Jurisprudence and Legal Studies of Media Journal
VOL. 8 / Autumn 2026 and Winter 2027 / NO. 2 (Serial 16)

User Content Moderation: From Concept and Rationale to Regulatory Frameworks and Legal Challenges

Zeinab Tavassoli[1] 0009-0004-6120-109X
Farshid Moghimiān Boroujeni[2] 0009-0000-9730-3419

Abstract

User content moderation is one of the central concepts in the governance of cyberspace and plays a decisive role in regulating the relationship between online platforms, users, and public authorities. This study aims to identify and analyze the most significant legal challenges arising from current models of user content moderation and to examine their impact on users' fundamental rights, particularly freedom of expression. The research adopts a descriptive–analytical method based on documentary analysis, with data collected through the review of legal sources, platforms' internal policies, and selected comparative legal systems. The findings indicate that prevailing content moderation models—largely grounded in platform self-regulation or limited public intervention—face multiple challenges, including violations of privacy, a lack of transparency in algorithmic decision-making, ambiguity in defining and determining permissible content, and difficulties in applying legal rules within a transnational digital environment. In practice, these challenges may lead to restrictions on freedom of expression, discriminatory decision-making, and reduced opportunities for effective user appeals. The study concludes that the absence of clear legal standards and robust accountability mechanisms risks transforming content moderation from a tool for safeguarding public interests into a potential threat to users' rights. Accordingly, the research emphasizes the need to design a balanced regulatory framework for content moderation that, while addressing public order concerns, is grounded in the principles of transparency, proportionality, accountability, and the protection of users' rights.

Keywords: Content moderation; Digital Services Act; Artificial intelligence; Online platforms.

1- MA. in Public Law, University of Science and Culture, Tehran, Iran
Z.tavassoli.379@gmail.com
2- Ph.D. in Public Law, Shahid Beheshti University, Tehran
f.moghimian.b@gmail.com

دوفصلنامه
مطالعات فقه و حقوق رسانه
دانشکده غیردولتی – غیرانتفاعی رفاه

دوره ۸، شماره ۲ (پیاپی ۱۶)
پاییز و زمستان ۱۴۰۵
صص: ۳۳۱-۳۵۶

تعدیل محتوای کاربران؛ از چیستی و چرایی تا چگونگی تنظیم‌گری و چالش‌های حقوقی آن

نوع مقاله: ترویجی
DOI: 10.22034/jlsm.2026.2064992.1019
تاریخ دریافت: ۱۴۰۴/۰۴/۱۶ **تاریخ بازنگری:** ۱۴۰۵/۰۲/۰۸ **تاریخ پذیرش:** ۱۴۰۵/۰۴/۲۰

زینب توسلی[۱] **فرشید مقیمیان بروجنی**[۲]

چکیده

تعدیل محتوای کاربران یکی از مفاهیم محوری در حکمرانی فضای مجازی است و نقش تعیین‌کننده‌ای در تنظیم رابطه میان سکو(پلتفرم)ها، کاربران و نهادهای عمومی ایفا می‌کند. این پژوهش با هدف شناسایی و تحلیل مهم‌ترین چالش‌های حقوقی ناشی از الگوهای فعلی تعدیل محتوای کاربران و بررسی تأثیر این چالش‌ها بر حقوق بنیادین کاربران، به‌ویژه آزادی بیان، انجام شده است. روش پژوهش توصیفی–تحلیلی و مبتنی بر مطالعات اسنادی است و داده‌ها از طریق بررسی منابع حقوقی، ضوابط داخلی سکوها و نمونه‌هایی از نظام‌های حقوقی منتخب گردآوری شده‌اند. تحلیل داده‌ها نشان می‌دهد که الگوهای غالب تعدیل محتوا که عمدتاً بر خودتنظیم‌گری سکوها یا مداخله محدود نهادهای عمومی استوارند، با چالش‌هایی نظیر نقض حریم خصوصی، کمبود شفافیت در تصمیمات مبتنی بر الگوریتم، ابهام در تعریف و تعیین محتوای مجاز و دشواری اعمال قواعد حقوقی در بستر فراملی مواجه‌اند. این چالش‌ها در عمل می‌توانند به محدودسازی آزادی بیان، تصمیم‌گیری‌های تبعیض‌آمیز و کاهش امکان اعتراض مؤثر کاربران منجر شوند. یافته‌های پژوهش حاکی از آن است که فقدان معیارهای روشن حقوقی و سازوکارهای پاسخگویی، تعدیل محتوا را از ابزاری برای حمایت از منافع عمومی به عاملی بالقوه تهدیدکننده حقوق کاربران تبدیل می‌کند. درنتیجه، پژوهش بر ضرورت طراحی چهارچوبی متوازن برای تنظیم‌گری تعدیل محتوا تأکید دارد که ضمن پاسخ‌گویی به ملاحظات نظم عمومی، بر اصول شفافیت، تناسب، مسئولیت‌پذیری و تضمین حقوق کاربران استوار باشد.

واژگان کلیدی:

تعدیل محتوا، قانون خدمات دیجیتال، هوش مصنوعی، فضای مجازی، تنظیم‌گری سکوها.

---

۱. دانش‌آموخته کارشناسی ارشد حقوق عمومی دانشگاه علم و فرهنگ، تهران، ایران. (نویسنده مسئول)
Z.tavassoli.379@gmail.com
۲. دانش‌آموخته دکتری حقوق عمومی، دانشگاه شهید بهشتی، تهران، ایران f.moghimian.b@gmail.com

مقدمه

توسعه سریع فناوری‌های دیجیتال و گسترش اینترنت، نحوه تولید، انتشار و مصرف اطلاعات را به‌طور بنیادین تغییر داده و اکوسیستم جدیدی از سکو(پلتفرم)های دیجیتال ایجاد کرده است که محیط اصلی تعاملات و ارتباطات برای میلیاردها نفر در سراسر جهان محسوب می‌شوند. شبکه‌های اجتماعی، سکوهای میزبانی ویدئو، بازارهای برخط، وبلاگ‌ها و سایر خدمات مبتنی بر محتوای تولیدشده توسط کاربران، فرصت‌های بی‌سابقه‌ای برای آزادی بیان، تبادل‌نظر، تعامل اجتماعی و فعالیت اقتصادی فراهم آورده‌اند.

آزادی ایجاد و انتشار محتوا همراه با فرصت‌ها، مخاطرات جدی نیز به همراه دارد. محتوای کاربران می‌تواند حقوق و منافع مشروع دیگران را تهدید کند و شامل مواردی همچون افترا، محتوای توهین‌آمیز، ترویج نفرت قومی و مذهبی، خشونت، تروریسم، پورنوگرافی کودکان، نقض مالکیت فکری و کلاهبرداری برخط باشد. حجم گسترده محتوای تولیدشده توسط کاربران، مرزهای سنتی آزادی بیان را دستخوش تغییر کرده و اهمیت «تعدیل محتوا» را برجسته ساخته است. افزون‌براین، فرایند تعدیل محتوا نیز مسئله‌ساز است؛ از یک‌سو، فقدان یا ضعف آن می‌تواند امنیت عمومی و اعتماد اجتماعی را به خطر اندازد، و از سوی‌دیگر، مداخله‌های گسترده و غیرشفاف در تعدیل، ممکن است آزادی بیان و حقوق بنیادین کاربران را محدود کند.

با افزایش اهمیت این مسئله، مطالعات پژوهشی متعددی به بررسی ابعاد فنی، اجتماعی و مدیریتی تعدیل محتوا پرداخته‌اند و در سال‌های اخیر، توجه پژوهشگران بیش‌ازپیش به جنبه‌های حقوقی این پدیده معطوف شده است. حسین حسنی (۱۴۰۳) در مقاله «حکمرانی و تنظیمگری سکوهای ویدئویی در ایران: از منازعه بر سر نهاد متولی تا تعدیل محتوا»، نشان داده است که تنظیم‌گری محتوا در ایران به یک چالش حقوقی تبدیل شده و مناقشات در سه سطح بررسی شده است: اختلاف در تعریف محتوای ویدئویی، تعیین نهاد تنظیم‌گر و شیوه‌های تعدیل محتوا. یافته‌ها حاکی از آن است که تدوین قانون جامع ارتباطات و تأسیس یک نهاد کلان تنظیم‌گر می‌تواند چهارچوب روشنی برای مسئولیت سکوها و کاربران ایجاد کرده و سازوکارهای تعدیل محتوا را بهبود بخشد. همچنین، پژمان الهامی

طالشمیکائیل و همکاران (۱۴۰۳) در مقاله «استفاده از هوش مصنوعی در تعدیل محتوا»، نشان داده‌اند که حذف خودکار محتوا با هوش مصنوعی اثربخشی محدودی دارد و ممکن است اصول آزادی بیان و مردم‌سالاری را نقض کند. یافته‌ها تأکید دارند که روش «پلتفرم‌زدایی» – که در آن هوش مصنوعی محتوای خشونت‌آمیز را شناسایی می‌کند؛ اما تصمیم نهایی حذف توسط مدیران یا مالکان سکو اتخاذ می‌شود – نسبت به حذف خودکار محتوا، اثرگذاری بیشتری دارد.

بر این‌اساس، نوآوری پژوهش حاضر در این است که منابع خارجی به شکل گسترده و با نگاه‌های مختلف به موضوع تعدیل محتوا پرداخته‌اند؛ لکن در نظام حقوقی ایران ادبیات پژوهشی مرتبط کمرنگ بوده و تنها به‌طور پراکنده به آن اشاره شده و بررسی این مفهوم تاکنون صورت نگرفته است. مقاله حاضر با بهره‌گیری از روش پژوهش توصیفی–تحلیلی و مبتنی بر مطالعات اسنادی، از طریق بررسی منابع حقوقی، مقررات داخلی سکوها و نمونه‌هایی از نظام‌های حقوقی منتخب می‌کوشد ضمن تبیین مفهوم و ماهیت تعدیل محتوا، ضرورت‌های آن را واکاوی کرده و مهم‌ترین چالش‌های حقوقی موجود را شناسایی نماید.

