Journal of
International Criminal Law
www.iclj.ir

Volume 2, Issue 1, 2024

Criminal Liability of Social Platforms in Promoting and Facilitating "Hate Speech" Leading to Ethnic Violence: A Comparative Study of International Court Jurisprudence and Digital Platform Principles

Mahmoud Darvishtorabi [1*]

1. Assistant Professor, Department of Law, Payame Noor University, Tehran, Iran.

ARTICLE INFORMATION

Type of Article: ***Original Research***
Pages: ***93-105***

Corresponding Author's Info
ORCID: *0000-0002-8819-3024*
TELL: *+989113530796*
Email: *darvishtorabi@pnu.ac.ir*

Article history:
Received: *23 Jan 2024*
Revised: *09 Feb 2024*
Accepted: *08 Mar 2024*
Published online: *20 Mar 2024*

Keywords:
Hate Speech, International Criminal Law, Social Platforms, Recommendation Algorithms, International Criminal Court.

ABSTRACT

The transformation of communication mechanisms and the expansion of interactive social networks have altered the nature of reproducing hate speech, turning the role of recommendation systems in exacerbating ethnic violence into a serious challenge within international criminal law. Aiming to assess the feasibility of attributing criminal liability to digital platform operators, this descriptive-analytical and comparative study examines the jurisprudence of international courts, human rights instruments, and the provisions of the Rome Statute of the International Criminal Court. The findings indicate that the structure of content distribution within algorithmic environments differs fundamentally from the prior editorial control of traditional media, and deficiencies or weaknesses in technical message moderation cannot be deemed a basis for criminal prosecution without proving specific constituent elements of the crime. Based on the results of this study, the realization of transnational criminal liability for natural persons operating platforms—in addition to obstacles arising from the Court's jurisdiction solely over individuals and the principle of legality of crimes and punishments (nullum crimen sine lege)—requires establishing actual knowledge, specific intent (dolus specialis), and a direct causal link between algorithmic performance and committed atrocities. Therefore, managerial errors, organizational leniency, or breaches of corporate standards and regulatory rules possess a distinct nature from criminal facilitation or aiding and abetting international crimes; thus, until criminal intent is established, addressing these challenges largely remains within the realms of civil liability and administrative regulatory systems.

This is an open access article under the CC BY license. © 2024 The Authors.

How to Cite This Article: Darvishtorabi, M (2024). "Criminal Liability of Social Platforms in Promoting and Facilitating "Hate Speech" Leading to Ethnic Violence: A Comparative Study of International Court Jurisprudence and Digital Platform Principles". *Journal of Comparative Criminal Jurisprudence*, 2(1): 93-105.

فصلنامه حقوق جزای بین‌الملل

www.iclj.ir

دوره دوم، شماره اول، بهار ۱۴۰۳

مسؤولیت کیفری پلتفرم‌های اجتماعی در ترویج و تسهیل «سخنان نفرت‌پراکن» منجر به خشونت‌های قومی: مطالعه تطبیقی رویه دادگاه‌های بین‌المللی و اصول پلتفرم‌های دیجیتال

محمود درویش‌ترابی [۱*]

۱. استادیار گروه حقوق، دانشگاه پیام نور، تهران، ایران.

اطلاعات مقاله

نوع مقاله: پژوهشی

صفحات: ۱۰۵-۹۳

اطلاعات نویسنده مسؤول

کد ارکید: ۰۰۰۰-۰۰۰۲-۸۸۱۹-۳۰۲۴

تلفن: ۹۸۹۱۱۳۵۳۰۷۹۶+

ایمیل: darvishtorabi@pnu.ac.ir

سابقه مقاله:

تاریخ دریافت: ۱۴۰۲/۱۱/۰۳

تاریخ ویرایش: ۱۴۰۲/۱۱/۲۰

تاریخ پذیرش: ۱۴۰۲/۱۲/۱۸

تاریخ انتشار: ۱۴۰۳/۰۱/۰۱

واژگان کلیدی:

سخ نان نفرت‌پراکن، م سؤولا یت کیفری بین‌المللی، پلتفرم‌های اجتماعی، الگوریتم‌های توصیه‌گر، دیوان کیفری بین‌المللی.

چکیده

تحول سازوکارهای ارتباطی و گسترش شبکه‌های اجتماعی تعاملی، ماهیت بازتولید سخنان نفرت‌پراکن را دگرگون ساخته و نقش سامانه‌های توصیه‌گر را در تشدید خشونت‌های قومی به چالشی جدی در حقوق کیفری بین‌المللی بدل کرده است. پژوهش حاضر با هدف سنجش قابلیت انتساب مسؤولیت کیفری به گردانندگان پلتفرم‌های دیجیتال و با روش توصیفی–تحلیلی و تطبیقی، به واکاوی رویه محاکم بین‌المللی، اسناد حقوق بشری و موازین اساسنامه دیوان کیفری بین‌المللی می‌پردازد. یافته‌ها نشان می‌دهد که ساختار توزیع محتوا در بستر الگوریتم‌ها با نظارت پیشینی رسانه‌های سنتی تفاوت بنیادین دارد و نارسایی یا ضعف در پایش فنی پیام‌ها را نمی‌توان بدون اثبات عناصر اختصاصی جرم، مبنای تعقیب کیفری قلمداد کرد. بر پایه نتایج این مطالعه، تحقق مسؤولیت کیفری فراملی اشخاص حقیقی اداره‌کننده پلتفرم‌ها، افزون بر موانع ناشی از انحصار صلاحیت دیوان بر افراد و اصل قانونی بودن جرائم، مستلزم احراز آگاهی واقعی، سوءنیت خاص و اثبات رابطه علیت مستقیم میان عملکرد الگوریتم و جنایات ارتکابی است. نوآوری پژوهش، ارائه مدل آستانه‌ای چهار مرحله‌ای برای ارزیابی قابلیت تعقیب کیفری مدیران پلتفرم‌هاست که به ترتیب بر ماهیت و آستانه مجرمانه محتوا، آگاهی مستند شخص تصمیم‌گیرنده، نقش علی تقویت الگوریتمی در وقوع جرم و امتناع عامدانه از مهار محتوا با وجود کنترل مؤثر تمرکز دارد. این مدل، برخلاف تسری مستقیم معیارهای مسؤولیت رسانه‌های سنتی، میان قصور فنی و مشارکت کیفری قابل انتساب به اشخاص حقیقی تمایز می‌گذارد. از این‌رو، خطاهای مدیریتی، تسامح سازمانی یا نقض ضوابط شرکتی و مقررات تنظیم‌گرانه، ماهیتی متمایز از تسهیل مجرمانه یا معاونت در جرائم بین‌المللی دارند و تا زمان عدم احراز قصد مجرمانه، پاسخ‌گویی به این چالش‌ها عمدتا در قلمرو مسؤولیت مدنی و نظامات حقوقی اداری باقی می‌ماند.

خوانندگان این مجله، اجازه توزیع، ترکیب مجدد، تغییر جزئی و کار روی مقاله حاضر به‌صورت غیرتجاری را دارند.

© تمامی حقوق انتشار این مقاله، متعلق به نویسنده می‌باشد.

