت به زمان معمول در حدود دو برابر افزایش پیدا کرده بود (Barker and Murphy, 2020).

با وجود این براساس مطالعه‌ای که کنوتیلا و همکارانش (۲۰۲۰) در مؤسسه اینترنت آکسفورد دربارهٔ انتشار ویدئوهای حاوی اطلاعات نادرست در رسانه‌های اجتماعی و میزان اثرگذاری سیاست‌های پلتفرم‌ها انجام داده‌اند، حذف ویدئوهای حاوی اطلاعات نادرست ۴۱ روز زمان برده بوده است. همچنین این دسته از ویدئوها پیش‌از آنکه در یوتیوب دیده شوند، عمدتاً در فیس‌بوک به اشتراک گذاشته شده بودند. علاوه‌بر این، ویدئوهای حاوی اطلاعات نادرست بیش از پنج منبع اصلی انگلیسی زبان شامل سی.ان.ان. ای.بی.سی. نیوز، بی.بی.سی، فاکس نیوز و الجزیره به اشتراک گذاشته شده بودند. این موضوع نشانگر وجود برخی شکاف‌ها در سیاست‌های تعدیل محتوای در یوتیوب است که نقش زیادی در انتشار اطلاعات نادرست مرتبط با کرونا در این پلتفرم داشته است.

نتیجه‌گیری

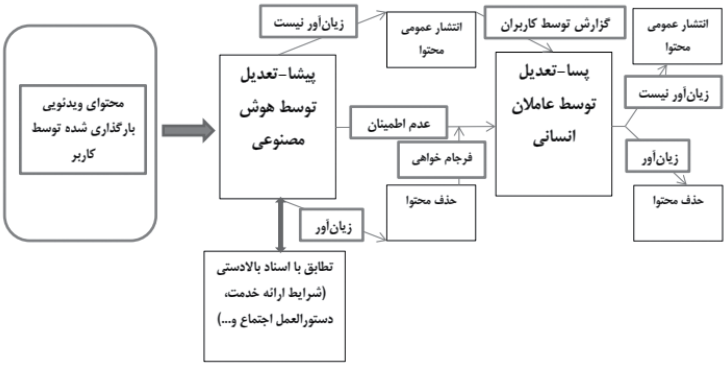
تعدیل محتوا به بخشی ضروری از پلتفرم‌های دیجیتال تبدیل شده است و از کاربران در برابر محتوای مضر محافظت می‌کند و درعین‌حال یک محیط آنلاین فراگیر و ایمن را تضمین می‌کند. چالش مدیریت حجم وسیعی از محتوای کاربرساخته به توسعه سازوکارهای پیشرفته تعدیل محتوا منجر شده است که هم هوش مصنوعی و هم تلاش‌های تعدیل انسانی را در بر می‌گیرد. سیستم‌های تعدیل مبتنی‌بر هوش مصنوعی کارایی استثنایی در شناسایی و پرچم‌گذاری محتوای مشکل‌ساز در مقیاس نشان داده‌اند. بااین‌حال، آن‌ها بدون محدودیت نیستند؛ زیرا اغلب با درک زمینه و تفاوت‌های ظریف در زبان مشکل دارند که گاهی به علامت‌گذاری اشتباه و حذف محتوای قانونی منجر می‌شود.

به‌طورکلی تعدیل محتوا دارای ابعاد نظری و عملی است. درابتدا لازم است تا پلتفرم‌ها براساس موارد گوناگون، شامل اسناد بالادستی، الزامات قانونی و رویکرد درون‌پلتفرمی مجموعه‌ای از اسناد شامل شرایط ارائه خدمت، اصول راهنمای اجتماع، سیاست حریم خصوصی، ایمنی، اطلاعات نادرست و گمراه‌کننده و غیره را توسعه دهند و برخی رویه‌های عملی را در پلتفرم مستقر کنند تا عمل بازبینی و پالایش محتواهای زیان‌آور از غیرزیان‌آور به روشنی و بدون ابهام انجام شود.

تعدیل‌کنندگان انسانی با واردکردن قضاوت و تشخیص خود به فرایند تعدیل، نقش

مهمی در رسیدگی به این محدودیت‌ها ایفا می‌کنند. آن‌ها برای درک پیچیدگی‌های زبان، فرهنگ و زمینه، مجهزتر هستند و ارزیابی دقیق‌تری از محتوا را تضمین می‌کنند.

