University of Tehran Press

Comparative Law Review

Homepage: https://jcl.ut.ac.ir

Online ISSN: 2423-3404

Volume: 17, **Issue:** 1
Spring & Summer
2026

Copyright Challenges in Training Data for Generative AI Systems: A Study in US, EU, and Iranian Law

Pouyan Ahmadi[1] | **Ebrahim Rahbari**[2]✉

1. PhD Student in International Trade and Investment Law, Faculty of Law, Shahid Beheshti University, Tehran, Iran. Email: pouyan.ahm@gmail.com
2. Corresponding Author; Assistant Professor of International Trade Law and Intellectual Property and Cyberspace Law Department, Faculty of Law, Shahid Beheshti University, Tehran, Iran. Email: e_rahbari@sbu.ac.ir

Article Info

Article Type:
Research Article

Received:
2025/09/14
Received in revised form:
2025/12/16
Accepted:
2026/01/11
Published online:
2026/08/23

Keywords:
Copyright, Fair Use Exception, Generative Artificial Intelligence, Intellectual Property Law, Law of Emerging Technologies.

Abstract

This article examines the emerging copyright challenges arising from the rapid development of generative artificial intelligence, particularly large language models (LLMs). The main research question is whether the extraction and use of copyright-protected data for training LLMs constitute an infringement of authors' rights. Employing a descriptive– analytical method and a comparative approach, this article examines the legal dimensions of using copyright-protected works in training these models. By analyzing recent lawsuits filed against AI developers, the authors explore how copyright exceptions have been interpreted and applied in response to this issue within the legal systems of the United States and the European Union. The findings indicate that, while the U.S. framework allows limited reliance on the fair use exception under certain conditions, the EU's regulatory system adopts a stricter approach, defining rule-based exceptions for text and data mining. Furthermore, given the growing adoption of AI in Iran, the article evaluates the current state of the Iranian intellectual property regime and highlights significant legal gaps concerning AI training. In Iran, not only is the concept of fair use undefined, but the existing legal framework also fails to address the evolving challenges posed by emerging technologies, underscoring the need for legislative reform in light of these developments.

How To Cite

Ahmadi, Pouyan; Rahbari, Ebrahim (2026). Copyright Challenges in Training Data for Generative AI Systems: A Study in US, EU, and Iranian Law. *Comparative Law Review*, 17 (1), 25-50.
DOI: https://doi.com/10.22059/jcl.2026.400962.634797

DOI

10.22059/jcl.2026.400962.634797

مطالعات حقوق تطبیقی

Homepage: https://jcl.ut.ac.ir

University of Tehran Press

شاپا الکترونیکی: ۲۴۲۳-۳۴۰۴

دوره: ۱۷، شماره: ۱

بهار و تابستان ۱۴۰۵

چالش‌های کپی‌رایت در داده‌های آموزشی سامانه‌های هوش مصنوعی مولد: مطالعه‌ای در حقوق امریکا، اتحادیهٔ اروپا و ایران

پویان احمدی[۱] | ابراهیم رهبری[۲]✉

۱. دانشجوی دکتری حقوق تجارت و سرمایه‌گذاری بین‌المللی، دانشکدهٔ حقوق، دانشگاه شهید بهشتی، تهران، ایران. رایانامه: pouyan.ahm@gmail.com

۲. نویسندهٔ مسئول؛ استادیار گروه حقوق تجارت بین‌الملل و حقوق مالکیت فکری و فضای مجازی، دانشکدهٔ حقوق، دانشگاه شهید بهشتی، تهران، ایران. رایانامه: e_rahbari@sbu.ac.ir

اطلاعات مقاله

نوع مقاله: پژوهشی

تاریخ دریافت:
۱۴۰۴/۰۶/۲۳

تاریخ بازنگری:
۱۴۰۴/۰۹/۲۵

تاریخ پذیرش:
۱۴۰۴/۱۰/۲۱

تاریخ انتشار برخط:
۱۴۰۵/۰۶/۰۱

کلیدواژه‌ها:
استثنای استفادهٔ منصفانه، حقوق فناوری‌های نوین، حقوق مالکیت فکری، کپی‌رایت، هوش مصنوعی مولد.