مقدمه

دگرگونی الگوهای ارتباطی در سده بیست‌ویکم، شیوه تولید، انتشار و دریافت گفتارهای خصمانه را به‌طور بنیادین تغییر داده است. در دوره‌ای که رسانه‌های جمعی سنتی، مانند رادیو و مطبوعات، ابزار اصلی انتقال پیام به شمار می‌رفتند، شمار محدودی از سردبیران و مدیران بر انتخاب و انتشار محتوا نظارت داشتند. این تمرکز نسبی، امکان شناسایی تصمیم‌گیرندگان و بررسی نقش آنان در بازتاب یا گسترش پیام‌های نفرت‌آمیز را فراهم می‌کرد. در مقابل، شبکه‌های اجتماعی ساختار ارتباطی پراکنده و تعاملی دارند؛ کاربران نه‌تنها دریافت‌کننده پیام، بلکه بازنشرکننده و تولیدکننده آن نیز هستند و یک پیام می‌تواند در زمانی کوتاه از مرزهای جغرافیایی و زبانی عبور کند. بنابراین، بررسی نقش این شبکه‌ها در خشونت‌های قومی مستلزم توجه هم‌زمان به رفتار کاربران، شیوه طراحی بسترها و تصمیم‌های مدیریتی درباره انتشار و برجسته‌سازی محتواست.

اهمیت این دگرگونی تنها به‌سرعت و گستره انتشار محدود نمی‌شود. سامانه‌های توصیه‌گر با انتخاب و رتبه‌بندی محتوا، بر آنچه کاربران می‌بینند اثر می‌گذارند. هنگامی‌که معیارهای افزایش مشارکت و ماندگاری مخاطب به برجسته‌شدن پیام‌های تند، تحریک‌آمیز یا هویت‌ستیز بینجامد، نقش بستر از میزبانی منفعلانه محتوا فراتر می‌رود و به چگونگی توزیع و تقویت آن پیوند می‌خورد. بااین‌حال، وجود چنین اثری به‌تنهایی برای اثبات مسؤولیت کیفری کافی نیست. باید روشن شود چه شخصی درباره طراحی یا اجرای سازوکارهای انتشار تصمیم گرفته، از پیامدهای آنچه اندازه آگاهی داشته، چه میزان کنترل بر آن داشته و رفتار او چگونه با وقوع یا گسترش خشونت ارتباط یافته است. همچنین، ناتوانی در حذف همه محتواهای زیان‌بار را نباید بی‌درنگ با تسهیل آگاهانه جرم یکسان دانست؛ زیرا چنین برداشتی می‌تواند میان خطای مدیریتی، کاستی در نظارت و مشارکت مجرمانه تمایز نگذارد.

بر این اساس، پرسش اصلی پژوهش آن است که آیا ترویج یا تقویت الگوریتمی سخنان نفرت‌پراکن، یا خودداری از مهار آن‌ها پس از دریافت هشدارهای معتبر، در شرایطی خاص می‌تواند در حقوق کیفری بین‌المللی وصف مجرمانه پیدا کند. پرسش‌های فرعی نیز ناظر بر معیارهای تشخیص تحریک مجرمانه از بیان زیان‌بار اما غیر مجرمانه، شرایط انتساب رفتار پلتفرم به اشخاص حقیقی اداره‌کننده آن، و نقش قصد، آگاهی، کنترل و پیوند میان رفتار متهم و جرم هستند. فرضیه پژوهش این است که صرف میزبانی یا انتشار محتوای کاربران، یا نقض تعهدات داخلی پلتفرم، به‌تنهایی مسؤولیت کیفری بین‌المللی ایجاد نمی‌کند؛ اما تصمیم آگاهانه و هدفمند شخص حقیقی برای تقویت محتوا یا تسهیل انتشار آن، در صورت احراز عناصر جرم و پیوند لازم با وقوع آن، می‌تواند زمینه بررسی مسؤولیت کیفری را فراهم سازد.

روش پژوهش، تحلیلی و تطبیقی است و با کنار هم نهادن رویه محاکم کیفری بین‌المللی و منطقه‌ای، اسناد حقوق بشری، قواعد ناظر بر آزادی بیان و معیارهای تنظیم‌گری بسترهای دیجیتال، به ارزیابی حدود این مسؤولیت می‌پردازد. در این چارچوب، مفاهیم حقوق حاکمیت داده و فناوری برای روشن کردن نقش تصمیم‌های طراحی، پردازش و توزیع محتوا به کار گرفته می‌شوند، بی‌آنکه جایگزین ارکان شناخته‌شده مسؤولیت کیفری شوند. پژوهش به جرائم بین‌المللی مندرج در اساسنامه رم، به‌ویژه نسل‌کشی و جنایات علیه بشریت، محدود است و سخنان نفرت‌پراکن را تنها هنگامی در کانون تحلیل قرار می‌دهد که نسبت آن با ارتکاب یا تسهیل این جرائم بررسی‌پذیر باشد. همچنین، مسؤولیت کیفری اشخاص حقیقی از مسؤولیت مدنی، تنظیم‌گرانه یا پاسخ‌گویی شرکتی پلتفرم‌ها جدا نگه‌داشته می‌شود؛ زیرا هر یک مبانی، اهداف و معیارهای اثبات متفاوتی دارند.

نوآوری پژوهش در تمرکز بر مرز میان نقش فنی پلتفرم و تصمیم انسانی قابل انتساب به فرد مشخص است. به‌جای فرض مسؤولیت خودکار شرکت‌ها یا مدیران، تحقیق می‌کوشد معیاری برای سنجش رفتار، میزان آگاهی، اختیار مؤثر و سهم اشخاص در تقویت یا مهار محتوا ارائه کند. بدین‌ترتیب، مقاله امکان بهره‌گیری از دستاوردهای رویه قضایی درباره نقش رسانه‌ها را بررسی می‌کند، اما تفاوت ساختاری رسانه‌های سنتی و شبکه‌های الگوریتم‌محور را نیز در نظر می‌گیرد. مقصود

نهایی، روشن‌کردن حدود پاسخ کیفری در برابر خشونت‌های قومی مرتبط با فضای دیجیتال است؛ پاسخی که از یک‌سو در برابر تسهیل آگاهانه جرائم بین‌المللی بی‌تفاوت نماند و از سوی دیگر، آزادی بیان و اصل قانونی بودن جرم را با توسعه بی‌ضابطه مسؤولیت کیفری محدود نکند.

۱- چارچوب مفهومی و استانداردهای حقوق بشری تفکیک بیان مجاز از رفتار مجرمانه

برای ورود به بحث مسؤولیت کیفری، نخست باید مرز باریک میان ابراز عقاید و رفتارهای مجرمانه را در اسناد بین‌المللی از یکدیگر تفکیک کرد تا مشخص شود چه نوع از گفتارها از دایره حمایت آزادی بیان خارج و مستوجب تعقیب می‌شوند.

۱-۱- مرزبندی مفهومی در اسناد بین‌المللی

سازوکار حقوق بین‌الملل بشر در تفکیک ابراز عقیده مشروع از رفتارهای نامشروع، توازنی ظریف میان اصل آزادی بیان و ضرورت مهار پیام‌های آسیب‌رسان برقرار ساخته است. ماده نوزدهم میثاق بین‌المللی حقوق مدنی و سیاسی، آزادی بیان را قاعده عام قرار داده و تحدید آن را تنها در چارچوب شروط مضیق قانونی مجاز می‌شمارد؛ حال آنکه بند دوم ماده بیستم همان سند، تعهدی بنیادین بر دوش دولت‌ها نهاده تا هرگونه ترغیب به نفرت ملی، نژادی یا مذهبی را که محرک تبعیض، خصومت یا خشونت باشد، منع کنند (Fino, 2023: 195). با وجود این تعهد الزامی، عدم ارائه تعریفی جامع و مانع از سخنان نفرت‌پراکن در حقوق معاهداتی، چالش‌های فراوانی در تبیین قلمرو مداخله پدید آورده و ابهام در تعیین مرز انحراف از آزادی به رفتار ممنوع را تشدید کرده است (Fino, 2020: 38).