۱. مفهوم و ماهیت تعدیل محتوای کاربران

تعدیل محتوا فرایند سازمان‌یافته‌ای است که توسط سکوها[1] و واسطه‌های اینترنتی[2] برای شناسایی، ارزشیابی و مدیریت محتوای تولیدشده توسط کاربران انجام می‌شود. این محتوا می‌تواند شامل متن، تصویر، ویدئو یا ترکیبی از انواع رسانه‌ها باشد که در تارنماها، شبکه‌های اجتماعی و سایر بسترهای برخط منتشر می‌شود (*De Gregorio, 2020; Roberts, 2019, quoted in Lagren, 2023, p. 11*). عملیات تعدیل محتوا مجموعه‌ای از اقداماتی است که متوجه محتوا و کاربران می‌شود. در حوزه محتوا می‌تواند شامل کاهش رتبه، محدودسازی دسترسی، کاهش امکان درآمدزایی یا حذف آن باشد و در ارتباط با کاربران نیز به مسدودسازی یا تعلیق حساب‌های کاربری منجر شود. افزون‌براین، سکوها با تدوین شرایط و ضوابط مشخص، فرایند مدیریت محتوا را سامان‌دهی می‌کنند تا تعامل کاربران

1. Platform
2. Internet Intermediaries

با محتوا به‌صورت کنترل‌شده انجام شود و حقوق آنان مورد حمایت قرار گیرد (*Grippo, 2025, p. 9*). هر سکو، سیاست‌ها و رویه‌های خاص خود را برای تعدیل محتوا تدوین می‌کند که بازتاب‌دهنده‌ی برند، اعتبار، میزان پذیرش ریسک و نوع تعامل کاربری مورد نظر آن است؛ بنابراین، شیوه‌های تعدیل محتوا از یک سکو به سکوی دیگر متفاوت است و می‌تواند از وضعیت‌های غیرسازمان‌یافته و غیرشفاف تا فرایندهای کاملاً نهادینه‌شده و شفاف متغیر باشد (*Roberts, 2017, p. 1; Grippo, 2025, p. 9*).

فناوری‌های مورد استفاده در تعدیل محتوا به نوع رسانه‌ای که محتوا در آن بیان شده (متن، صدا، تصویر، ویدئو یا ترکیبی از آن‌ها) وابسته است و از رویکردهای ساده تا روش‌های بسیار پیچیده هوش مصنوعی متغیر است (*Michalski et al., 2013, quoted in Galli et al., 2023, p. 68*).

برخی از شیوه‌های رایج تعدیل محتوا که معمولاً به‌کار می‌روند عبارت‌اند از:

۱. تعدیل پیش از انتشار[1]: در این روش، محتوا پیش از انتشار عمومی بررسی می‌شود تا از انتشار مطالب نامناسب، غیرقانونی یا مضر جلوگیری شود. این شیوه بیشترین سطح کنترل را فراهم می‌کند و معمولاً در محیط‌هایی به‌کار می‌رود که حساسیت محتوایی بالایی دارند؛ مانند فضاهای آموزشی یا سکوهای کودکان. با این‌حال، این روش می‌تواند باعث کاهش سرعت تعامل کاربران و محدود شدن جریان آزاد اطلاعات شود (*Imagga, 2024; Veglis, 2014, P. 143*).

۲ . تعدیل پس از انتشار[2]: در این روش، محتوای منتشر شده است و سپس توسط ناظران انسانی یا سیستم‌های خودکار بررسی می‌گردد. در صورت شناسایی تخلف، محتوا حذف یا محدود می‌شود. این روش امکان تعامل سریع و پویای کاربران را فراهم می‌کند؛ اما خطر مواجهه کاربران با محتوای مضر تا زمان حذف آن وجود دارد (*Gillespie, 2018; Roberts, 2019*).

۳. تعدیل واکنشی[3]: تعدیل واکنشی بر اساس گزارش کاربران انجام می‌شود. در این مدل،

1. Pre-Moderation
2. Post-Moderation
3. Reactive Moderation

بررسی محتوا تنها زمانی آغاز می‌شود که کاربران آن را به‌عنوان نامناسب پرچم‌گذاری[1] یا گزارش کنند (*Crawford & Gillespie, 2016, P. 411*). این شیوه هزینه‌های نظارتی سکو را کاهش می‌دهد؛ اما ممکن است به دلیل تأخیر در گزارش یا سوءاستفاده کاربران، کارایی محدودی داشته باشد (*Crawford & Gillespie, 2016; Seering, 2020, P. 17*).

٤ . تعدیل توزیعی[2]: در تعدیل توزیعی، فرایند ارزیابی محتوا میان کاربران توزیع می‌شود و از سازوکارهایی مانند رأی‌گیری، امتیازدهی یا رتبه‌بندی استفاده می‌شود. در این روش، محتوایی که امتیاز پایینی دریافت کند یا گزارش‌های متعددی داشته باشد، به‌طور خودکار مخفی یا محدود می‌شود. این شیوه مشارکت جمعی کاربران را تقویت می‌کند؛ اما ممکن است تحت تأثیر فشار گروهی یا رفتارهای هماهنگ‌شده قرار گیرد (*Veglis, 2014, P. 143; Seering, 2020, P. 17*).

٥ . تعدیل خودکار[3]: تعدیل خودکار مبتنی بر استفاده از الگوریتم‌ها و سیستم‌های هوش مصنوعی برای شناسایی محتوای نامناسب است. این سیستم‌ها از روش‌هایی مانند پردازش زبان طبیعی، تشخیص تصویر و تطبیق اثر انگشت دیجیتال استفاده می‌کنند. مزیت اصلی این روش، توانایی پردازش حجم عظیم محتوا در زمان کوتاه است" اما دقت آن در تشخیص زمینه‌های فرهنگی، طنز یا کنایه محدود است (*Llansó, 2020, P. 4; Halevy et al., 2022*).

٦ . تعدیل ترکیبی[4]: در تعدیل ترکیبی، از ترکیب روش‌های خودکار، انسانی و کاربرمحور استفاده می‌شود تا دقت و کارایی فرایند تعدیل افزایش یابد. در این مدل، سیستم‌های هوش مصنوعی محتوای مشکوک را شناسایی کرده و موارد پیچیده برای بررسی انسانی ارسال می‌شود. این رویکرد به‌عنوان یکی از کارآمدترین شیوه‌های تعدیل شناخته می‌شود، هرچند اجرای آن مستلزم سرمایه‌گذاری گسترده است (*Crawford & Gillespie, 2016; Roberts, 2019*).

1. Flagging
2. Distributed Moderation
3. Automated Moderation
4. Hybrid Moderation

۷ . تعدیل مبتنی بر سیاست[1]: در این شیوه، تصمیمات تعدیل محتوا بر اساس مجموعه‌ای از سیاست‌ها، دستورالعمل‌ها و قوانین داخلی سکوها اتخاذ می‌شود. این سیاست‌ها معمولاً شامل ممنوعیت‌هایی درباره نفرت‌پراکنی، خشونت، اطلاعات نادرست و محتوای جنسی هستند. تعدیل مبتنی بر سیاست به شفافیت تصمیم‌گیری کمک می‌کند؛ اما تفسیر این قوانین می‌تواند در موارد مرزی با چالش همراه باشد (*Gillespie, 2018; Díaz & Hecht-Felella, 2021*). تعدیل محتوا در فضای برخط بر اساس آنچه «استانداردهای جامعه» نامیده می‌شود تنظیم می‌گردد (برای‌مثال: استانداردهای جامعه فیس‌بوک و دستورالعمل‌های جامعه تیک‌تاک). این سیاست‌ها مشخص می‌کنند چه نوع محتوایی در یک سکوی برخط مجاز یا غیرمجاز است و تلاش می‌کنند میان حفظ یک محیط امن برخط و حمایت از آزادی بیان تعادل برقرار کنند (*Halevy et al., 2022*). برای‌نمونه، نحوه برخورد با محتوای دارای برهنگی به‌گونه‌ای است که چنین محتوایی معمولاً ممنوع است؛ اما برای محتوای دارای ارزش پزشکی، مانند ویدئوی یک عمل جراحی، ممکن است استثناهایی در نظر گرفته شود.

۸. تعدیل مبتنی بر جامعه[2]: در تعدیل مبتنی بر جامعه، کاربران نقش فعالی در شناسایی، گزارش و ارزیابی محتوا دارند. این مدل به‌ویژه در سکوهایی مانند ویکی‌پدیا[3] رایج است. مشارکت کاربران می‌تواند حس مسئولیت‌پذیری جمعی را تقویت کند؛ اما هم‌زمان احتمال بروز تعصب، گزارش‌های نادرست یا حذف هدفمند محتوا نیز وجود دارد (*Seering, 2020, P. 8*).

۲. ضرورت‌ها و اهداف تعدیل محتوای کاربران

پویایی‌های جوامع برخط نشان داده است که اینترنت به‌عنوان یک «بازار آزاد اندیشه‌ها» عمل نمی‌کند؛ بازاری که در آن ادعاهای درست، ایده‌های ارزشمند و محتوای اجتماعیِ مفید «کالاها» به‌طور طبیعی در رقابت با اخبار جعلی، محتوای نامناسب و مطالب ضد اجتماعی

1. Policy-Based Moderation
2. Community-Based Moderation
3. Wikipedia

برتری یابند. در سال‌های اخیر، گردش محتوای توهین‌آمیز تولیدشده توسط کاربران به‌طور جدی مورد توجه قرار گرفته است. این امر موجب شده است که سکوهای اشتراک‌گذاری محتوای برخط، به‌جای ارائه «کالاهای اطلاعاتی»، به ابزاری برای انتشار «آسیب‌های اطلاعاتی» تبدیل شوند؛ یعنی محتوایی که ارزش اجتماعی منفی دارد و زیان آن بیش از فایده‌اش است. رفتارهای توهین‌آمیز برخط می‌تواند شامل انواع رفتارهای اطلاعاتی غیرقانونی یا ضداجتماعی باشد؛ از جمله نقض حریم خصوصی یا حیثیت کاربران یا اشخاص ثالث، ترویج خشونت، نژادپرستی یا تروریسم، انتشار محتوای خشونت‌آمیز یا هرزه‌نگاری کودکان و موارد مشابه (*Galli, Loreggia, & Sartor, 2023, Pp. 65-66*).