مطابق شکل ۱ در توضیح فرایند تعدیل محتوا توسط هوش مصنوعی در پلتفرم یوتیوب باید گفت که ابتدا محتوای بارگذاری‌شده توسط کاربر که می‌تواند انواع محتوای ویدئویی ازقبیل گیف، چت زنده، ویدئوی زنده، میم و دیپ‌فیک یا جعل عمیق باشد، وارد مرحله پیشاتعدیل توسط هوش مصنوعی می‌شود. اگر در این مرحله محتوای بارگذاری‌شده زیان‌آور یا غیرمجاز تشخیص داده نشود در معرض دید عموم قرار می‌گیرد. این ویدئو ممکن است از طرف برخی کاربران زیان‌آور قلمداد می‌شود به اصطلاح آن را گزارش کنند. در این حالت محتوای مورد نظر وارد مرحلهٔ پساتعدیل یا تعدیل واکنشی می‌شود که توسط عاملان انسانی صورت می‌گیرد. اگر تعدیل‌کنندگان انسانی آن محتوا را زیان‌آور شناسایی کنند، محتوا حذف می‌شود و در غیر این صورت همچنان در معرض دید عموم کاربران قرار می‌گیرد. در وضعیت دوم که هوش مصنوعی از همان ابتدا محتوا را زیان‌آور تشخیص دهد، محتوا حذف می‌شود، اما ممکن است در این مرحله کاربری که آن‌ها محتوا را به اشتراک گذاشته است، اعتراض کند و به‌اصطلاح تقاضای فرجام‌خواهی کند؛ در این حالت، آن محتوا بار دیگر توسط عاملان انسانی بازبینی می‌شود و در این مرحله آن‌ها تصمیم نهایی را مبنی‌بر انتشار یا عدم انتشار محتوا خواهند گرفت. این بدان خاطر است که هوش مصنوعی با وجود پیشرفت‌های اخیر همچنان در حال یادگیری است؛ بنابراین ممکن است در تشخیص زیان‌آور یا زیان‌آور نبودن یک محتوا دچار تردید شود، در این حالت این محتوا به‌طور مستقیم به عاملان انسانی ارجاع داده می‌شود تا آن‌ها براساس دانش زمینه‌ای در مورد سرنوشت آن محتوا تصمیم بگیرند.



شکل ۱. فرایند سازوکار تعدیل محتوا در پلتفرم یوتیوب

از آنجایی که پلتفرم‌های دیجیتال ایرانی به تکامل خود ادامه می‌دهند، یادگیری از بهترین شیوه‌های جهانی در تعدیل محتوا، اتخاذ یک رویکرد متعادل که به‌طور مؤثر تلاش‌های مبتنی‌بر هوش مصنوعی و تعدیل انسانی را ترکیب می‌کند، برای آن‌ها ضروری است. تعدیل محتوا یک سازوکار پیچیده دارد و شامل طیفی از دستورالعمل‌ها تا رویه‌های عملی پالایش محتوا است. ازآنجاکه محتواهای نابهنجار جدید به شکل فزایندهٔ تولید و به اشتراک گذاشته می‌شوند، لازم است که این دستورالعمل‌ها به‌طور مکرر به‌روز شوند و با تا حد امکان شفاف و دقیق باشند تا عمل تعدیل هم توسط عوامل انسانی و هم توسط عوامل ماشینی با سهولت و بدون ابهام انجام شود.

برای دستیابی به این هدف، پلتفرم‌های ایرانی باید روی فناوری پیشرفته هوش مصنوعی سرمایه‌گذاری کنند و درعین‌حال محیطی همدلانه و حمایت‌کننده را برای تعدیل‌کنندگان انسانی ایجاد کنند. علاوه‌بر این، پلتفرم‌های ایرانی باید شفافیت، مسئولیت‌پذیری و همکاری با کاربران، ذی‌نفعان صنعت و مقامات نظارتی را ارتقا دهند تا اطمینان حاصل شود که شیوه‌های تعدیل محتوا با ارزش‌های اجتماعی و الزامات قانونی مطابقت دارد. این اقدام درنهایت باعث ایجاد یک محیط آنلاین امن و فراگیر برای کاربران ایرانی می‌شود و درعین‌حال خطرات مرتبط با محتوای کاربرساخته، به‌ویژه در مورد محتواهای ویدئویی را کاهش می‌دهد.

تعارض منافع

تعارض منافع ندارم.