نقطه عطف این تمایزگذاری مفهومی، در جداسازی سخنان نفرت‌پراکن با ضمانت‌اجراهای تنظیمی از تحریک مستقیم و علنی به ارتکاب نسل‌کشی متجلی می‌شود. بند سوم ماده سوم کنوانسیون پیشگیری و مجازات جنایت نسل‌کشی، تنها صورتی از گفتار را به‌عنوان جرمی مستقل و قائم‌به‌ذات مجرم‌انگاری می‌کند که حامل ترغیبی صریح، بی‌واسطه و غیرمبهم به نابودی فیزیکی تمام یا بخشی از یک گروه حمایت‌شده باشد (de Souza Dias, 2024: 402). بنابراین، اگرچه سخنان نفرت‌پراکن فضای روانی مستعد برای خصومت جمعی را فراهم می‌سازند، اما تا زمانی که قصد خاص امحای گروهی و صراحت بیانی محقق نگردد، ماهیتی کیفری در سطح جرائم بین‌المللی نمی‌یابند و پاسخ‌گویی به آن‌ها در سطوح هنجاری پایین‌تری سامان می‌یابد.

۱-۲- آزمون آستانه‌ای شش‌گانه طرح اقدام رباط

طرح اقدام رباط با طراحی یک آزمون آستانه‌ای شش‌گانه، چارچوبی منسجم برای تفکیک بیان مجاز از سخنان غیرقانونی مستوجب ضمانت‌اجراهای کیفری فراهم آورده است. این الگو بر ارزیابی متوازن شش متغیر بنیادین تکیه دارد: زمینه سیاسی و اجتماعی حاکم بر پیام، موقعیت و میزان نفوذ گوینده، قصد صریح مرتکب برای تحریک، محتوا و لحن کلام، دامنه دسترسی به پیام، و درجه احتمال وقوع آسیب قریب‌الوقوع (Mahdi & Rinwigati, 2024: 328). هدف بنیادین این شاخص‌ها، پیشگیری از محدودسازی سلیقه‌ای ابراز عقیده و تضمین ضرورت و تناسب در مداخلات محدودکننده حقوقی است (Salama & Wiener, 2025: 7).

با این حال، انتقال این آزمون سنتی به زیست‌بوم پلتفرم‌های دیجیتال با چالش‌های فنی و ساختاری عمیقی روبرو شده است. در شبکه‌های اجتماعی، پدیده بازنشر فراگیر و سازوکارهای محاسباتی توصیه‌گر، معنای ایستای متغیرهایی چون موقعیت گوینده و دامنه نفوذ را بازتعریف کرده‌اند؛ به‌طوری‌که پیامی تولیدشده توسط کاربری ناشناس، در صورت برخورداری از ماهیت برانگیزاننده، از طریق بازخورد تعاملی سامانه به‌سرعت شتاب می‌گیرد و به گستره وسیعی از مخاطبان عرضه می‌شود (Hatano, 2023: 135). چنین سرعتی در انتشار و بازتولید برخط، سنجش احتمال آسیب قریب‌الوقوع و جداسازی عاملیت پلتفرم از نیت کاربر را در عمل دشوار می‌سازد. از این‌رو، سنجه‌های رباط بدون بازتعریف متغیرهای فناوری‌پایه و در نظر گرفتن پویایی‌های شبکه‌ای، قادر به ارزیابی کامل نقش میانجی‌گر سامانه‌های نوین در بازتولید خشونت‌های قومی نخواهند بود.

۲- مطالعه تطبیقی رویه قضایی محاکم بین‌المللی و منطقه‌ای

برای سنجش حدود مسؤولیت در فضای دیجیتال، بررسی رویه دادگاه‌هایی که با نقش رسانه‌ها و واسطه‌های انتشار در

گسترش نفرت روبه‌رو شده‌اند، اهمیت ویژه‌ای دارد. آرای دادگاه کیفری روآندا و دادگاه اروپایی حقوق بشر، با وجود تفاوت در موضوع و مبنای مسؤولیت، زمینه‌ای فراهم می‌کنند تا معیارهای انتساب رفتار مجرمانه و حدود تعمیم آن به پلتفرم‌های اجتماعی روشن شود.

۲-۱- رویه دادگاه کیفری بین‌المللی برای روآندا: پیشینه سرنوشت‌ساز رسانه‌ها

رویه‌های قضایی دادگاه کیفری بین‌المللی برای روآندا نقطه عطفی در شکل‌گیری دکترین حقوقی ناظر بر مسؤولیت ارکان رسانه‌ای در جرائم بین‌المللی فراهم آورده است. در پرونده آکایسو، برای نخستین بار چارچوب مفهومی روشنی از عناصر مادی و روانی جرم تحریک مستقیم و علنی به نسل‌کشی ترسیم شد و تأثیر کلام در ترغیب توده‌ها به خشونت ساختاری مورد شناسایی قرار گرفت (Gordon, 2017: 112). توسعه این رویه در پرونده مدیران رسانه‌ای رادیو و نشریه کانگورا، گامی فراتر در تبیین مسؤولیت کیفری عوامل پشت صحنه انتشار پیام گذاشت؛ جایی که کنترل ابزار نشر و مدیریت محتوای نفرت‌بار به‌عنوان مصداقی از مشارکت مستقیم در فجایع جمعی ارزیابی شد (Mosely, 2026: 15).

با این وجود، تبیین دقیق مرز میان مالکیت ساختار انتشار و تحقق اراده مجرمانه در پرونده سایمون بیکیندی با پیچیدگی‌های بیشتری روبه‌رو شد، زیرا تفکیک تأثیر بافتاری سرودها و پیام‌های فرهنگی بر مخاطبان از نیت درونی و آگاهانه سازنده یا توزیع‌کننده، نیازمند احراز ارتباطی مستقیم میان محتوا و وقوع جنایت عینی بود (Parker, 2015: 88). این تحولات تاریخی نشان می‌دهد که انتساب مسؤولیت کیفری به واسطه‌های نشر، همواره نیازمند عبور از صافی سخت‌گیرانه احراز قصد مجرمانه و اثبات نقش بی‌واسطه در تحریک خشونت است؛ معیاری که تسری آن به شبکه‌های اجتماعی الگوریتم‌محور و تصمیم‌گیرندگان پشت پلتفرم‌ها با موانع جدی تئوریک مواجه است.

۲-۲- آموزه‌های دادگاه اروپایی حقوق بشر پیرامون واسطه‌ها و فضای مجازی

دیوان اروپایی حقوق بشر با تدوین رویه‌های کلیدی پیرامون مسؤولیت واسطه‌های دیجیتال، چالش‌های خطیر نظارت بر محتوای تولیدشده توسط کاربران را به تصویر کشیده است. در پرونده درگاه اینترنتی استونی، مسؤولیت مدنی و نظارتی پلتفرم در قبال دیدگاه‌های نفرت‌آمیز اشخاص ثالث مورد تأیید قرار گرفت و معیار ضرورت اقدام فوری به محض کسب اطلاع از محتوای مجرمانه به‌عنوان شاخصی برای ارزیابی قصور واسطه‌ها مطرح شد (FitzGerald, 2025: 52). تسری این رویکرد سخت‌گیرانه در پرونده دارنده حساب فرانسوی، ابعاد تازه‌ای از بار مسؤولیت کیفری اشخاص حقیقی را به دلیل عدم پالایش نظرات توهین‌آمیز در فضای مجازی برجسته ساخت و مرزهای شکننده مسؤولیت مدیران زیرساخت‌های کلان را به چالش کشید (Husovec et al., 2024: 390).