برای‌مثال، سکوهای شبکه‌های اجتماعی از سازوکارهای تعدیل محتوا برای مقابله با رفتارهای سوءاستفاده‌گرانه کاربران و حذف محتوای غیرقانونی یا مضر بهره می‌گیرند. این فرایند عمدتاً به دو شیوه اصلی انجام می‌شود: تعدیل انسانی که در آن ناظران انسانی محتوای گزارش‌شده یا حساس را بررسی می‌کنند، و تعدیل الگوریتمی که مبتنی بر سامانه‌های هوش مصنوعی برای تحلیل و تصمیم‌گیری خودکار درباره محتوا است. درحال‌حاضر، با توجه به حجم عظیم داده‌ها، تعداد بالای کاربران و تنوع گسترده فعالیت‌ها، سکوهای بزرگ به‌طور فزاینده‌ای به سامانه‌های الگوریتمی متکی شده‌اند. این وابستگی ناشی از ضرورت مدیریت سریع و مؤثر اطلاعات در مقیاس وسیع است. بدون وجود نظام‌های کارآمد تعدیل محتوا، سکوها به‌سرعت با مشکلاتی نظیر هرزنامه، اطلاعات نادرست و محتوای مضر مواجه می‌شوند و کارکرد اصلی آن‌ها مختل خواهد شد. با این‌حال، بسیاری از شرکت‌ها تنها در شرایطی حاضر به سرمایه‌گذاری گسترده در این حوزه می‌شوند که تحت فشار افکار عمومی، رسانه‌ها یا نهادهای نظارتی قرار گیرند؛ زیرا مدیریت محتوا اغلب پرهزینه تلقی شده و مستقیماً با توسعه محصول مرتبط دانسته نمی‌شود (*Díaz & Hecht-Felella, 2021, p. 4*). به‌طور کلی، سکوها از طریق تدوین مجموعه‌ای از دستورالعمل‌ها و استانداردهای جامعه، چهارچوب محتوای مجاز و غیرمجاز را مشخص می‌کنند. این قواعد معمولاً شامل ممنوعیت سخنان نفرت‌آمیز، محتوای افراط‌گرایانه، برهنگی و آزار و اذیت است. اگرچه نحوه بیان این مقررات در سکوهای مختلف متفاوت

است، اصول بنیادین آن‌ها شباهت‌های زیادی دارد. در عمل، تصمیم یک سکوی بزرگ برای حذف محتوای بحث‌برانگیز اغلب موجب واکنش زنجیره‌ای سایر سکوها می‌شود. همچنین، طی دهه اخیر، سیاست‌های تعدیل محتوا از نظر حجم و پیچیدگی به‌طور چشمگیری گسترش یافته‌اند که این امر ناشی از جهانی‌شدن سکوهای دیجیتال، تنوع فرهنگی کاربران و افزایش فشارهای اجتماعی و دولتی است (*Díaz & Hecht-Felella, 2021, p. 4*).

ازسوی‌دیگر، تعدیل محتوا به یکی از چالش‌برانگیزترین موضوعات در حوزه سیاست‌گذاری، ارتباطات و اقتصاد دیجیتال تبدیل شده است؛ زیرا مستقیماً با مباحثی همچون آزادی بیان، سانسور و مشروعیت مداخله سکوها در جریان اطلاعات مرتبط است (*Gorwa et al., 2020, P. 12*). افزون‌برآن، تعدیل محتوا نقش مهمی در تضمین رعایت قوانین ملی و جلوگیری از میزبانی محتوای غیرقانونی ایفا می‌کند (*Roberts, 2019, quoted in Lagren, 2023, p. 12*).

در این چهارچوب، اهداف تعدیل محتوا را می‌توان در چند محور اساسی طبقه‌بندی کرد:

پیشگیری و کاهش آسیب‌های فضای مجازی: یکی از مهم‌ترین اهداف تعدیل محتوا، کاهش آسیب‌های روانی، اجتماعی و فرهنگی ناشی از مواجهه کاربران با محتوای نامطلوب است. این آسیب‌ها به‌ویژه نوجوانان و گروه‌های آسیب‌پذیر را تهدید می‌کند. با این‌حال، برداشت از مفهوم «آسیب» در جوامع مختلف متفاوت است و همین مسئله اجرای سیاست‌های یکسان را دشوار می‌سازد. پژوهش‌ها نشان می‌دهند که مواجهه با سخنان نفرت‌آمیز، تصاویر خشونت‌آمیز و محتوای توهین‌آمیز می‌تواند آثار منفی قابل‌توجهی بر کاربران داشته باشد (*Jhaver et al., 2023, p. 5*).

حفظ فضای امن و سالم برای کاربران: ایجاد محیطی امن برای تعاملات اجتماعی از دیگر اهداف اساسی تعدیل محتوا است. سکوها تلاش می‌کنند از انتشار محتوایی مانند خودآسیب‌رسانی، خشونت، محتوای جنسی نامناسب و نفرت‌پراکنی جلوگیری کنند. در این‌راستا، سیستم‌های تعدیل محتوا نقش پیشگیرانه دارند و می‌کوشند محتوای مضر را پیش از دسترسی گسترده کاربران شناسایی و محدود سازند (*Haimson et al., 2021, p. 25*).

مقابله با اطلاعات نادرست در فرایندهای سیاسی و اجتماعی: گسترش اطلاعات

نادرست یکی از مهم‌ترین چالش‌های رسانه‌های اجتماعی در سال‌های اخیر بوده است. مطالعات متعدد نشان داده‌اند که اخبار جعلی و اطلاعات نادرست به‌سرعت در این فضا منتشر می‌شوند و بر نگرش‌ها و رفتارهای سیاسی و اجتماعی کاربران تأثیر می‌گذارند (*Lazer et al., 2018, Pp. 2-4*). در انتخابات ریاست‌جمهوری ایالات متحده در سال ۲۰۱٦، اطلاعات نادرست نقش قابل‌توجهی در شکل‌دهی به افکار عمومی ایفا کرد و بسیاری از کاربران آن‌ها را به‌عنوان واقعیت پذیرفتند (*Allcott & Gentzkow, 2017,P212*). از این‌رو، مقابله با اطلاعات نادرست به یکی از اهداف سیاست‌های تعدیل محتوا است.

۳. ساختار تنظیم‌گری تعدیل محتوای کاربران

کشورهای مختلف رویکردهای متفاوتی نسبت به تنظیم‌گری تعدیل محتوا اتخاذ کرده‌اند. در اتحادیه اروپا، قوانین سخت‌گیرانه‌ای مانند «قانون خدمات دیجیتال» بر شفافیت و مسئولیت‌پذیری سکوها تأکید دارد. از سال ۲۰۰۰، خدمات برخط در سطح اتحادیه اروپا توسط دستورالعمل خدمات تجارت الکترونیکی[1] تنظیم شده‌اند. این دستورالعمل شامل معافیت از مسئولیت برای ارائه‌دهندگان خدمات میزبانی است؛ زمانی که ارائه‌دهندگان اطلاعی واقعی از محتوای غیرقانونی یا فعالیت غیرقانونی نداشته باشند یا هنگامی که پس از آگاهی از چنین محتوایی سریعاً اقدام کنند. با این‌حال، کشورهای عضو نمی‌توانند الزام عمومی بر ارائه‌دهندگان خدمات جامعه اطلاعاتی برای نظارت بر اطلاعاتی که منتقل یا ذخیره می‌کنند یا الزام عمومی برای جستجوی فعال برای شواهد یا شرایطی که نشان‌دهنده فعالیت غیرقانونی باشد وضع کنند. این قوانین عمومی در دوره ۲۰۰۰ تا ۲۰۲۰ با قوانین خاص اتحادیه اروپا و مقرراتی که با مشارکت فعال بخش‌های مربوطه تدوین شده‌اند، در زمینه خدمات رسانه‌ای صوتی و تصویری[2]، حق تکثیر[3]، تروریسم[4]، سوءاستفاده جنسی از کودکان[5]، نفرت‌پراکنی و اطلاعات نادرست تکمیل شده‌اند (*Grippo, 2025, P. 5*). در سال

1. Directive 2000/31/EC
2. Directive 2010/13/EU
3. Directive (EU) 2019/790
4. Regulation (EU) 2021/784
5. Directive 2011/93/EU

۲۰۱۹، کمیسیون اروپا فرایند تصویب یک بسته‌ی جامع مقرراتی با عنوان «بستهی قانون خدمات دیجیتال» را آغاز کرد. در نتیجه این فرایند، دو مقرره جدید، قانون خدمات دیجیتال[1] و قانون بازارهای دیجیتال[2]، در سال ۲۰۲۲ تصویب شدند.