Orcid

Hossein Hassani https://orcid.org/0000-0003-1255-9533

منابع و مآخذ

اخوان، منصوره، روشندل اربطانی، طاهر، خواجه ئیان، داتیس و عقیلی، سید وحید (۱۴۰۲). طراحی چارچوب حکمرانی صوت‌وتصویر فراگیر در جمهوری اسلامی ایران. *فصلنامه علمی رسانه‌های دیداری و شنیداری*، ۱۷ (۴۷)، ۳۷ـ۶۴. Doi: 10.22085/javm.2023.402680.2083

اخوان، منصوره، روشندل اربطانی، طاهر، عقیلی، سید وحید و خواجه ئیان، داتیس (۱۴۰۲). مطالعه تطبیقی الگوی حکمرانی صوت‌وتصویر فراگیر در روسیه، ترکیه و کره جنوبی. *مدیریت دولتی*، ۱۵ (۲)، ۲۵۸ـ۲۹۲.

Doi: 10.22059/jipa.2023.351991.3251

اکبری نوری، ح (۱۳۹۹). تنظیم‌گری پلتفرم‌های برخط رسانه‌ای صوتی و تصویری در ایران (گزارش نخست)، ارائه چارچوبی هنجاری برای تنظیم‌گری پلتفرم‌های رسانه‌ای صوتی و تصویری. تهران: *ساترا*.

https://satra.ir/fa/wp-content/uploads/2019/12/02-plat-majazi.pdf

حسنی، حسین (۱۳۹۹). کرونا، فاصله‌گذاری اجتماعی و فرهنگ ویدیویی پلتفرمی در کرونا و جامعه ایران: سویه‌های فرهنگی‌ـ اجتماعی (مجموعه مقالات). تهران: پژوهشگاه فرهنگ، هنر و ارتباطات.

حسنی، حسین، کلانتری، عبدالحسین (۱۳۹۹). سیاست‌گذاری اینترنت: مرور سیستماتیک رویکردهای حکمرانی پلتفرم‌های آنلاین و رسانه‌های اجتماعی. *سیاستگذاری عمومی*، ۶(۳)، ۵۹ـ۷۹.

Doi: 10.22059/jppolicy.2021.79492

خرم‌دل مهدی، استوارسنگری کوروش، علائی حسین، ضرابی حمید (۱۴۰۱). چالش‌های مربوط به مرجع تنظیم‌گر و ناظر بر تولیدات صوت‌وتصویر فراگیر درفضای مجازی. *حقوق اداری*، ۹(۳۰)، ۳۱ـ۵۶.

سرحدی‌کاظم، طاهری، محسن (۱۳۹۹). جستاری در اعمال نظارت مطلوب بر انتشار صوت‌وتصویر در فضای مجازی از منظر حقوقی. *حقوق اداری*، ۸(۲۵)، ۱۳۹ـ ۱۶۰.

طحان نظیف، هادی، علی پور، محمدرضا (۱۴۰۱). جایگاه و آثار حقوقی خودتنظیم‌گری پلتفرم‌های دیجیتال. *حقوق فناوری‌های نوین*، ۳(۶)، ۱۲۷ـ۱۴۱. Doi: 10.22133/mtlj.2022.366647.1131

فن‌دایک، یوزه (۱۳۹۶). فرهنگ اتصال: تاریخ انتقادی رسانه‌های اجتماعی. ترجمهٔ حسین حسنی. تهران: سوره مهر.

قاسم‌زاده عراقی، مرتضی، خجسته بافرزاده، حسن، سلمانی‌شاه محمدی، عبدالرضا (۱۴۰۲). تبیین مفهومی خدمات رسانه‌ای صوت‌وتصویر فراگیر در ایران (با تأکید بر تجربه اتحادیه اروپا). *فصلنامه علمی رسانه‌های دیداری و شنیداری*، ۱۷ (۴۷)، ۱۶۱ـ۱۹۴. Doi: 10.22085/javm.2022.292068.1808

Bradford, A. (2023). Digital Empires: The Global Battle to Regulate Technology. Oxford. Oxford University Press.

Brubaker, Rogers. (2023). Hyperconnectivity and Its Contents, Cambridge: Polity.

Ceci, Laura (2024) Most popular video content type worldwide in 3rd quarter 2023, by weekly usage

reach, https://www.statista.com/statistics/1254810/top-video-content-type-by-global-reach/

Crowe, S., Cresswell, K., Robertson, A., Huby, G., Avery, A., & Sheikh, A. (2011). The case study approach. BMC medical research methodology, 11, 100. https://doi.org/10.1186/1471-2288-11-100

DeNardis, L. & Hackl, A.M. (2015). Internet Governance by Social Media Platforms. In The Oxford Handbook of Internet Studies. Edited by William H. Dutton. Oxford: Oxford University Press.

Dixon, Stacy Jo. (2024). Most popular social networks worldwide as of January 2024, ranked by number of monthly active users, https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

Flew, T. (2021). *Regulating Platforms*. Cambridge & Medford: Polity.