با این حال، تعمیم این الگوهای قضایی منطقه‌ای به ساحت حقوق کیفری بین‌المللی و پلتفرم‌های کلان فناوری با موانع ساختاری روبروست، زیرا استانداردهای ارزیابی در این محاکم عمدتا بر پایه مسؤولیت مدنی، نظارت پسینی و تخلفات شبه‌جرم بنا شده‌اند و نمی‌توانند ارکان روانی سنگین و عناصر اختصاصی جرائم بین‌المللی را توجیه کنند (Enarsson, 2024: 394). این ناهماهنگی نشان می‌دهد که رویکردهای حمایتی محاکم حقوق بشری در کنترل مجاری نشر، ظرفیت لازم برای انتساب مسؤولیت کیفری فراملی به طراحان الگوریتم‌ها را دارا نیستند.

۲-۳- مقایسه تطبیقی ساختار رسانه‌های سنتی با پلتفرم‌های دیجیتال الگوریتم‌محور

ارزیابی تحولات ساختاری در عرصه نشر اطلاعات نشان می‌دهد که انتقال از رسانه‌های سنتی به بسترهای دیجیتال، بنیادین‌ترین تغییر را در الگوهای اعمال کنترل بر محتوا پدید آورده است. در رسانه‌های سنتی، نظارت پیشینی سردبیری به‌عنوان سد دفاعی محکمی در برابر انتشار محتوای مجرمانه عمل می‌کرد، درحالی‌که در محیط پلتفرم‌های برخط، این

الگوی سنتی جای خود را به مدیریت پسینی و خودکار داده است (Schroeder, 2024: 12).

این دگرگونی ساختاری باعث می‌شود که پلتفرم‌های فناوری پیش از انتشار سراسری پیام‌ها مداخله‌ای نداشته باشند و پالایش محتوا به‌شدت به ابزارهای خودکار و گزارش‌های کاربران وابسته شود (Karlsson et al., 2022: 198). همین تفاوت بنیادین میان کنترل پیشینی و مدیریت پسینی است که تعمیم مستقیم احکام تاریخی دادگاه‌های نورنبرگ و رواندا به واسطه‌های فناورانه امروز را با بن‌بست‌های حقوقی جدی روبه‌رو می‌سازد (Park & Rohatgi, 2024: 15). درحالی‌که در پرونده‌های سنتی، مدیران رسانه با آگاهی کامل و به‌صورت مستقیم بر انتشار متن نظارت داشتند، پلتفرم‌های نوین با حجم انبوهی از داده‌های الگوریتمی روبه‌رو هستند که بازتولید خودکار آن‌ها مانع از تحقق ارکان سنتی مسؤولیت کیفری می‌شود. این شکاف ساختاری اثبات می‌کند که استانداردهای پیشین فاقد کارایی لازم برای انطباق با پویایی شبکه‌های دیجیتال مدرن هستند.

۳- اصول، سازوکارهای تعدیل محتوا و مسؤولیت‌های پلتفرم‌های دیجیتال

برای سنجش نقش پلتفرم‌ها در مهار یا تشدید خشونت‌های قومی، نخست باید دید این شرکت‌ها چه سازوکارهایی برای پالایش محتوا طراحی کرده‌اند و تا چه اندازه به تکالیف حقوق بشری و مقررات فراملی پایبند بوده‌اند. بااین‌حال، چالش بنیادین این بخش در تفکیک دقیق میان نقض سیاست‌های داخلی یا استانداردهای شرکتی از یک‌سو، و پیدایش رفتاری با قابلیت تعقیب کیفری بین‌المللی از سوی دیگر نهفته است.

۳-۱- استانداردهای رفتاری و سیاست‌های جامعه

بررسی ساختار کلان تنظیم‌گری در شبکه‌های اجتماعی بزرگ نشان می‌دهد که سیاست‌های موسوم به استانداردهای جامعه، به‌عنوان خط مقدم مواجهه با محتوای مجرمانه و سخنان نفرت‌انگیز عمل می‌کنند (Schaffner et al., 2024: 6). مطالعه تطبیقی ضوابط حاکم بر پلتفرم‌های کلان مانند متا، ایکس و یوتیوب در مواجهه با خشونت‌های قومی و رفتارهای تروریستی بیانگر وجود افتراق و ناهماهنگی در الگوهای اعمال محدودیت است، به‌طوری‌که رویکردهای متفاوتی در تعریف و تشخیص مصادیق محتوای آسیب‌زا اتخاذ شده است (Flynn et al., 2025: 12). این ضوابط همواره با ابهام مفهومی در اصطلاحات کلیدی، گستره وسیع استثناهای مرتبط با منافع عمومی و خطاهای ساختاری در غربالگری خودکار توسط سامانه‌های هوش مصنوعی دست به گریبان هستند (Narrain, 2026: 345). وجود چنین خلأهایی موجب می‌شود که مرز میان بیان مشروع و محتوای مستوجب مسدودسازی مخدوش گردد و ابزارهای تعدیل محتوا در صیانت از گروه‌های آسیب‌پذیر ناکارآمد بمانند. همین ابهام‌آلود بودن و عدم قطعیت در معیارهای درونی شرکت‌ها نشان می‌دهد که اتکا به سیاست‌های داخلی پلتفرم‌ها به‌هیچ‌وجه نمی‌تواند جایگزین استانداردهای دقیق حقوق کیفری در انتساب مسؤولیت بین‌المللی به اداره‌کنندگان آن‌ها شود.

۳-۲- اسناد تنظیم‌گرانه و چارچوب‌های حقوق بشری شرکتی

سیر تحولات تنظیم‌گری در عرصه فناوری نیازمند انطباق فعالیت شرکت‌های بزرگ با تکالیف بنیادین حقوق بشری از طریق مکانیسم‌های سنجش و پایش ریسک است. اصول راهنمای سازمان ملل متحد درباره تجارت و حقوق بشر، شرکت‌های فناوری را مکلف می‌سازد تا با اجرای دقیق فرآیندهای ارزیابی ریسک حقوق بشری، از بازتولید پیامدهای مخرب محتوایی جلوگیری کنند (Office of the United Nations High Commissioner for Human Rights, 2025: 14). با این وجود، ماهیت داوطلبانه این اصول و تکیه بر تعهدات شرکتی با محدودیت‌های ذاتی مواجه است و در غیاب سازوکارهای الزام‌آور، کارایی لازم را در مهار پیامدهای مرگبار سخنان نفرت‌انگیز از دست می‌دهد (Álvarez-Ugarte, 2025: 618). به همین دلیل، قوانین نوینی مانند قانون خدمات دیجیتال اتحادیه اروپا با تمرکز بر پلتفرم‌های بسیار بزرگ، تکالیف ویژه‌ای را در زمینه مدیریت ریسک‌های سیستمیک، ارزیابی الگوریتمی و شفاف‌سازی خطرات انتشار محتوا وضع کرده‌اند (Griffin, 2025: 235). این چارچوب‌های تنظیمی جدید تلاش می‌کنند تا شرکت‌ها را از جایگاه صرفا تجاری خارج کرده و در قبال آثار کلان انتشار نفرت در جامعه مسؤول بدانند. با این وجود، میان این الزامات

نظارتی شرکتی و ارکان سخت‌گیرانه حقوق کیفری بین‌المللی فاصله عمیقی وجود دارد که نشان می‌دهد ضمانت‌اجراهای تنظیم‌گری به‌تنهایی نمی‌توانند به‌عنوان مبنایی برای تعقیب کیفری بین‌المللی مدیران فناوری مورد استناد قرار گیرند.