چکیده

این مقاله به بررسی چالش‌های نوظهور کپی‌رایت در پرتو توسعهٔ فزایندهٔ هوش مصنوعی مولد، به‌ویژه مدل‌های زبانی بزرگ می‌پردازد. پرسش اصلی پژوهش این است که چگونه استخراج و استفاده از داده‌های دارای کپی‌رایت برای آموزش مدل‌های زبانی بزرگ، نقض حقوق پدیدآورندگان محسوب می‌شود. این مقاله با روش توصیفی- تحلیلی و با بهره‌گیری از رویکرد تطبیقی، ابعاد حقوقی استفاده از آثار دارای کپی‌رایت در آموزش این مدل‌ها را بررسی می‌کند. نگارندگان با تحلیل پرونده‌های قضایی اخیر علیه توسعه‌دهندگان سامانه‌های هوش مصنوعی، نحوهٔ تفسیر و اعمال استثناهای حقوق مؤلف در مواجهه با این موضوع را در نظام‌های حقوقی ایالات‌متحده و اتحادیهٔ اروپا مطالعه کرده‌اند. یافته‌ها نشان می‌دهد که در امریکا، ضمن حمایت از حقوق دارندگان آثار، امکان بهره‌گیری از استثنای استفادهٔ منصفانه در شرایط خاص وجود دارد؛ حال آنکه مقررات اتحادیهٔ اروپا سخت‌گیرانه‌تر بوده و استثناهای قاعده‌محور داده‌کاوی را تعریف کرده است. در ادامه، باتوجه به گسترش استفاده از هوش مصنوعی در ایران، پژوهش حاضر وضعیت موجود در نظام مالکیت فکری ایران را ارزیابی کرده، خلأهای قانونی مرتبط با آموزش مدل‌های هوش مصنوعی را برجسته می‌سازد. در حقوق ایران نه‌تنها مفهوم «استفادهٔ منصفانه» تعریف روشنی ندارد، بلکه چارچوب‌های کنونی نیز پاسخگوی نیازهای حقوقی در مواجهه با فناوری‌های نوین نیستند و به بازنگری در پرتو تحولات این حوزه نیاز دارند.

استناد

احمدی، پویان؛ رهبری، ابراهیم (۱۴۰۵). چالش‌های کپی‌رایت در داده‌های آموزشی سامانه‌های هوش مصنوعی مولد: مطالعه‌ای در حقوق امریکا، اتحادیهٔ اروپا و ایران. *مطالعات حقوق تطبیقی*، ۱۷ (۱)، ۲۵- ۵۰.
DOI: https://doi.com/10.22059/jcl.2026.400962.634797

DOI 10.22059/jcl.2026.400962.634797

ناشر مؤسسه انتشارات دانشگاه تهران.

۱. مقدمه

در عصر پرشتاب فناوری، هوش مصنوعی به نیرویی تحول‌آفرین در عرصه‌های گوناگون ازجمله صنعت، بهداشت، ارتباطات و سرگرمی بدل شده و تأثیرات ژرفی بر این حوزه‌ها بر جای گذاشته است. در این میان، مدل‌های زبان بزرگ مانند چت‌جی‌پی‌تی[1] با توانایی چشمگیر خود در تولید متنی شبیه به زبان انسان، مرز میان خلاقیت انسانی و محتوای تولیدشده از سوی ماشین را بیش‌ازپیش کم‌رنگ کرده‌اند. این مدل‌ها که بر پایهٔ حجم عظیمی از داده‌های متنی و کد، آموزش دیده‌اند، قادرند انواع محتوای متنی خلاقانه‌ای همچون شعر، کد برنامه‌نویسی، فیلم‌نامه و غیره را تولید کنند (Haleem, et al., 2022: 3).