Gillespie, T. (2018). in The Sage Handbook of Social Media. Edited by Jean Burgess; Thomas Poell & Alice E. Marwick. London: Sage Publication.

Google. (2021). Google Transparency Report. Google.com. Avalable at: https://transparencyreport.google.com/youtube-policy/removals

Gorwa, R. (2019). What is platform governance?*Information, Communication & Society*,*22*(6), 854-871. https://doi.org/10.1080/1369118X.2019.1573914

Gorwa, R. (2024). The Politics of Platform Regulation: How Governments Shape Online Content Moderation. Oxford: Oxford University Press.

Gosztonyi, G. (2023). Censorship From Plato to Social Media: The Complexity of Social Media''s Content Regulation and Moderation Practices. Switzerland. Springer.

Roberts, S.T. (2017). Content Moderation. In: Schintler, L., McNeely, C. (eds) Encyclopedia of Big Data. Springer, Cham. https://doi.org/10.1007/978-3-319-32001-4_44-1

Akbari Nouri, H2021) .). Regulation of Online Audiovisual Media Platforms in Iran: A First Report. Presenting a Normative Framework for the Regulation of Audiovisual Media Platforms. Tehran. https://satra.ir/fa/wp-content/uploads/2019/12/02-plat-majazi.pdf **[in Persian]**

Akhavan, M., Roshandel Arbatani, T., Aqili, S. V., & Khajeheian, D. (2023). A Comparative Study of Audiovisual Governance Frameworks in Russia, Turkey and South Korea. *Journal of Public Administration*, 15(2), 258-292. doi: 10.22059/jipa.2023.351991.3251 **[in Persian]**

Akhavan, M., Roshandel Arbatani, T., Khajeheian, D., & Aqili, S. V. (2023). Designing a Framework of Immersive Audiovisual Governance in the Islamic Republic of Iran. *Quarterly Scientific Journal of Audio-Visual Media*, 17(47), 37-64. doi: 10.22085/javm.2023.402680.2083 **[in Persian]**

Ghasemzadeh, M., khojastehBagherzadeh, H., & Salmani Shah mohammadi, A. (2023). Conceptual Explanation of Immersive Audiovisual Media Services in Iran (Emphasizing the Experience of the European Union). *Quarterly Scientific Journal of Audio-Visual Media*, 17(47), 161-194. doi: 10.22085/javm.2022.292068.1808 **[in Persian]**

Hasani, H. (2021). Corona, Social Distancing, and Platform Video Culture. In *Corona and Iranian Society: Socio-cultural Dimensions*. Tehran: Research Institute for Culture, Arts, and Communications **[in Persian]**

Hassani, H., & Kalantari, A. (2020). Internet Policy: A Systematic Review of Approaches to Governance of Online and Social Media Platforms. *Iranian Journal of Public Policy*, 6(3), 59-79. doi: 10.22059/jppolicy.2021.79492 **[in Persian]**

Khoramdel M, Ostovarsangari K, Alayee H, Zarabi H. (2022). Challenges Related to the Reference Regulator and Supervisor of Pervasive Audio and Video Productions in Cyberspace. *Journal of Administrative Law*; 9 (30) :31-56 **[in Persian]**

Sarhaddi, K., Taheri, M. (2021). Investigating the Proper Monitoring of Audio and Video Broadcasts in Cyberspace from a Legal Perspective. *Journal of Administrative Law*; 8 (25) :139-160 **[in Persian]**

Singh, S. (2019). Assessing YouTube, Facebook and Twitter's Content Takedown Policies: How Internet Platforms Have Adopted the 2018 Santa Clara Principles. https://www.newamerica.org/oti/reports/assessing-youtube-facebook-and-twitters-content-takedown-policies/

Tan, C. (2018). Regulating content on social media: Copyright, terms of service, and technological features. London, United Kingdom: UCL Press.

Van Dijck, J. (2017). *The Culture of Connectivity: A Critical History of Social Me*dia (Translated by H. Hosseini). Tehran: Sooreh Mehr. **[in Persian]**

YouTube. (2021). Community Guidelines. YouTube. Available at: https://www.youtube.com/howyoutubeworks/policies/community-guidelines/#community-guidelines

YouTube. (2021). Terms of Service.YouTube. available at: https://www.youtube.com/t/terms?archive=20210317

YouTube. (2024). Overview of the Guidelines. YouTube. Available at: https://www.youtube.com/howyoutubeworks/policies/overview/

YouTube. (2024a). YouTube Help. YouTube.com. Available at: https://support.google.com/youtube/answer/7650329?hl=en&ref_topic=9282435