۳-۳- تفکیک ساختاری

رعایت دقیق مرزهای میان نظام‌های حقوقی گوناگون مستلزم تفکیک میان نقض مقررات داخلی پلتفرم‌ها و وقوع مسؤولیت کیفری بین‌المللی است (Di Stefano, 2025: 524). تخطی از سیاست‌های شرکتی یا استانداردهای رفتاری معمولا به ضمانت‌اجراهای درون‌سیستمی، مسدودسازی محتوا یا جریمه‌های مدنی و تنظیم‌گرانه منتهی می‌شود، درحالی‌که تحقق مسؤولیت کیفری بین‌المللی نیازمند احراز عناصر سخت‌گیرانه حقوق جزای بین‌الملل و ارتکاب جنایات شدید علیه بشریت است (van Sliedregt, 2025: 20). ابزارهای نوین تنظیم‌گری و نظارت بر الگوریتم‌های توصیه‌گر تلاش می‌کنند تا از بازتولید محتوای سمی جلوگیری کنند، اما این سازوکارها ذاتا ماهیت بازدارنده و مدیریتی دارند و نباید با سازوکارهای تعقیب کیفری اشخاص حقیقی اشتباه گرفته شوند (Kohl, 2024: 312). در واقع، خلط میان این دو سطح از مسؤولیت موجب انحراف در نظام دادرسی و تعمیم ناروای مفاهیم تجاری به حقوق کیفری می‌شود. پلتفرم‌های دیجیتال ممکن است با نقض تعهدات تنظیمی خود مستوجب مسؤولیت حقوقی یا مدنی شرکت‌ها گردند، اما این امر به‌خودی‌خود نمی‌تواند به‌عنوان دلیلی بر انتساب مستقیم جنایات بین‌المللی به مدیران یا تصمیم‌گیرندگان آن‌ها تلقی شود، مگر آنکه ارکان دقیق مادی و روانی جرم به‌طور مستقل به اثبات برسد.

۴- موانع حقوقی و ارکان انتساب مسؤولیت کیفری به اشخاص اداره‌کننده پلتفرم

گذر از خطاهای سازمانی یا مدنی پلتفرم‌ها به‌سوی مسؤولیت کیفری بین‌المللی گردانندگان آن‌ها، مستلزم عبور از صافی‌های دقیق و سنتی حقوق جزا است. در این میان، انحصار صلاحیت محاکم بر اشخاص حقیقی، لزوم انطباق تصمیم‌های فنی یا مدیریتی با ارکان مادی و روانی مشارکت در جرائم بین‌المللی، و صیانت از اصل قانونی بودن جرم و مجازات، موانعی جدی در مسیر انتساب کیفری پدید می‌آورند که نیازمند موشکافی است.

۴-۱- صلاحیت شخصی و موضوعی در دیوان کیفری بین‌المللی

بررسی صلاحیت دیوان کیفری بین‌المللی نشان می‌دهد که نظام حقوق کیفری بین‌الملل بر اصل مسؤولیت اشخاص حقیقی بنا شده است و این نهاد قضایی فاقد صلاحیت برای تعقیب شرکت‌ها یا اشخاص حقوقی است (Tsilonis, 2024: 45). بر اساس مقررات اساسنامه رم، صلاحیت شخصی دیوان منحصرا انسان‌ها را در برمی‌گیرد و امکان انتساب مستقیم مسؤولیت کیفری به شرکت‌های تجاری وجود ندارد (van Baar, 2024: 995). این ساختار حقوقی موجب می‌شود که در مواجهه با تخلفات کلان پلتفرم‌های دیجیتال، شناسایی اشخاص قابل تعقیب با چالش‌های جدی روبه‌رو شود (Nieprzecka, 2024: 112).

در این میان، پرسش اساسی این است که آیا باید مدیران ارشد اجرایی، مهندسان طراح الگوریتم‌های توصیه‌گر و بیشینه‌ساز تعامل، یا مدیران بخش امنیت و تعدیل محتوا را به‌عنوان عاملان اصلی یا شرکای جرم تحت تعقیب قرار داد. تعیین دقیق این مرزها برای جلوگیری از مسؤولیت‌ناپذیری یا تعقیب‌های خودسرانه اهمیت حیاتی دارد. از یک‌سو، مدیران ارشد استراتژی کلان شرکت را هدایت می‌کنند و از سوی دیگر، طراحان الگوریتم نقش مستقیمی در بازنشر ویروسی و هدایت هدفمند محتوای نفرت‌انگیز دارند. با این حال، احراز عنصر روانی و رابطه علیت میان تصمیمات فنی این اشخاص و وقوع خشونت‌های قومی، نیازمند تفکیک دقیق سطوح تصمیم‌گیری سازمانی و انطباق آن با چارچوب‌های سخت‌گیرانه صلاحیت شخصی و موضوعی در نظام کیفری بین‌المللی است تا از خلط میان مسؤولیت شرکتی و مسؤولیت کیفری فردی جلوگیری شود.

۴-۲- رکن مادی و اشکال مشارکت در جرم

بررسی ارکان مادی جرم در بستر پلتفرم‌های دیجیتال مستلزم تحلیل دقیق نحوه انطباق رفتارهای فناورانه با الگوهای سنتی مشارکت در جرائم بین‌المللی است (Anand, 2024: 170).

تحریک مستقیم و علنی به نسل‌کشی زمانی در فضای مجازی تحقق می‌یابد که انتشار محتوای نفرت‌انگیز به‌صورت صریح و عمومی مخاطبان را به ارتکاب جنایت ترغیب کند، هرچند چالش اصلی در این حوزه تشخیص مرز میان اظهارنظر تند و تحریک مؤثر است (Zarmsky, 2024: 175). از سوی دیگر، معاونت و تسهیل ارتکاب جرم از طریق به‌کارگیری الگوریتم‌های بیشینه‌ساز تعامل رخ می‌دهد؛ جایی که طراحی عامدانه سیستم برای افزایش بازدید و بازنشر ویروسی محتوا، عملا به تقویت و گسترش کانون‌های خشونت یاری می‌رساند (Holvoet, 2022: 235). در کنار این موارد، مسؤولیت ناشی از ترک فعل و خودداری از مهار محتوا زمانی مطرح می‌شود که مدیران پلتفرم با وجود داشتن وظیفه حقوقی اقدام و توانایی فنی مؤثر برای پالایش محیط، عمدا از مداخله سر باز زنند. احراز این ارکان مادی نیازمند اثبات این واقعیت است که امتناع شرکت در حذف محتوا یا تغییر نکردن تنظیمات الگوریتمی، سهمی عینی و مؤثر در وقوع جنایات کلان قومی ایفا کرده است تا از این طریق بتوان رفتار فنی و تجاری اداره‌کنندگان را به‌عنوان یکی از اشکال مشارکت در جرائم بین‌المللی قلمداد نمود.

۳-۴- رکن روانی و بار اثبات

بررسی ارکان روانی در جرائم بین‌المللی مستلزم انطباق دقیق رفتار اشخاص با شروط مقرر در ماده ۳۰ اساسنامه رم است که بر احراز همزمان قصد و آگاهی دلالت دارد (McDermott, 2024: 115). در خصوص مدیران فناوری و طراحان الگوریتم، اثبات «سوءنیت ویژه» یا همان قصد خاص برای نابودی یک گروه قومی با چالش‌های بنیادینی روبه‌روست؛ زیرا فعالیت‌های تجاری و فنی شرکت‌ها معمولا با هدف بیشینه‌سازی سود و گسترش تعاملات کاربران انجام می‌شود و نه نابودی عامدانه گروهی خاص (Pisillo Mazzeschi & Carli, 2024: 435). از سوی دیگر، نظام حقوق کیفری بین‌المللی برای جرائمی همچون نسل‌کشی، احراز قصد صریح و آگاهی دقیق را الزامی می‌داند و به همین دلیل، استانداردهایی نظیر بی‌پروا بودن یا غفلت فاحش معمولا برای انتساب جنایات بین‌المللی به مدیران فناوری کفایت نمی‌کند (Savard, 2026: 1180). این فاصله میان انگیزه تجاری ارتقای الگوریتمی و سوءنیت مجرمانه ویژه، مانع اصلی در تعقیب کیفری مدیران پلتفرم‌هاست؛ مگر آنکه از طریق مستندات روشن اثبات شود که تصمیم‌گیرندگان با وجود آگاهی قطعی از وقوع خشونت‌های قومی، عامدا به بازتولید و ترویج محتوای نفرت‌آفرین ادامه داده‌اند تا از این مسیر، قصد مجرمانه آن‌ها در تحقق جنایت قابل احراز و انتساب باشد.