باوجود این پیشرفت‌های فناورانه، چالش‌های حقوقی قابل‌توجهی نیز پدید آمده است که شاید مهم‌ترین آنها در تعامل پیچیدهٔ میان هوش مصنوعی و حقوق ناظر به کپی‌رایت قرار دارد. مجموعه‌های وسیع داده که برای آموزش مدل‌های زبانی بزرگ استفاده می‌شوند، اغلب شامل محتوای دارای کپی‌رایت است. این امر چالش‌هایی را در مورد میزان نقض کپی‌رایت و پیامدهای آن برای مالکیت و حفاظت از محتوای تولیدشده از طریق هوش مصنوعی مطرح می‌سازد (Craig, 2021: 13-17). پرونده‌های حقوقی اخیر این چالش‌ها را برجسته‌تر کرده‌اند. برای مثال، در دو دعوای *ترمبلی و سیلورمن علیه شرکت توسعه‌هندهٔ هوش مصنوعی اوپن‌اِی‌آی*، شاکیان مدعی‌اند که برای آموزش هوش مصنوعی این شرکت، بدون اجازه از آثار دارای کپی‌رایت استفاده شده است[2]. استفاده از محتوای دارای کپی‌رایت در داده‌های آموزشی چت‌جی‌پی‌تی نگرانی‌هایی جدی دربارهٔ نقض احتمالی کپی‌رایت ایجاد کرده است. در صورت بازتولید این محتوای بدون مجوز، امکان مسئولیت حقوقی اوپن‌اِی‌آی وجود دارد. این وضعیت پرسش‌هایی اساسی دربارهٔ پدیدآورنده، مالکیت و حدود استفادهٔ منصفانه مطرح می‌سازد که مستلزم بازنگری در چارچوب‌های حقوقی و اتخاذ رویکردهای نوین است (Lennartz & Kraetzig, 2024: 2-5).

چالش‌های کپی‌رایت هوش مصنوعی، به‌ویژه در مدل‌های زبانی بزرگ، پیچیده و چندجانبه هستند و تعادل میان نوآوری فناورانه و حفاظت از مالکیت فکری را دشوار می‌سازند. ازجمله چالش‌های مهم می‌توان به استفادهٔ غیرمجاز از آثار دارای کپی‌رایت در داده‌های آموزشی، ابهام در مالکیت آثار تولیدشده از سوی هوش مصنوعی و احتمال نقض حریم خصوصی در فرایند جمع‌آوری داده‌ها، اشاره کرد (Savrai, 2025: 12-16). با توسعهٔ کاربست هوش مصنوعی، رسیدگی به این مقولات برای حمایت از حقوق پدیدآورندگان و توسعهٔ مسئولانهٔ فناوری ضروری است. در ایران نیز استفاده از هوش مصنوعی رو به افزایش است و از محتوایی که برای اولین بار منتشر شده و تحت حمایت کپی‌رایت قرار گرفته است نیز

1. ChatGPT'
2. Tremblay v. OpenAI, Inc., 2023; Silverman v. OpenAI, Inc., 2023

در مجموعه‌داده‌هایی که برای آموزش سیستم‌های هوش مصنوعی- که منحصر به چت‌جی‌پی‌تی نیست- استفاده می‌شود. از همین رو، بررسی این موضوع در چارچوب نظام حقوق مالکیت فکری ایران نیز برای عیان ساختن چالش‌ها ضرورت دارد.

در این مقاله به روش تحلیلی- توصیفی و مبتنی بر مطالعهٔ تطبیقی، به بررسی عمیق‌تر این چالش‌ها در پرتو انگاره‌ها و تحولات حقوق امریکا، اتحادیهٔ اروپا و ایران پرداخته شده و تمرکز بر ورودی‌های هوش مصنوعی- نه خروجی‌های آن- قرار گرفته است. نگارندگان این پژوهش تلاش می‌کنند به این پرسش پاسخ دهند که آیا استخراج و استفاده از داده‌های دارای کپی‌رایت برای آموزش مدل‌های زبانی بزرگ، نقض حقوق پدیدآورندگان محسوب می‌شود و در صورت مثبت بودن پاسخ، با چه سازوکارهای حقوقی می‌توان میان حمایت از آفرینندگان اثر و پیشرفت فناوری تعادل برقرار کرد؟