قانون خدمات دیجیتال به واسطه‌های اینترنتی کمک می‌کند تا کاربران بتوانند حقوق خود، به‌ویژه آزادی بیان و دسترسی به اطلاعات را حفظ کنند. در زمینه مسئولیت، این قانون، معافیت‌های قانونی قبلی که در دستورالعمل تجارت الکترونیک برای واسطه‌های اینترنتی وجود داشت را حفظ کرده است. به‌علاوه، واسطه‌های اینترنتی مجبور نیستند همه محتوای غیرقانونی را به‌طور مداوم بررسی یا جستجو کنند. همزمان، طبق ماده ٦ اگر واسطه‌ها به‌طور داوطلبانه و با حسن‌نیت تلاش کنند محتوای غیرقانونی را شناسایی یا حذف کنند، همچنان از معافیت‌های مسئولیت بهره‌مند می‌شوند و این اقدامات باعث از دست رفتن حمایت قانونی آن‌ها نمی‌شود. ماده ١٤ ارائه‌دهندگان خدمات واسطه‌ای را ملزم می‌کند که تمامی محدودیت‌های مربوط به استفاده از خدمت، به‌ویژه در زمینه محتوای تولیدی کاربران، را به‌صورت شفاف در شرایط و ضوابط[3] خود بیان کنند. این اطلاعات باید شامل سیاست‌ها، رویه‌ها و ابزارهای مورد استفاده برای تعدیل محتوا، اعم از تصمیم‌گیری الگوریتمی و بازبینی انسانی و نیز سازوکار داخلی رسیدگی به شکایات باشد و به زبانی روشن، ساده، قابل‌فهم و بدون ابهام منتشر شود. همچنین هرگونه تغییر اساسی در شرایط استفاده باید به اطلاع کاربران برسد. درصورتی‌که خدمت عمدتاً برای کودکان ارائه می‌شود، توضیحات باید متناسب با درک آنان تنظیم گردد. این ماده تأکید می‌کند که اعمال محدودیت‌ها باید با رعایت اصول دقت، عینیت و تناسب و با توجه به حقوق بنیادین کاربران، ازجمله آزادی بیان، انجام شود. در این چهارچوب، هرگاه ارائه‌دهنده خدمات میزبانی به دلیل غیرقانونی بودن محتوا یا مغایرت آن با شرایط استفاده، اقدام به حذف، محدودسازی، تعلیق حساب یا قطع درآمد کاربر کند، باید هم‌زمان یک «بیانیه دلایل[4]» روشن و مشخص به کاربر ارائه دهد. این بیانیه باید نوع اقدام، مبنای قانونی یا قراردادی

1. Digital Services Act
2. Digital Markets Act
3. Terms and Conditions
4. Statement of Reasons

تصمیم، نقش احتمالی ابزارهای خودکار، و راه‌های اعتراض و جبران را به‌صورت قابل‌فهم توضیح دهد تا کاربر بتواند به‌طور مؤثر از حق دفاع خود استفاده کند.

جهت حفظ حقوق کاربران، ماده ۲۰ واسطه‌ها را ملزم به ایجاد سیستم‌های مؤثر رسیدگی به شکایات کاربران درباره محتوای حذف یا محدودشده می‌کند و بر اساس ماده ۲۱، کاربران حق دارند برای شکایات خود، مرجع حل اختلاف خارج از دادگاه معتبر را انتخاب کنند.

بر اساس مواد ٣٤ و ٣٥، این سکوها موظف‌اند ریسک‌های سیستمی ناشی از طراحی و عملکرد سرویس، سیستم‌های الگوریتمی و استفاده کاربران را شناسایی، ارزیابی و با اقدامات متناسب و مؤثر کاهش دهند؛ این ریسک‌ها شامل انتشار محتوای غیرقانونی، تهدید حقوق بنیادین، تأثیر بر گفت‌وگوهای مدنی و سلامت کاربران است. علاوه‌براین، ماده ۳۷ این سکوها را موظف به ممیزی مستقل توسط نهاد ثالث کرده و مواد ٢٤ و ٤٢ آن‌ها را ملزم به گزارش‌دهی دقیق‌تر درباره فعالیت‌های تعدیل محتوا می‌نماید. بر اساس بند توضیحی شماره ٤٥، ٤٩ و ٦٦ مقدمه، ارائه‌دهندگان خدمات واسطه‌ای موظف‌اند شرایط و محدودیت‌های اعمال‌شده بر محتوای کاربران و روش‌های تعدیل محتوا را به‌طور شفاف، قابل فهم و در دسترس ارائه کنند و تغییرات مهم در این شرایط و ضوابط را به کاربران اطلاع‌رسانی نمایند. افزون‌براین، کمیسیون اروپا پایگاه داده‌ای را مدیریت می‌کند که شامل تصمیمات و دلایل سکوها هنگام حذف یا محدودسازی محتوا است و این پایگاه برای نظارت بر فرایند تعدیل محتوا و کنترل انتشار محتوای غیرقانونی در فضای برخط به کار می‌رود.

بررسی موارد فوق نشان می‌دهند در قانون خدمات دیجیتال اروپا، شاخص‌هایی مانند اطلاع‌رسانی به کاربران، سیستم‌های رسیدگی به شکایات، امکان انتخاب مرجع حل اختلاف، شفافیت الگوریتمی، ارزیابی و کاهش ریسک‌های سیستمی، گزارش‌دهی و شفافیت عمومی و همچنین تضمین رعایت استانداردهای تعدیل محتوا پیش‌بینی شده است.

تقریباً همه سکوهایی که امکان تولید محتوای کاربران را فراهم می‌کنند، سیاست‌هایی درباره محتوایی که حاضر به میزبانی آن هستند یا نیستند تدوین کرده‌اند؛ حتی اگر این سیاست‌ها همیشه به‌صورت عمومی منتشر نشوند؛ البته سیاست‌های محتوایی سکوهای

مختلف از جهات گوناگون تفاوت‌های زیادی با یکدیگر دارند. سکوها از نظر نوع محتوای مجاز یا غیرمجاز، میزان سخت‌گیری در اجرای قوانین و سازوکارهای اجرایی با یکدیگر متفاوت‌اند. با این‌حال، تقریباً همه سکوها دارای دو مجموعه معیار محتوایی هستند: ۱- یک مجموعه قوانین حداقلی که مشخص می‌کند چه محتوایی اصولاً اجازه انتشار در سکو را دارد؛ ۲- مجموعه‌ای سخت‌گیرانه‌تر که استانداردهای مربوط به محتوای تبلیغاتی را تعیین می‌کند (*Edelson, 2024, p. 150*).

در ایالات متحده، آزادی بیان جایگاه ویژه‌ای دارد و بیشتر بر خودتنظیمی شرکت‌ها تکیه می‌شود. در متمم اول قانون اساسی ایالات متحده تصریح شده است که کنگره نباید هیچ قانونی وضع کند که آزادی بیان یا آزادی مطبوعات را محدود کند.

شرکت‌های بزرگ فناوری در ایالات متحده از جمله متا[1] به‌طور گسترده از سازوکار خودتنظیمی برای مدیریت و تعدیل محتوا استفاده می‌کنند، به‌گونه‌ای‌که شرکت‌ها خود چهارچوب‌های رفتاری، معیارهای محتوایی و سازوکارهای اجرایی را تدوین و اجرا می‌نمایند. این رویکرد خودتنظیمی در فضای حقوقی آمریکا ریشه در حفاظت گسترده از آزادی بیان و تکیه بر مسئولیت شرکت‌ها برای تدوین استانداردهای داخلی دارد. متا به‌عنوان مالک سکوهایی مانند فیسبوک، اینستاگرام، مسنجر و تردز، یک نمونه بارز این مدل است؛ این شرکت چهارچوب‌های کلی مدیریت محتوا را در قالب سیاست‌ها[2] تدوین کرده که محور آن‌ها شامل مقابله با محتوا و رفتارهای مضر مانند هماهنگی برای آسیب، ترویج جرم، خشونت، سوءاستفاده جنسی، رفتار نفرت‌آمیز، اطلاعات نادرست، نقض حریم خصوصی، نقض مالکیت فکری و سایر موارد است.[3] در کنار سیاست‌ها، متا استانداردهای جامعه را برای تعیین محتوای مجاز، محدود یا ممنوع تدوین کرده که این استانداردها در همه سکوهای مرتبط اجرا می‌شود و بر اصولی مانند احترام به کرامت انسانی، ایمنی، اصالت و حریم خصوصی استوار است. درواقع، این چهارچوب‌های خودتنظیمی به متا

1. Meta Platforms
2. Policies
۳. متا، «سیاست‌ها»، مرکز شفافیت متا، در https://transparency.meta.com/policies/ : تاریخ بازیابی: ۴ اسفند ۱۴۰۴.

امکان می‌دهد تا پاسخگویی نسبت به کاربران و جامعه جهانی خود را مدیریت کند، درحالی‌که ایالات متحده به‌طور سنتی بیشتر به اعتماد به شرکت‌ها برای تدوین و اجرای سیاست‌های محتوا تمایل دارد تا تحمیل قواعد دقیق و یکسان از سوی دولت فدرال[1].

درمقابل، برخی کشورها مدل کنترل دولتی را دنبال می‌کنند و نظارت مستقیم بر محتوای برخط اعمال می‌نمایند. دولت چین، با اعمال سیاست‌های سختگیرانه در زمینهٔ امنیت سایبری و مدیریت فضای مجازی، تلاش می‌کند تا از نفوذ خارجی و تهدیدات سایبری جلوگیری کند. این سیاست‌ها شامل نظارت بر زیرساخت‌های کلیدی و حفظ حریم خصوصی شهروندان است و به توسعه فناوری‌های داخلی با استفاده از فناوری‌های پیشرفته نظیر هوش مصنوعی کمک می‌کند *(حسینی و رامشه، ۱٤۰٤، ص. ۵۹)*.