۴-۴- اصل قانونی بودن جرائم و مجازات‌ها

رعایت اصل قانونی بودن جرائم و مجازات‌ها به‌عنوان یکی از ارکان بنیادین دادرسی منصفانه، مانع از تسری خودسرانه عناوین مجرمانه به حوزه‌های نوظهور فناوری می‌شود (Davidson, 2017: 55). در نظام حقوق کیفری بین‌المللی، تفسیر مضیق قوانین جزایی ایجاب می‌کند که مفاهیمی چون نسل‌کشی و جنایت علیه بشریت به‌واسطه ویژگی‌های ساختاری خود، فراتر از مصادیق سنتی تعبیر نشوند (Akhavan, 2022: 50). چالش اصلی در این میان، انطباق تصمیم‌های فنی و تجاری شرکت‌های فناوری، نظیر بهینه‌سازی الگوریتم‌های توصیه‌گر و طراحی معماری نرم‌افزاری پلتفرم‌ها، بر ارکان عینی مصادیق بین‌المللی تسهیل جنایات است (Manley et al., 2023: 785). ازآنجاکه ساختار این تصمیم‌ها ریشه در اهداف تجاری و افزایش تعاملات کاربران دارد، تعمیم دادن آن‌ها به‌عنوان مصداق بارز ارتکاب یا تسهیل جنایات کلان، بدون وجود متن صریح قانونی و پیش‌بینی‌پذیری حقوقی، با موازین اصل قانونی بودن در تضاد آشکار قرار می‌گیرد. هرگونه تلاش برای مسؤولیت‌کیفری مدیران فناوری بر اساس این‌گونه تصمیم‌ها مستلزم وجود نص صریح است تا از گسترش تفسیرهای قضایی سلیقه‌ای جلوگیری شود.

۵- ارزیابی موردی و ارائه مدل تحلیلی تصمیم‌گیری قضایی

برای آزمودن عملی مباحث نظری پیشین، بررسی یک بستر عینی نظیر بحران میانمار می‌تواند شکاف میان قصور سیستمی و رفتار مجرمانه فردی را به‌خوبی آشکار کند. از دل این تجربه عینی و در پیوند با موازین حقوق کیفری، مدلی گام‌به‌گام ترسیم می‌شود تا معیاری سنجش‌پذیر برای ارزیابی قابلیت تعقیب قضایی تصمیم‌گیران پلتفرم‌ها در بحران‌های مشابه فراهم آورد.

۵-۱- واکاوی بحران میانمار (روهینگیا) به‌مثابه آزمون تجربی

واکاوی بحران میانمار و نقشه‌برداری از رخدادهای منتهی به خشونت علیه روهینگیا، بستر تجربی مهمی برای ارزیابی مسؤولیت پلتفرم‌های دیجیتال فراهم می‌سازد (Independent Investigative Mechanism for Myanmar, 2024: 45). یافته‌های نهادهای بین‌المللی گویای آن است که انتشار گسترده محتوای نفرت‌آفرین در شبکه‌های اجتماعی نقش مؤثری در تشدید تنش‌های قومی داشته است. با وجود این، میان دریافت هشدارهای متعدد از سوی نهادهای مدنی و تحقق عنصر آگاهی مجرمانه در سطح مدیریت ارشد، فاصله حقوقی معناداری وجود دارد (Amnesty International, 2022: 32). دریافت گزارش‌های هشداردهنده از سوی سازمان‌های حقوق بشری، لزوما به معنای احراز سوءنیت خاص یا قصد عامدانه برای تسهیل جنایت نیست و نمی‌تواند به تنهایی مبنای استواری برای تعقیب کیفری بین‌المللی مدیران باشد (Independent Investigative Mechanism for Myanmar, 2024: 72). تفکیک دقیق میان خطاهای ساختاری در مدیریت محتوا و مسؤولیت مبتنی بر همدستی در جنایات بین‌المللی، مستلزم احراز رابطه علیت مستقیم و تصمیم آگاهانه برای بهره‌برداری از بحران است (Tambu, 2026: 170). نادیده گرفتن این مرز باریک، خطر تبدیل خطاهای مدیریتی شرکت‌ها به مسؤولیت‌های سنگین کیفری را به همراه دارد که با اصول پذیرفته‌شده انتساب مسؤولیت در حقوق بین‌الملل سازگار نیست.

۵-۲- مدل آستانه‌ای چهار مرحله‌ای برای ارزیابی قابلیت تعقیب کیفری

برای سنجش دقیق قابلیت تعقیب کیفری اشخاص حقیقی اداره‌کننده شبکه‌های اجتماعی، بهره‌گیری از یک چارچوب نظام‌مند چهار مرحله‌ای ضروری به نظر می‌رسد. در گام نخست، بررسی درجه تهدید واقعی محتوا و انطباق مفاد آن با جنایات مشمول صلاحیت دیوان کیفری بین‌المللی اهمیت اساسی دارد (Office of the Prosecutor of the International Criminal Court, 2025: 12). گام دوم بر اثبات آگاهی اختصاصی و مستند شخص متهم تمرکز می‌کند تا مشخص شود آیا هشدارهای رسمی نهادهای بین‌المللی با تصمیم آگاهانه برای استمرار روند موجود همراه بوده است یا خیر (Anand, 2024: 170). در گام سوم، میزان تأثیرگذاری مستقیم الگوریتم‌ها و طراحی‌های نرم‌افزاری در بازنشر غیرعادی و فزاینده پیام‌های نفرت‌آفرین اندازه‌گیری می‌شود تا نقش علی این ابزارها در وقوع جرم احراز گردد (Coco, 2026: 242). سرانجام در گام چهارم، امتناع عامدانه از مهار محتوا با وجود داشتن ابزار و کنترل فنی مؤثر مورد ارزیابی قرار می‌گیرد تا از این طریق، مرز میان خطاهای سیستمی و مشارکت کیفری محرز شود (Office of the Prosecutor of the International Criminal Court, 2025: 18). این ساختار منطقی مانع از تعمیم نادرست مسؤولیت‌ها می‌شود و چارچوب روشنی برای ارزیابی حقوقی این دست از پرونده‌ها به دست می‌دهد.

۵-۳- مسیرهای جایگزین و مکمل

دشواری‌های حقوقی و موانع ساختاری در تعقیب کیفری بین‌المللی مدیران فناوری، ضرورت طراحی و تقویت مسیرهای جایگزین و مکمل را برای مقابله با بازتولید نفرت در فضای مجازی دوچندان می‌سازد (Hamilton, 2022: 1375). در غیاب سازوکارهای مستقیم انتساب کیفری، پاسخ‌های تنظیمی و نظارت‌های نهادی به ابزارهای کارآمدی برای الزام شرکت‌ها به رعایت استانداردهای ایمنی مبدل می‌شوند (Santos et al., 2025: 740). از سوی دیگر، اعمال ضمانت‌اجراهای شبه‌کیفری، از جمله جرائم مالی سنگین و محدودیت‌های بازرگانی، فشار اقتصادی مؤثرتری بر پلتفرم‌ها وارد می‌کند تا از بازنشر محتوای مخرب جلوگیری نمایند (Nave, 2026: 55). تقویت مسؤولیت مدنی شرکتی و ایجاد سازوکارهای جبران خسارت برای قربانیان خشونت‌های قومی، مسیر دیگری است که امکان پاسخگویی مدنی پلتفرم‌ها را بدون نیاز به اثبات سوءنیت ویژه فراهم می‌آورد (Hamilton, 2022: 1400). تلفیق این سازوکارهای تنظیمی با نظارت‌های دقیق نهادی، خلأهای حقوق کیفری را پوشش داده و ضمانت اجرای پایداری برای صیانت از حقوق انسانی در بسترهای دیجیتال ایجاد می‌کند.