۲. مدل‌های زبانی بزرگ و پیوند با نظام کپی‌رایت

مدل‌های زبانی بزرگ (LLM)[1]، نوعی الگوریتم هوش مصنوعی یادگیری عمیق هستند که بر روی حجم عظیمی از داده‌های متنی و کدی آموزش دیده‌اند. کارکرد بنیادین آنها پیش‌بینی «توکن»[2] بعدی در یک دنباله است. توکن‌ها کوچک‌ترین واحدهای معنایی قابل پردازش (مانند کلمات، بخشی از کلمات یا علائم نگارشی) هستند و مدل بر اساس دنباله‌ای از این توکن‌ها که به‌عنوان ورودی دریافت کرده است، به تولید محتوا می‌پردازد. این مدل‌ها با تکیه بر توانایی پیش‌بینی، ساختارهای پیچیدهٔ زبان انسانی را فرا می‌گیرند و قادر به تولید متن منسجم، ترجمهٔ دقیق و خلاصه‌سازی هوشمندانه می‌شوند (Zhao et al., 2023: 60). مدل‌های پیش‌گفته عمدتاً بر اساس ساختار ترانسفورمر طراحی شده‌اند. این ساختار با بهره‌گیری از مکانیسم «خودتوجهی»[3]، به مدل اجازه می‌دهد هنگام پردازش هر بخش از متن، به اهمیت بخش‌های دیگر، ازجمله بخش‌های دورتر در همان متن توجه کند. این توانایی پردازش سریع و درک وابستگی‌های دوربرد، امکانات مدل را متحول کرده است (Chang et al., 2023: 5).

مدل‌های زبانی بزرگ در واقع شبکه‌های عصبی مصنوعی هستند که بر روی مجموعه‌داده‌های عظیمی از متن و کد آموزش می‌بینند و الگوها و ظرافت‌های زبان انسانی را درک می‌کنند (Hadi et al., 2023: 7-8). از شناخته‌شده‌ترین نمونه‌های این مدل‌ها می‌توان به جی‌پی‌تی-۴[4] (توسعه‌یافته از سوی شرکت اوپن‌اِی‌آی) و بِرت[5] (توسعه‌یافته از سوی شرکت گوگل) اشاره کرد. این مدل‌ها با استفاده از

1. Large Language Models
2. ‘Token’
3. ‘Self-Attention’
4. GPT-4
5. BERT

یادگیری خودنظارتی (بدون نیاز به داده‌های برچسب‌گذاری‌شده) و مدل‌سازی خودبازگشتی، متن‌های منسجم و معنادار تولید می‌کنند (Hadi et al., 2023: 7-8). قابلیت‌های آنها فراتر از پیش‌بینی کلمات است و شامل ترجمهٔ ماشینی، خلاصه‌سازی متن، پاسخ به سؤالات و تولید محتوای خلاقانه مانند شعر و کد می‌شود که پتانسیل تحول در صنایع مختلف را دارد (Keary, 2023).

از مهم‌ترین کاربردهای مدل‌های زبانی بزرگ، به‌کارگیری آنها در ربات‌های گفتگو (چَت‌بات)[1] است که چت‌جی‌پی‌تی (توسعه‌یافته از سوی شرکت اوپن‌اِی‌آی در سال ۲۰۲۲)، شاخص‌ترین نمونهٔ آن است. نحوهٔ عملکرد چت‌جی‌پی‌تی با استفاده از معماری ترانسفورمر و ترکیب یادگیری تقویتی از بازخورد انسانی (RLHF)[2]، برای مکالمات بهینه شده است (Gertner, 2023; Uszkoreit, 2017). فرایند آموزش این مدل از داده‌های عظیمی شامل آثار دارای کپی‌رایت مانند کتاب‌ها و منابع آنلاین استفاده می‌کند و این استفاده از داده‌ها، نقطهٔ آغاز پیوند مدل‌های زبانی با نظام کپی‌رایت محسوب می‌شود. این فرایند آموزشی چندمرحله‌ای، شامل آموزش تحت نظارت اولیه و سپس یادگیری تقویتی است که در آن مربیان انسانی به‌عنوان کاربر و دستیار عمل می‌کنند تا مدل را به تولید پاسخ‌های طبیعی و جذاب هدایت نمایند (Schulman et al., 2017: 1). این فرایند شامل مراحل پیش‌پردازش داده‌ها، آموزش مدل، ارزیابی انسانی، پاداش‌دهی برای تقویت پاسخ‌های مطلوب و تکرار برای بهبود مستمر است (Roumeliotis & Tselikas, 2023: 2-6). با این حال، شباهت تولیدات این مدل‌ها با آثار دارای کپی‌رایت، نگرانی‌هایی جدی را در مورد نقض حقوق مالکیت فکری ایجاد کرده و موارد متعددی از اتهام به سرقت ادبی و استفاده از داده‌های دارای کپی‌رایت را برای آموزش مطرح ساخته است (Ortiz, 2023).