چین دارای مجموعه‌ای گسترده از قوانین، مقررات و سیاست‌هاست که چهارچوب مدیریت و تعدیل محتوای برخط را مشخص می‌کنند. از جمله می‌توان به «مقررات مربوط به مدیریت اکوسیستم محتوای اطلاعات برخط[2]» اشاره کرد بر اساس ماده ٦ این مقررات، محتوای محتوای غیرقانونی شامل مواردی مانند: مخالفت با اصول اساسی قانون اساسی[جمهوری خلق چین]؛ تهدید امنیت ملی، افشای اسرار دولتی، براندازی یا خدشه به وحدت ملی؛ لطمه به حیثیت و منافع کشور؛ توهین یا تخریب جایگاه قهرمانان و شهدا؛ ترویج تروریسم و افراط‌گرایی؛ تحریک نفرت یا تبعیض قومی؛ تضعیف سیاست‌های دینی دولت؛ انتشار شایعات مخل نظم اقتصادی و اجتماعی؛ اشاعه هرزه‌نگاری، خشونت، قمار یا تحریک به جرم؛ توهین، افترا یا نقض حقوق و حریم خصوصی اشخاص؛ و سایر موارد ممنوعه در قوانین و مقررات است. همچنین در ماده ۷ محتوای منفی به مواردی اطلاق می‌شود؛ ازجمله: استفاده از عناوین اغراق‌آمیز یا گمراه‌کننده؛ برجسته‌سازی شایعات و رسوایی‌ها؛ اظهارنظر نامناسب درباره بلایای طبیعی یا حوادث بزرگ؛ محتوای دارای اشارات یا تحریکات جنسی؛ تصاویر خشن یا هولناک آزاردهنده؛ مطالب تحریک‌آمیز علیه

1. متا، «استانداردهای انجمن»، مرکز شفافیت متا در:
https://transparency.meta.com/policies/community-standards/. تاریخ بازیابی: ۴ اسفند ۱۴۰۴.

2. Provisions on the Governance of the Online Information Content Ecosystem. Order No. 5, National Internet Information Office, December 15, 2019

گروه‌ها یا مناطق؛ محتوای مبتذل و عامیانه؛ مطالبی که موجب الگوبرداری رفتارهای پرخطر توسط کودکان شود؛ و سایر مطالبی که آثار منفی بر اکوسیستم برخط داشته باشد. در این چهارچوب، سکوهای خدمات محتوای برخط مکلف‌اند محتوای غیرقانونی را شناسایی، حذف، ثبت و به مراجع ذی‌صلاح گزارش کنند و نسبت به پیشگیری و مقابله با انتشار محتوای منفی نیز تدابیر مدیریتی لازم را اتخاذ نمایند.

همچنین «قانون امنیت سایبری جمهوری خلق چین»[1] در ماده ۱۲ به‌طور صریح به محتوای غیرقانونی اشاره کرده است. بر اساس این ماده، هیچ فرد یا سازمانی حق ندارد از اینترنت برای انجام فعالیت‌هایی که امنیت ملی، حیثیت و منافع کشور را به خطر می‌اندازد استفاده کند؛ ازجمله تحریک به براندازی حاکمیت ملی یا تجزیه‌طلبی، ترویج تروریسم و افراط‌گرایی، اشاعه نفرت قومی، انتشار اطلاعات خشونت‌آمیز، مستهجن یا جنسی، انتشار اطلاعات کذب که نظم اقتصادی یا اجتماعی را مختل می‌کند و نیز انتشار مطالبی که به آبرو، حریم خصوصی، مالکیت فکری یا سایر حقوق و منافع قانونی اشخاص لطمه می‌زند. این تصریح قانونی، مبنای حقوقی شناسایی، حذف و جلوگیری از انتشار محتوای غیرقانونی در فضای برخط را فراهم می‌کند. نظارت و فیلتر محتوا با استفاده از ابزارهای خودکار و انسانی برای شناسایی و حذف محتوای غیرقانونی یا مضر، ثبت‌نام با نام واقعی کاربران و اطمینان از صحت هویت آنان، نگهداری داده‌ها و همکاری با مقامات قانونی و همچنین سرمایه‌گذاری در امنیت سایبری برای حفاظت از کاربران در برابر تهدیدهای سایبری، محتوای غیرقانونی و کلاهبرداری برخط صورت می‌گیرند.

وی‌چت به‌عنوان یکی از پرکاربردترین سکوهای ارتباطی در چین، نه‌تنها یک ابزار پیام‌رسان، بلکه یک زیست‌بوم دیجیتال جامع برای خدمات مالی، اداری، تجاری و اجتماعی محسوب می‌شود – توسط شرکت تنسنت[2] توسعه یافته است، نقش مهمی در اجرای سیاست‌های حاکمیتی ایفا می‌کنند و ازاین‌رو، تحت نظارت نهادهای دولتی قرار دارد. در این چهارچوب، شرکت‌هایی مانند تنسنت موظف‌اند سازوکارهای مؤثر برای پایش، پالایش

1. Cybersecurity Law of the People’s Republic of China. Promulgated by the Standing Committee of the National People’s Congress, November 7, 2016. Effective June 1, 2017.
2. Tencent

و حذف محتوای مغایر با سیاست‌های دولتی، نظم عمومی و امنیت ملی ایجاد کنند. این سازوکارها شامل استفاده گسترده از الگوریتم‌های هوشمند، فیلترهای کلمات کلیدی، سامانه‌های تشخیص خودکار محتوا و شبکه‌ای از ناظران انسانی است که به‌صورت مستمر فعالیت کاربران را رصد می‌کنند *(حسینی و رامشه، ۱٤۰٤، صص. ٦٧-٦٨)*.

در ایران، مواجهه با شبکه‌های اجتماعی از بیش از یک دهه گذشته یکی از چالش‌های اصلی سیاست‌گذاران فضای مجازی بوده است *(بصیریان و خانیکی، ۱۳۹۳، ص. ۲)*. دولت ایران برای مدیریت محتوا اقداماتی مانند فیلترینگ وبگاه‌ها، توسعه پیام‌رسان‌های داخلی، تشکیل شورای عالی فضای مجازی و راه‌اندازی شبکه ملی اطلاعات انجام داده است *(اسلامی و جعفری‌مقدم، ۱۳۹۹)*. از سال ۱۳۸۰، کمیته‌هایی مانند «کمیته مصادی پایگاه‌های اطلاع‌رسانی رایانه‌ای غیرمجاز» با هدف شناسایی و مسدودسازی محتوای مجرمانه مرتبط با عفت عمومی، امنیت، مقدسات و جرایم رایانه‌ای تشکیل شده‌اند *(فرهنگی و دیگران، ۱۳۹۸)*. این اقدامات عمدتاً متمرکز و دولتی هستند و مشارکت کاربران در فرایند تعدیل محتوا محدود است.

روش‌های محدودسازی محتوا، به‌ویژه بر اساس کلمات کلیدی، گاه موجب مسدود شدن محتوای مفید و دسترسی محدود به منابع علمی می‌شوند *(فرهنگی و دیگران، ۱۳۹۸)*. در کنار تعدیل، ایران تلاش کرده با تنظیم مقررات، محتوای متناسب با جنسیت، سن، ملیت و مذهب کاربران ارائه دهد؛ اما خلأهای قانونی و ضعف در حوزه‌هایی مانند حقوق مالکیت معنوی، حفاظت از کودکان، حریم خصوصی و مسئولیت ارائه‌دهندگان خدمات اینترنتی مانع اجرای کامل این سیاست‌ها شده است *(هاتف، ۱۳۹۵؛ شریفی، ۱۳۹۷، ص. ۲۵۸)*.

ساختار نهادی و تقنینی ایران در حوزه فضای مجازی چندسطحی است. در سطح کلان، شورای عالی فضای مجازی و شورای عالی انقلاب فرهنگی بازیگران اصلی سیاست‌گذاری هستند و نهادهایی مانند شورای اجرایی فناوری اطلاعات، کارگروه تعامل‌پذیری دولت الکترونیک، کمیسیون تنظیم مقررات ارتباطات و مرکز ملی فضای مجازی سیاست‌های اجرایی را پیاده‌سازی می‌کنند *(رجبی، ۱٤۰۳، ص. ۷)*. با این‌حال، سکوها و ارائه‌دهندگان خدمات میزبانی در ایران فاقد چهارچوب قانونی منسجم هستند و قوانین و مقررات موجود تکالیف

بسیار محدود و پراکنده‌ای برای آن‌ها تعیین کرده‌اند؛ به‌عنوان نمونه:

در سطح کیفری، ماده ۷۴۹ قانون مجازات اسلامی ارائه‌دهندگان خدمات دسترسی را مکلف به پالایش محتوای مجرمانه بر اساس تصمیمات «کارگروه تعیین مصادیق محتوای مجرمانه» کرده و برای استنکاف، ضمانت اجرای کیفری پیش‌بینی نموده است. ماده ۷۵۱ نیز ارائه‌دهندگان خدمات میزبانی را در صورت عدم اجرای دستور قضائی برای حذف یا مسدودسازی محتوا، مسئول می‌داند. مقصود از «محتوای مجرمانه» در این چهارچوب، صرفاً یک مفهوم کلی نیست بلکه دامنه‌ای گسترده از مصادیق قانونی را دربرمی‌گیرد که در قوانین مختلف احصا شده‌اند؛ ازجمله: جرایم علیه عفت و اخلاق عمومی (مانند اشاعه فحشا)، جرایم علیه مقدسات اسلامی و توهین به مقدسات، جرایم علیه امنیت و آسایش عمومی (تبلیغ علیه نظام، افشای اسناد محرمانه، تحریک به برهم زدن امنیت)، توهین و افترا نسبت به اشخاص و مقامات، جرایم رایانه‌ای نظیر دسترسی غیرمجاز و انتشار ابزار ارتکاب جرم، جرایم مرتبط با انتخابات، قمار و شرط‌بندی، نقض حقوق مالکیت معنوی، جرایم علیه کودکان و نوجوانان *(دادستانی، بی‌تا)*.