نتیجه‌گیری

پژوهش حاضر در پی پاسخ به این پرسش بنیادین سامان یافت که آیا ترویج الگوریتمی، تقویت تعامل‌محور یا ناتوانی در پالایش به‌موقع سخنان نفرت‌پراکن منتهی به خشونت‌های قومی، می‌تواند وصف مجرمانه در حقوق کیفری بین‌المللی بیابد و گردانندگان پلتفرم‌های اجتماعی را در جایگاه مرتکب یا معاون جنایات بین‌المللی بنشاند. واکاوی موشکافانه تحولات فنی و تطبیق آن با هنجارهای حقوق بشری و رویه قضایی نشان داد که تفاوت ماهوی ساختار شبکه‌های اجتماعی با رسانه‌های کلاسیک، مانع از تسری مکانیکی اصول سنتی مسؤولیت کیفری به زیست‌بوم دیجیتال است. در رسانه‌های سنتی، نظارت پیشینی سردبیری و کنترل انحصاری بر محتوا حاکم بود؛ درحالی‌که پلتفرم‌های نوین بر مبنای بازنشر پسینی، تولید محتوا توسط صدها میلیون کاربر و چینش خودکار بر پایه داده‌های رفتاری فعالیت می‌کنند. از این‌رو، خطای پایش فنی، کمبود نیروی انسانی مسلط به زبان‌های بومی یا تأخیر در حذف پست‌های التهاب‌آفرین، پیش از آنکه دلالت بر رفتار مجرمانه داشته باشد، بیانگر نارسایی سیستمی و ضعف در طراحی سازمانی است.

بر همین اساس، فرضیه پژوهش با تکیه بر یافته‌های تطبیقی به اثبات رسید؛ بدین معنا که مسؤولیت مدنی یا پاسخ‌گویی تنظیم‌گرانه شرکت‌های فناوری را نمی‌توان به‌گونه‌ای خودکار به قلمرو مسؤولیت کیفری بین‌المللی مدیران و طراحان تعمیم داد. عبور از آستانه مسؤولیت مدنی و ورود به دایره جرائم مشمول اساسنامه دیوان کیفری بین‌المللی، مستلزم احراز ارکان سه‌گانه جرم با استانداردهای بسیار سخت‌گیرانه است. نخست، بر پایه اصل صلاحیت شخصی، تنها اشخاص حقیقی دارای اختیار تصمیم‌گیری قابلیت پیگرد دارند و شخصیت حقوقی پلتفرم‌ها از دایره صلاحیت مستقیم خارج است. دوم، در بعد رکن مادی، انتساب عناوینی چون معاونت یا تسهیل ارتکاب جرم نیازمند اثبات نقش اساسی و مستقیم سامانه نرم‌افزاری در وقوع جنایت عینی است، به گونه‌ای که بتوان رابطه‌ای سبب‌ساز میان معماری عامدانه سامانه و آسیب‌های وارده در دنیای واقعی برقرار کرد. سوم و مهم‌تر از همه، سد نفوذناپذیر رکن روانی مانع از مجرم‌انگاری رفتارهای مبتنی بر سهل‌انگاری یا بی‌احتیاطی تجاری می‌شود. حقوق کیفری بین‌المللی برای انتساب جنایت علیه بشریت یا تحریک به نسل‌کشی، استانداردهای مطلق قصد و آگاهی و حتی سوءنیت ویژه را طلب می‌کند؛ عناصری که اثبات آن‌ها در مورد کارگزارانی که در پی بیشینه‌سازی سود اقتصادی و جذب مخاطب بوده‌اند و انگیزه‌ای برای نابودی فیزیکی یک گروه قومی نداشته‌اند، با موازین فعلی ناممکن می‌نماید. ناتوانی از تفکیک میان سودجویی نامشروع اقتصادی با هم‌راستایی ذهنی در جنایت بین‌المللی، نقض آشکار اصل بنیادین قانونی بودن جرائم و مجازات‌ها خواهد بود.

با توجه به این واقعیت‌های حقوقی، پر کردن خلأ پاسخ‌گویی در عصر الگوریتم‌ها نیازمند بازتعریف ابتکارات تقنینی و سیاست‌گذاری کیفری فراملی است. در گام نخست، دفتر دادستانی دیوان کیفری بین‌المللی در سند سیاست‌گذاری جرائم سایبری و محیط‌های دیجیتال باید به تدوین رهنمودهایی شفاف در خصوص آستانه مداخله کیفری اشخاص حقیقی اداره‌کننده فناوری‌ها روی آورد. این سند باید مرز دقیق میان هشدارهای کلی نهادهای مدنی با «آگاهی اختصاصی و غیرقابل انکار فردی» را ترسیم کند تا تعلل تعمدی پس از دریافت اخطارهای معتبر بین‌المللی، شاخصی برای ارزیابی قابلیت انتساب تلقی گردد. در گام دوم، تمرکز نظام بین‌الملل باید بر وادارسازی پلتفرم‌ها به رعایت اصل شفافیت الگوریتمی، ممیزی بی‌طرفانه کدهای توصیه‌گر و ارزیابی ریسک‌های حقوق بشری پیش از عرضه خدمات در مناطق بحران‌زده قرار گیرد. پاسخ‌گویی متناسب با عصر اطلاعات، از مسیر خروج از دگم‌های کلاسیک و از طریق تلفیق هوشمندانه ضمانت‌اجراهای تنظیمی الزام‌آور با چارچوب‌های هدفمند مسؤولیت مدنی شرکتی محقق می‌شود، نه با تلاش نافرجام برای گنجاندن رفتارهای فناوری‌محور در قالب‌های غیر منطبق حقوق کیفری.

ملاحظات اخلاقی: موارد مربوط به اخلاق در پژوهش و نیز امانت‌داری در استناد به متون و ارجاعات مقاله تماما رعایت گردید.

تعارض منافع: در تدوین این مقاله، هیچ‌گونه تعارض منافعی وجود نداشته است.

سهم نویسندگان: نگارش مقاله به‌صورت مشترک توسط نویسندگان انجام گرفته است.

تشکر و قدردانی: از تمام کسانی که در تهیه این مقاله یاری رسانده‌اند، سپاسگزاریم.

تأمین اعتبار پژوهش: پژوهش حاضر بدون تأمین مالی انجام شده است.

منابع

- Akhavan, P (2022). "The Perils of Progressive Jurisprudence: The Nullum Crimen Sine Lege Principle in International Criminal Law". *Current Legal Problems*, 75(1): 45–70.

- Álvarez-Ugarte, R (2025). "Bad Cover Versions of Law: On the Inherent Limits of Voluntary Human Rights Obligations, as Applied to Internet Companies doing Content Moderation". *Transnational Legal Theory*, 16(4): 611–635.

- Amnesty International (2022). *The Social Atrocity: Meta and the Right to Remedy for the Rohingya* (Index No. ASA 16/5933/2022).

- Anand, N (2024). "Of Code and Consequence: Assessing the Impact of Artificial Intelligence on International Criminal law Norms Governing the Direct and Public Incitement to Genocide". *International Journal of Legal Information*, 52(2): 166–175.

- Coco, A (2026). "Modes of Liability for AI-Enabled Crimes in International Criminal Law". *International Law Studies*, 107(1): 228–255.

- Davidson, C (2017). How to Read International Criminal Law: Strict Construction and the Rome Statute of the International Criminal Court". *St. John's Law Review*, 91(1), 37–103.

- de Souza Dias, T (2024). "Finding Common Ground: The Right to be Free from Incitement to Discrimination, Hostility, and Violence in the Digital age". *Global Responsibility to Protect*, 16(4): 391–425.

- Di Stefano, S (2025). "International Human Rights Law in Content Moderation and the Risks of 'Misdiagnosing' its Limits". *Transnational Legal Theory*, 16(4): 519–545.

- Enarsson, T (2024). "Navigating Hate Speech and Content Moderation under the DSA: Insights from ECtHR Case law". *Information & Communications Technology Law*, 33(3): 384–401.

- Fino, A (2020). "Defining Hate Speech: A Seemingly Elusive Task". *Journal of International Criminal Justice*, 18(1): 31–57.

- Fino, A (2023). "A Critique of the UN Strategy and Guidance on 'Hate Speech': Some Legal Considerations". *Netherlands Quarterly of Human Rights*, 41(4): 190–213.

- FitzGerald, M (2025). "Not Hollowed by a Delphic Frenzy: European Intermediary Liability from the Perspective of a Bad Man: A Response to Martin Husovec et al.". *Maastricht Journal of European and Comparative Law*, 32(1): 47–64.

- Flynn, A; Vakhitova, Z; Wheildon, L; Harris, B & Robards, B (2025). "Content moderation and Community Standards: The Disconnect between Policy and User Experiences Reporting Harmful and Offensive Content on Social Media". *Policy & Internet*, 17(3): 1-20.

- Gordon, G. S (2017). *Atrocity Speech Law: Foundation, Fragmentation, Fruition*. Oxford University Press.

- Griffin, R (2025). "Governing Platforms through Corporate Risk Management: The Politics of Systemic Risk in the Digital Services Act". *European Law Open*, 4(2): 223–253.

- Hamilton, R. J (2022). "Platform-Enabled Crimes: Pluralizing Accountability when Social Media Companies Enable Perpetrators to Commit Atrocities". *Boston College Law Review*, 63(4): 1349–1420.

- Hatano, A (2023). "Regulating Online Hate Speech through the Prism of Human Rights Law: The Potential of Localised Content Moderation". *Australian Year Book of International Law*, (41): 127–156.

- Nave, Eva (2026). "Criminal Hate Speech Attributable to Online Platforms: A Call for a thorough Corporate Remedial Responsibilities framework in Europe". *Business and Human Rights Journal*, 11(1): 46–71.

- Nieprzecka, A (2024). *Participation in Crime Falling within the Subject-Matter Jurisdiction of the International Criminal Court*. T.M.C. Asser Press.

- Office of the Prosecutor of the International Criminal Court. (2025, December 3). *Policy on Cyber-Enabled Crimes under the Rome Statute*. International Criminal Court.

- Office of the United Nations High Commissioner for Human Rights. (2025). *Practical Application of the Guiding Principles on Business and Human Rights to the Activities of Technology Companies, Including Activities Relating to Artificial Intelligence* (A/HRC/59/32). United Nations.

- Park, T. J & Rohatgi, A (2024). "Balancing the Platform Responsibility Paradox: A Case for Amplification Regulation to Mitigate the Spread of Harmful But Legal Content Online". *Computer Law & Security Review*, 52(1): 1-28.

- Parker, J. E. K (2015). *Acoustic jurisprudence: Listening to the Trial of Simon Bikindi*. Oxford University Press.

- Pisillo Mazzeschi, R & Carli, E (2024). "Proof of Specific Intent in the Crime of Genocide: The Case of South Africa v. Israel before the International Court of Justice". *Journal of International Criminal Justice*, 22(2): 429–443.

- Salama, I., & Wiener, M. (2025). "Faith for Rights in Armed Conflict: Lessons from Practice". *Journal of Human Rights Practice*, 17(2): 1-20.

- Santos, A; Cazzamatta, R & Napolitano, C. J (2025). "Holding Platforms Accountable in the Fight against Misinformation: A Cross-National Analysis of State-Established Content Moderation Regulations". *International Communication Gazette*, 87(8): 729–750.

- Holvoet, M (2022). "International Criminal Liability for Spreading Disinformation in the Context of Mass Atrocity". *Journal of International Criminal Justice*, 20(1): 223–250.

- Husovec, M; Grote, T; Mazhar, Y; Mikhaeil, C; Miñarro Escalona, H; Sinha Kumar, P & Sreenath, S (2024). "Grand Confusion after Sanchez v. France: Seven Reasons for Concern about Strasbourg Jurisprudence on Intermediaries". *Maastricht Journal of European and Comparative Law*, 31(3): 385–411.

- Independent Investigative Mechanism for Myanmar. (2024). *Anti-Rohingya Hate Speech on Facebook: Content and Network Analysis*. United Nations.

- Karlsson, M; Van Couvering, E & Lindell, J (2022). "Publishing, Sharing, and Spreading Online News: A Case Study of Gatekeeping Logics in the Platform Era". *Nordicom Review*, 43(2): 190–213.

- Kohla, U (2024). "Toxic Recommender Algorithms: Immunities, Liabilities and the Regulated Self-Regulation of the Digital Services Act and the Online Safety Act". *Journal of Media Law*, 16(2): 301–335.

- Mahdi, O. R & Rinwigati, P (2024). "The Urgency to Use the Rabat Plan of Action in the Hate Speech Offense". *Media Iuris*, 7(2): 323–342.

- Manley, S; Moslemzadeh Tehrani, P & Rasiah, R (2023). "Mapping Interpretation by the International Criminal Court". *Leiden Journal of International Law*, 36(3): 771–806.

- McDermott, Y (2024). *Proving International Crimes*. Oxford University Press.

- Mosely, E (2026). "Becoming the "High Priest of Hutu Supremacy": The Case against Historian Ferdinand Nahimana at the International Criminal Tribunal for Rwanda". *The Journal of African History*, 67(1): 1-20.

- Narrain, S (2026). "Facebook's Community Standards on Hate Speech, Virality and "Real-World" Harm in Contemporary India". *Asian Journal of Law and Society*, 13(Special Issue 2): 340–359.

- Savard, C (2026). “What Mens rea for Ecocide in the Rome Statute? ” *The International Journal of Human Rights*, 30(6): 1175–1192.

- Schaffner, B; Bhagoji, A. N; Cheng, S; Mei, J., Shen, J. L; Wang, G; Chetty, M; Feamster, N; Lakier, G & Tan, C (2024). Community Guidelines Make this the Best Party on the Internet: An In-Depth Study of Online Platforms’ Content Moderation Policies. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Article 486, pp. 1–16). Association for Computing Machinery.

- Schroeder, R (2025). “Content Moderation and the Digital Transformations of Gatekeeping”. *Policy & Internet*, 17(2): 1-24.

- Tambu, L. K (2026). “Social Media as a Digital Weapon in Modern Warfare: Rethinking Incitement to Genocide and Complicity in the Digital Age”. *Journal of Conflict and Security Law*, 31(2): 159–178.

- Tsilonis, V (2024). *The Jurisdiction of the International Criminal Court* (2nd ed.). Springer.

- van Baar, A (2024). “Corporate Involvement in Atrocity Crimes—Trends and Patterns in Regulation”. *Crime, Law and Social Change*, 82(4): 989–1010.

- van Sliedregt, E (2025). “The Future of International Criminal Justice is Corporate”. *Journal of International Criminal Justice*, 23(1): 17–26.

- Zarmsky, S (2024). “Is International Criminal Law Ready to Accommodate Online Harm? Challenges and Ppportunities”. *Journal of International Criminal Justice*, 22(1): 169–184.