در سطح مقرراتی نیز، ماده ۷ «آیین‌نامه ساماندهی پایگاه‌های اطلاع‌رسانی اینترنتی مصوب ۱۳۸۵ هیئت وزیران» انتشار و نگهداری هرگونه داده (اعم از متن، تصویر، صوت و...) حاوی مضامینی چون: مطالب الحادی و تضعیف ارزش‌های اسلامی، توهین به ادیان و مقدسات، اقدام علیه امنیت و وحدت ملی، تحریف بیانات بنیان‌گذار جمهوری اسلامی یا رهبری، افشای اسناد طبقه‌بندی‌شده، اشاعه فحشا، توهین و افترا، انتشار اطلاعات خصوصی اشخاص بدون اجازه، فعالیت‌های اقتصادی غیرقانونی، آموزش مقابله با فیلترینگ و نشر اکاذیب را ممنوع اعلام کرده است. مطابق ماده ۸ این آیین‌نامه، انتشار داده‌های ممنوع موجب ارسال گزارش از سوی وزارت فرهنگ و ارشاد اسلامی به کارگروه تعیین مصادیق پایگاه‌های غیرمجاز اینترنتی برای تصمیم‌گیری درباره مسدودسازی خواهد شد. ماده ۹ نیز طیفی از ضمانت اجراها را پیش‌بینی می‌کند: از تذکر رسمی و الزام به حذف محتوا یا عذرخواهی، تا درخواست مسدودسازی موقت یا تعطیلی دائم پایگاه اطلاع‌رسانی؛ همچنین امکان درخواست تجدیدنظر برای پایگاه‌های مسدودشده

را پیش‌بینی کرده است.

در این چهارچوب، حذف یا پالایش محتوای مجرمانه ماهیتی حاکمیتی دارد نه رضایت قراردادی کاربر و هیچ مقرره‌ای رضایت کاربر را شرط اجرای دستور قضائی یا تصمیم کارگروه ندانسته است؛ البته بسیاری از سازوکارهای تعدیل محتوا در عمل مبتنی بر پردازش داده‌های رفتاری و هویتی کاربران‌اند؛ ازجمله الگوریتم‌های تشخیص تخلف، نظام‌های گزارش‌دهی کاربران و سازوکارهای احراز هویت. از این‌حیث، تعدیل محتوا صرفاً یک اقدام محتوایی نیست بلکه با پردازش داده‌های شخصی نیز تلاقی پیدا می‌کند. در این چهارچوب، «دستورالعمل اجرایی بهبود حفاظت از حریم خصوصی کاربران مصوب ۱۴۰۲ شورای عالی فضای مجازی» تصریح می‌کند که جمع‌آوری و پردازش داده‌ها باید با رضایت صریح و آگاهانه کاربر، بر مبنای اصل حداقل‌گرایی داده و برای هدف معین و شفاف انجام شود؛ بنابراین سکوها نمی‌توانند به بهانه اجرای سیاست‌های تعدیل محتوا، هر نوع داده‌ای را بدون ضابطه، تناسب و اطلاع‌رسانی شفاف گردآوری یا تحلیل کنند. از سوی‌دیگر، بند ۱ـ۳ آیین دستورالعمل حق درخواست حذف داده را برای کاربر شناسایی کرده است؛ اما این حق را مقید به «عدم مغایرت با قوانین و مقررات کشور» و تکالیف مقرر در مواد ۶۶۷ تا ۶۷۰ قانون آیین دادرسی کیفری (در حوزه ادله جرایم رایانه‌ای) می‌داند. این قید نشان می‌دهد که حتی در حوزه حریم خصوصی نیز اراده کاربر مطلق نیست و در برابر الزامات کیفری محدود می‌شود.

در بعد قضائی نیز، با توجه به دشواری نظارت پیشینی در سکوهای کاربرمحور، رویه به سمت پذیرش «تعدیل پسینی» متمایل شده است؛ چنان‌که در رأی دیوان عالی کشور در پرونده شکایت سازمان صداوسیما علیه آپارات به شماره دادنامه ۱۴۰۳۰۶۳۹۰۰۰۰۴۹۳۹۰۱، مسئولیت اولیه متوجه کاربر دانسته و مسئولیت میزبان منوط به عدم اجرای دستور قضائی شده است. این رویکرد تفسیری، اجرای تکالیف قانونی را واقع‌بینانه‌تر کرده است.

باوجوداین، چالش‌هایی همچون گستردگی و تفسیرپذیری برخی مفاهیم مجرمانه و ممنوعه، تعدد مراجع تصمیم‌گیر و فقدان تفکیک روشن میان انواع ارائه‌دهندگان خدمات،

سطح مداخله آن‌ها که ریشه در نبود یک چهارچوب تقنینی منسجم، تفکیک‌شده و متناسب با واقعیت‌های فنی سکوهای دیجیتال دارند. در مقایسه با تجربه سایر کشورها، وضعیت ایران در تعدیل محتوا ترکیبی از کنترل توسط دولتی، تلاش‌های مقرراتی بسیار محدود و خلأهای قانونی گسترده است که چالش‌هایی جدی برای تضمین شفافیت، حفظ حریم خصوصی و حفاظت از حقوق کاربران ایجاد می‌کند.

۴. چالش‌های حقوقی تعدیل محتوای کاربران

۴-۱. کمبود شفافیت در سیاست‌های تعدیل محتوا سکوها

برای آنکه حذف یا مسدودسازی محتوای برخط به‌صورت منصفانه و بدون تبعیض انجام شود، وجود شفافیت در فرایندهای تعدیل محتوا ضروری است. این شفافیت باید هم متوجه افرادی باشد که محتوایشان حذف یا محدود می‌شود و هم کل جامعه مخاطبان را دربربگیرد. در سطح فردی، سکوها موظف‌اند به‌طور روشن به کاربران اطلاع دهند که محتوای آن‌ها حذف یا مسدود شده است، دلیل اتخاذ این تصمیم چه بوده و کاربر از چه طریقی می‌تواند در صورت اعتراض، شکایت یا درخواست بازبینی ثبت کند. در سطح عمومی نیز سکوها باید اطلاعات کلی و آماری منتشر کنند؛ ازجمله تعداد محتواهای حذف‌شده، تعداد حساب‌های تعلیق یا مسدودشده و دلایل کلی این اقدامات. بدیهی است که ارائه این اطلاعات باید بدون افشای هویت یا داده‌های شخصی کاربران انجام شود تا حریم خصوصی آن‌ها حفظ گردد (*Galli et al., 2023, p. 74*).

در سال‌های اخیر، ابتکارهای حقوقی غیرالزام‌آوری با هدف افزایش شفافیت در تعدیل محتوا شکل گرفته‌اند که از مهم‌ترین آن‌ها می‌توان به «اصول سانتا کلارا» اشاره کرد. بر اساس این اصول، شرکت‌هایی که اقدام به تعدیل محتوای کاربران می‌کنند باید آمار مربوط به حذف پست‌ها و تعلیق موقت یا دائم حساب‌ها را منتشر کنند، به هر کاربری که محتوایش حذف یا حسابش تعلیق شده است توضیحی مشخص و قابل‌فهم درباره دلیل این اقدام ارائه دهند و امکان اعتراض مؤثر و به‌موقع به تصمیم‌های اتخاذشده را برای کاربران فراهم کنند (*Santa Clara Principles, 2018*).

برخی پژوهش‌ها نشان می‌دهند که توضیح و شفاف‌سازی فرایندهای نظارتی می‌تواند

به کاهش تولید محتوای مضر در آینده کمک کند؛ بااین‌حال، این شفافیت ممکن است پیامدهای ناخواسته‌ای نیز داشته باشد؛ زیرا آگاهی کاربران از نحوه عملکرد فیلترها می‌تواند به یافتن راه‌هایی برای دور زدن آن‌ها منجر شود (*Gerrard, 2018, pp. 14–19*).

بااین‌حال، در عمل بسیاری از شرایط خدمات سکوها مبهم و نامشخص است و کاربران غالباً نمی‌دانند چه نوع محتوایی مجاز است و چه محتوایی ممکن است حذف یا مسدود شود. افزون‌براین، شرکت‌ها معمولاً اطلاعات دقیقی درباره سازوکارهای تصمیم‌گیری خود در خصوص حذف محتوا یا تعلیق حساب‌ها ارائه نمی‌دهند (*CDMSI, 2021, p. 17*).

۴-۲. اختلاف‌نظر در تعریف و تعیین محتوای مجاز

کاربران شبکه‌های اجتماعی دارای دیدگاه‌های متنوع و گاه متعارضی در مورد محتوای قابل‌قبول هستند. اختلاف‌نظر در تعریف و تعیین محتوای مجاز یکی از پیچیده‌ترین چالش‌های حقوقی و سیاست‌گذاری در مدیریت سکوهای برخط به‌شمار می‌رود. این اختلاف‌نظرها عمدتاً ریشه در تفاوت‌های فرهنگی، اجتماعی و سیاسی میان کاربران دارند و همین امر تعریف مرزهای «مجاز» و «غیرمجاز» را با عدم قطعیت مواجه می‌کند. از منظر حقوقی، مسئله محوری در این حوزه، تدوین معیارهایی شفاف، قابل پیش‌بینی و منصفانه برای محتوای مجاز است؛ معیارهایی که در عین‌حال بتوانند مرزهای قانونی و اخلاقی متفاوت در جوامع گوناگون را نیز در نظر بگیرند. از آن‌جا که این مرزها ممکن است از کشوری به کشور دیگر یا از جامعه‌ای به جامعه دیگر تفاوت داشته باشند، نه‌تنها سیاست‌ها و قواعد داخلی هر سکو، بلکه میزان انطباق آن‌ها با قوانین ملی و بین‌المللی نیز به مسئله‌ای چالش‌برانگیز تبدیل می‌شود. برای‌مثال، سکوها ممکن است قوانینی وضع کنند که بر اساس آن‌ها محتوای تروریستی یا نفرت‌پراکن حذف شود؛ اما این مفاهیم اغلب دارای تعاریف گسترده، مبهم و قابل تفسیر هستند. در چنین شرایطی، ابهام در تشخیص اینکه چه نوع محتوایی مجاز تلقی می‌شود و چه نوعی باید حذف گردد، می‌تواند به مناقشات و مشکلات حقوقی منجر شود و در نهایت آزادی بیان کاربران را به‌طور مستقیم یا غیرمستقیم محدود کند. علاوه‌براین، در نبود یک مرجع مستقل، شفاف و الزام‌آور برای نظارت بر اجرای سیاست‌های تعدیل محتوا، بسیاری از کاربران احساس می‌کنند که تصمیم‌های اتخاذشده

از سوی سکوها ناعادلانه، سلیقه‌ای یا جانبدارانه است. یکی دیگر از مسائل حقوقی مهم در این زمینه، ادعاهای مکرر برخی گروه‌ها درباره سانسور ناعادلانه محتوای آن‌هاست. به‌طور خاص، گروه‌هایی مانند محافظه‌کاران، اقلیت‌های نژادی و کاربران تراجنسیتی بارها اعلام کرده‌اند که محتوای آن‌ها به‌صورت نامتناسب و تبعیض‌آمیز حذف می‌شود. برای نمونه، محافظه‌کاران فیس‌بوک را به جانبداری از جریان‌های خاص متهم کرده‌اند و معتقدند که این سکو برای کنترل یا محدودسازی محتوای سیاسی، سیاست‌های خود را به نفع دیدگاه‌های خاصی تنظیم می‌کند. در همین راستا، کاربران تراجنسیتی نیز از سانسور محتوای خود در تیک‌تاک شکایت کرده‌اند و اذعان داشته‌اند که محتوای آنان بدون دلایل شفاف یا موجه حذف می‌شود (*Haimson et al., 2021*).

۴-۳. مسائل ناشی از الگوریتم‌های تعدیل محتوا

۴-۳-۱. اشتباه‌های فیلترهای خودکار: با توجه به اینکه فیلترهای محتوا ممکن است اشتباه کنند به‌طوری‌که محتوای نامطلوب گاهی از فیلتر عبور می‌کند و محتوای قابل‌قبول گاهی حذف می‌شود هر سیستم اجتماعی-فنی برای تعدیل محتوا باید راهکاری برای اعتراض کاربران به تصمیم‌های فیلتر داشته باشد. چنین اشتباه‌هایی می‌تواند منجر به تضییع حقوق کاربران و مشکلات قانونی برای سکوها شود. راهکار اصلی برای مقابله با این چالش، توانمند کردن کاربران برای اعتراض به تصمیم‌های فیلتر است. کاربران باید بتوانند محتوایی را که غیرقانونی یا نامناسب می‌دانند علامت‌گذاری کرده و دلایل خود برای حذف آن را ارائه کنند، و هم‌زمان، کاربران بارگذارنده محتوا باید امکان اعتراض به حذف پست‌های خود و ارائه دلایل برای نگه داشتن آن را داشته باشند. در هر دو حالت، سادگی فرایند اعتراض و پاسخ سریع ناظران از اهمیت بالایی برخوردار است تا کاربران بتوانند به راحتی و به موقع، تصمیم‌های تعدیل محتوا را به چالش بکشند (*Galli et al., 2023, p. 75*).

۴-۳-۲. سوگیری الگوریتمی: سوگیری الگوریتمی به‌معنای وجود تصمیمات ناعادلانه یا تبعیض‌آمیز در نتایج سیستم‌های هوش مصنوعی است. یعنی وقتی یک الگوریتم محتوای کاربران را مدیریت یا فیلتر می‌کند، ممکن است به‌طور ناخواسته برخی گروه‌ها یا افراد را ترجیح داده یا نادیده بگیرد. الگوریتم‌ها بر اساس داده‌هایی که برای آموزششان استفاده

می‌شود، تصمیم می‌گیرند. اگر این داده‌ها نمایانگر گروه خاصی باشند یا شامل سوگیری‌های انسانی باشند، الگوریتم نیز این سوگیری‌ها را یاد می‌گیرد و اعمال می‌کند. مثلاً الگوریتمی که فقط روی متن‌های انگلیسی آموزش دیده باشد، ممکن است محتوای سایر زبان‌ها را اشتباه حذف کند (*Sawyer, 2024, P. 5*).

۴-۳-۳. عدم درک ظرافت‌ها و پیچیدگی‌های ارتباط انسانی: الگوریتم‌های هوش مصنوعی قادر به تحلیل متن‌ها و شناسایی الگوها هستند؛ اما اغلب در درک جزئیات و پیچیدگی‌های ارتباط انسانی ناکام می‌مانند. این ضعف می‌تواند منجر به طبقه‌بندی نادرست محتوا شود، به‌طوری‌که بیان قانونی و مشروع به اشتباه به‌عنوان محتوای مضر شناسایی گردد. برای‌نمونه، طنز، پارودی یا تحلیل‌های سیاسی ظریف ممکن است توسط هوش مصنوعی به شکل نادرست تفسیر شوند و حذف نامناسب محتوا یا محدود شدن گفتگوهای ارزشمند را به‌دنبال داشته باشند. برای رفع این مشکل، ضروری است که سیستم‌های تعدیل مبتنی بر هوش مصنوعی با سازوکارهایی طراحی شوند که حساسیت به زمینه و انعطاف‌پذیری را در نظر بگیرند. یکی از مهم‌ترین این سازوکارها، نظارت انسانی بر محتوای علامت‌گذاری‌شده است تا تصمیمات اتخاذشده با اصول آزادی بیان هماهنگ باشند و از حذف ناعادلانه محتوای مشروع جلوگیری شود (*Sawyer, 2024, Pp. 3-4*).

۴-۴. نقض حریم خصوصی: با استفاده روزافزون از هوش مصنوعی در مدیریت و تعدیل محتوای تولیدشده توسط کاربران، مسائل مربوط به حفظ حریم خصوصی اهمیت فزاینده‌ای پیدا کرده است. سیستم‌های هوش مصنوعی برای شناسایی و مدیریت محتوای مضر، نیازمند تحلیل حجم بالایی از داده‌های کاربران هستند که اغلب شامل اطلاعات حساس شخصی می‌شوند، مانند اطلاعات پزشکی، مالی یا سایر شناسه‌های شخصی. یکی از نگرانی‌های اصلی، جمع‌آوری و پردازش گسترده داده‌ها است. راهکارهای کاهش خطر نقض حریم خصوصی شامل حداقل‌سازی داده‌ها و حذف یا ناشناس‌سازی داده‌ها است. حداقل‌سازی داده‌ها به معنای جمع‌آوری تنها اطلاعات لازم برای هدف مشخص است و ناشناس‌سازی شامل حذف یا مخفی کردن شناسه‌های شخصی کاربران می‌شود، بدون اینکه اثرگذاری سیستم در شناسایی محتوای مضر کاهش یابد (*Sawyer, 2024, Pp. 3-4*).

۴-۵. ابهام در مقررات بین‌المللی و تطابق با قوانین ملی: سکوهای جهانی موظف است با قوانین و مقررات متنوع کشورهای مختلف هماهنگ عمل کند، امری که معمولاً به ناهماهنگی و پیچیدگی‌های اجرایی قابل‌توجه منجر می‌شود؛ برای‌مثال، قوانینی که در یک کشور اجازه انتشار محتوای خاص را می‌دهند، در کشور دیگر ممکن است ممنوعیت داشته باشند، و این باعث می‌شود سکوها در تصمیم‌گیری درباره حذف یا مجاز دانستن محتوا با ابهامات جدی مواجه شوند. در این‌راستا، مقررات جدیدی مانند قانون خدمات دیجیتال اتحادیه اروپا تلاش کرده‌اند با تفکیک مسئولیت‌ها و الزامات بر اساس اندازه و نقش سکوها (سکوهای بزرگ در مقابل کوچک، یا سکوهای عمومی در مقابل تخصصی)، چهارچوب حقوقی روشن‌تری ارائه کنند. با این‌حال، همچنان ابهاماتی در نحوه تفسیر و اجرای این الزامات وجود دارد و برخی رویه‌های عملی می‌تواند باعث تردید یا تناقض در اعمال قانون شود (*Allakuliev, 2025, Pp. 25-26*).

چالش‌های مطرح‌شده نشان می‌دهد که نبود معیارهای حقوقی شفاف و سازوکارهای مؤثر پاسخ‌گویی در حوزه تعدیل محتوا، خطر انحراف این فرایند را افزایش می‌دهد و می‌تواند به مداخله‌های نامتناسب، تبعیض‌آمیز و غیرقابل پیش‌بینی در فعالیت کاربران بینجامد. در چنین شرایطی، تصمیم‌های سکوها نه‌تنها اعتماد عمومی را تضعیف می‌کند بلکه ممکن است به خدشه‌دار شدن حقوق بنیادینی همچون آزادی بیان، حق دسترسی به اطلاعات و حق اعتراض مؤثر منجر شود. از این‌رو، شکل‌دهی چهارچوبی متوازن برای تنظیم‌گری تعدیل محتوا ضرورت دارد؛ چهارچوبی که همزمان با صیانت از نظم عمومی، تضمین‌کننده شفافیت، تناسب تصمیم‌ها، مسئولیت‌پذیری سکوها و حمایت مؤثر از حقوق کاربران باشد.

نتیجه‌گیری

تعدیل محتوا فرایندی سازمان‌یافته است که توسط سکوها و واسطه‌های اینترنتی برای شناسایی، ارزیابی و مدیریت محتوای تولیدشده توسط کاربران انجام می‌شود. این محتوا می‌تواند شامل متن، تصویر، ویدئو یا ترکیبی از رسانه‌ها باشد و هدف اصلی آن کاهش

انتشار محتوای آسیب‌زا، حفظ فضای امن و سالم برای کاربران، پیشگیری از آسیب‌های روانی و اجتماعی و مقابله با اطلاعات نادرست است. روش‌های تعدیل محتوا متنوع‌اند و از تعدیل پیش از انتشار، پس از انتشار، واکنشی، توزیعی، خودکار، ترکیبی، مبتنی بر سیاست و مبتنی بر جامعه تشکیل می‌شوند که هر یک مزایا و محدودیت‌های خاص خود را دارند و بسته به نوع سکو و محتوای منتشرشده به‌کار گرفته می‌شوند.

تنظیم‌گری تعدیل محتوا در سطح جهانی به مدل‌های مختلفی تقسیم می‌شود: اروپا با قوانینی مانند قانون خدمات دیجیتال بر شفافیت، پاسخگویی و حقوق کاربران تأکید دارد؛ ایالات متحده عمدتاً بر خودتنظیم‌گری شرکت‌ها و حفاظت از آزادی بیان تکیه می‌کند؛ و چین مدل کنترل دولتی و نظارت مستقیم بر محتوا را دنبال می‌کند. در ایران نیز، برای تعدیل محتوا الزام قانونی وجود ندارد. هرچند اقداماتی مانند فیلترینگ، توسعه پیام‌رسان‌های داخلی و ایجاد نهادهای سیاستگذار در حوزه فضای مجازی صورت گرفته است؛ لکن همچنان چهارچوب قانونی منسجم و جامع برای تعدیل محتوا شکل نگرفته است. این خلأهای قانونی، محدودیت مشارکت کاربران و نبود سازوکارهای شفاف و پاسخگو، چالش‌های جدی در تضمین حقوق کاربران و اجرای مؤثر سیاست‌های مدیریت محتوا ایجاد کرده است. چالش‌های حقوقی عمده تعدیل محتوای کاربران شامل کمبود شفافیت در سیاست‌های تعدیل محتوا، اختلاف در تعریف محتوای مجاز، مسائل ناشی از الگوریتم‌های خودکار، نقض احتمالی حریم خصوصی و ابهام در تطابق مقررات بین‌المللی با قوانین ملی هستند.

منابع

الهامی طالشمیکائیل، پژمان، کردکریمی، امیر، شیرپور، ولی و اوروجی، حسام (۱٤۰٤). استفاده از هوش مصنوعی در تعدیل محتوا. *شناخت‌پژوهی مطالعات سیاسی*، 2(5)، ٦٥-۸۳.

بصیریان جهرمی، حسین و خانیکی، هادی (۱۳۹۳). سیاست‌گذاران ایرانی و سیاست‌گذاری رسانه‌های اجتماعی: چالش‌ها، الگوها و ارائه یک مدل پیشنهادی. *فصلنامه برنامه‌ریزی رفاه و توسعه اجتماعی*، ش۲۱، ۷-۲۵.

حسنی، حسین (۱٤۰۳). حکمرانی و تنظیمگری پلتفرم‌های ویدئویی در ایران: از منازعه بر

سر نهاد متولی تا تعدیل محتوا .*حقوق فناوری‌های نوین*، (10)5، ۳۳-۴۸.

حسینی، محمدرضا و نوریان رامشه، میلاد (۱٤۰٤). بومی‌سازی رسانه‌های اجتماعی در چین: تحلیل تأثیر سیاست‌های دولتی و فناوری محلی بر توسعه پلتفرم‌های بومی؛ مطالعه موردی: وی‌چت و سینا ویبو .*مطالعات فضای مجازی و رسانه‌های اجتماعی*، 2(4)، ۵۷-۸٦

اسلامی، روح‌الله و جعفری‌مقدم، حجت (۱۳۹۹). ساخت حکومت و سیاست‌گذاری شبکه‌های اجتماعی در ایران، *سیاست‌گذاری عمومی*، ش (٦)٤، ۱۵۱-۱۷۱.

فرهنگی، محمدمهدی، میرترابی، سعید، و گلشنی، علیرضا (۱۳۹۸) تبیین سیاست‌گذاری فضای مجازی در جمهوری اسلامی ایران (الزامات و اصلاح در چهارچوب اهداف مورد نظر در اسناد بالادستی). *سیاست‌گذاری عمومی*، ش (۸)۳، ۱۱۱-۱۳۲.

هاتف، میدی (۱۳۹۵). الزامات و شاخص‌های امنیت در فضای مجازی، *مجله پژوهش‌های کاربردی*، ش ٦.

رجبی، احمد (۱٤۰۳). مروری بر مفاهیم و نهادهای مرتبط با فضای مجازی (رایاسپهر) و ارائه چشم‌انداز تقنینی و نظارتی مجلس دوازدهم،*گزارش‌های نظارتی* مرکز پژوهش‌های مجلس شورای اسلامی، ش (۳۲)۱۱.

شریفی، سیدمهدی؛ مرزبان، بیتا و لبافی، سمیه (۱۳۹۷) آسیب‌شناسی سیاست‌گذاری تولید محتوا در فضای مجازی در ایران، *مدیریت دولتی*، ش (۱۰)۲، ۲۵۱-۲٦۸.

اصول سانتا کلارا (۲۰۱۸). اصول سانتا کلارا درباره شفافیت و پاسخ‌گویی در تعدیل محتوا، (تاریخ مشاهده: ۱۸ بهمن ۱٤۰٤) در:/https://santaclaraprinciples.org

دادستانی (بی‌تا)، قوانین و مقررات فضای مجازی (تاریخ مشاهده: ۱۸ بهمن ۱٤۰٤) در: https://dadsetani.ir/قوانین-فضای-مجازی

متا (بی‌تا). «استانداردهای انجمن»، مرکز شفافیت متا، (تاریخ مشاهده: ۱۸ بهمن ۱٤۰٤) در: https://transparency.meta.com/policies/community-standards/

متا (بی‌تا). سیاست‌ها، مرکز شفافیت متا. (تاریخ مشاهده: ۱۸ بهمن ۱٤۰٤) در: https://transparency.meta.com/policies/

Allakuliev, M,. (2025). Tort Liability of Digital Platforms for User-Generated Content. *Uzbek Journal of Law and Digital Policy*, 3(5), 20-35. https://doi.org/10.59022/ujldp.376

Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2), 211–236.

CDMSI (2021). Content moderation – Guidance Note [PDF]. Council of Europe. (Retrieved February 9, 2026) from https://rm.coe.int/content-moderation-guidance-note-gbr/1680a3b7c3

Crawford, K., & Gillespie, T. (2016). What is a flag for? Social media reporting tools and the vocabulary of complaint. *New Media & Society*, 18(3), 410–428.

De Gregorio, G. (2020). Democratising online content moderation: A constitutional framework. *Computer Law & Security Review*, 36. https://doi.org/10.1016/j.clsr.2019.105374

Díaz, Á., & Hecht-Felella, L. (2021). Double standards in social media content moderation. *Brennan Center for Justice at New York University School of Law*, 1–23.

Edelson, L. (2024). Content moderation in practice. In K. Langvardt & J. (G.) Hurwitz (Eds.), Platform governance (Part III, pp. 149–165). Cambridge University Press. https://doi.org/10.1017/9781009174411.017

Galli, F., Loreggia, A., & Sartor, G. (2023). The regulation of content moderation. In The legal challenges of the fourth industrial revolution (pp. 63–87). Cham: Springer. https://doi.org/10.1007/978-3-031-40516-7_5

Grippo, V., (2025). Regulating content moderation on social media to safeguard freedom of expression. Report Doc. 16089. Committee on Culture, Science, Education and Media, Parliamentary Assembly of the Council of Europe (PACE).

Gerrard, Ysabel (2018). Beyond the hashtag: Circumventing content moderation on social media. *New Media & Society*, 20(12), 4492–4511. https://doi.org/10.1177/1461444818776611

Gillespie, T. (2018). *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions that Shape Social Media*. New Haven: Yale University Press.

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), 2053951719897945. https://doi.org/10.1177/2053951719897945

Haimson, O. L., Delmonaco, D., Nie, P., & Wegner, A. (2021). Disproportionate removals and differing content moderation experiences for conservative, transgender, and Black social media users: Marginalization and moderation gray areas. Proc. ACM Hum.-Comput. Interact., 5(CSCW2), Article 466, 35 pages. https://doi.org/10.1145/3479610

Halevy, A., Canton-Ferrer, C., Ma, H., Ozertem, U., Pantel, P., Saeidi, M., Silvestri, F., & Stoyanov, V. (2022). Preserving Integrity in Online Social Networks. *Communications of the ACM*, 65(2), 92–98. https://doi.org/10.1145/3462671

Imagga Blog (2024). Types of content moderation: Automated, pre-moderation, post-moderation, reactive, distributed. Imagga. بازیابی شده در ۲۴ آوریل ۲۰۲۵ از https://imagga.com

Lagren, E,. (2023). Artificial intelligence as a tool in social media content moderation.

Bachelor's Thesis. University of Jyväskylä.

Jhaver, S., Zhang, A. Q., Chen, Q. Z., Natarajan, N., Wang, R., & Zhang, A. X. (2023). Personalizing content moderation on social media: User perspectives on moderation choices, interface design, and labor. Proceedings of the ACM on Human-Computer Interaction, 7(CSCW2), 1–33.

Llansó, E. J. (2020). No amount of "AI" in content moderation will solve filtering's prior-restraint problem. *Big Data & Society*, 7(1), 205395172092068.

Lazer, D. M., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., ... & Zittrain, J. L. (2018). The science of fake news. *Science*, 359(6380), 1094–1096.

Roberts, S. T. (2017). *Content moderation*. ISBN 978-3-319-32010-6. https://escholarship.org/uc/item/7371c1hf

Roberts, S. T. (2019). *Behind the Screen: Content Moderation in the Shadows of Social Media*. Yale University Press.

Sawyer, J. (2024). Legal and Ethical Strategies for AI in Content Moderation: Safeguarding Free Speech and Privacy Rights. October 2024. https://doi.org/10.13140/RG.2.2.35076.33928

Santa Clara Principles (2018). The Santa Clara Principles: On Transparency and Accountability in Content Moderation. https://santaclaraprinciples.org/

Seering, J. (2020). Reconsidering self-moderation. Proceedings of the ACM on Human-Computer Interaction, 4(CSCW2), 1–28. https://doi.org/10.1145/3415178

Veglis, A. (2014). Moderation Techniques for Social Media Content. In Social Computing and Social Media. SCSM 2014. Lecture Notes in Computer Science, vol 8531. Cham: Springer. https://doi.org/10.1007/978-3-319-07632-4_13

WeChat Training Institute (2020). WeChat surveillance explained. WeChatTraining.com. https://wechattraining.com/newsfeed/wechat-surveillance-explained/