۳. چالش‌های کپی‌رایت با مدل‌های زبانی بزرگ

مسائل پیش‌آمده دربارهٔ کپی‌رایت در مدل‌های زبانی بزرگ مانند چت‌جی‌پی‌تی را می‌توان در دو بخش داده‌های آموزشی و محتوای خروجی بررسی کرد. اگرچه تمرکز حقوقی بیشتر بر مالکیت خروجی است، استفاده از آثار دارای کپی‌رایت در آموزش نیز نگرانی‌زا است. این مدل‌ها بر روی مجموعه‌های عظیمی از داده‌ها که ممکن است شامل آثار دارای کپی‌رایت باشند، آموزش می‌بینند. این موضوع، به‌ویژه در دعاوی اخیر علیه توسعه‌دهندگان هوش مصنوعی مانند اوپن‌اِی‌آی، به مسئله‌ای بزرگ تبدیل شده است. ماهیت فنی این مدل‌ها که مستقیماً از داده‌های آموزشی یاد می‌گیرند و خروجی‌هایی شبیه به آثار انسانی تولید می‌کنند (Bender & Koller, 2020: 85-88)، تعیین مالکیت کپی‌رایت را دشوار می‌سازد. ازآنجا که

1. ChatBot'
2. 'Reinforcement Learning from Human Feedback'

این مدل‌ها فاقد آگاهی و نیت هستند و عملکردشان مصداق «عمل بدون فهم»[1] است (Floridi, 2023: 15)، نمی‌توانند به‌عنوان پدیدآورندهٔ انسانی شناخته شوند.

نمونهٔ بارز این محدودیت در پروندهٔ *دکتر ثالر* در امریکا دیده شده است. در این پرونده، دادگاه فدرال و دفتر کپی‌رایت امریکا درخواست ثبت نقاشی خلق‌شده از سوی هوش مصنوعی را به دلیل فقدان شرط اساسی تألیف انسانی برای برخورداری از حمایت کپی‌رایت رد کردند (Drawdy, 2023; U.S. Copyright Office Review Board, n.d. 2023)[2]. این رویکرد تأکید می‌کند که هوش مصنوعی نمی‌تواند خالق قانونی یک اثر هنری یا ادبی باشد (Ginsburg & Budiardjo, 2019: 111-115).

مالکیت محتوای تولیدی نیز مسئله‌برانگیز است. در توافقنامهٔ کاربری اوپن‌اِی‌آی، مالکیت خروجی چت‌جی‌پی‌تی به کاربر واگذار شده و مسئولیت کپی‌رایت آن نیز برعهدهٔ کاربر است (OpenAI, n.d-c; for Terms of Use). اگر این واگذاری معتبر نباشد، مسائل حقوقی تازه‌ای ایجاد می‌شود. با این حال، مسئلهٔ اصلی در داده‌های ورودی است (Radford et al., 2019: 491-494). هوش مصنوعی برای آموزش خود به حجم عظیمی از داده‌های عمومی، ازجمله محتوای دارای کپی‌رایت متکی است. پرونده‌های حقوقی متعددی در سال ۲۰۲۳ علیه اوپن‌اِی‌آی مطرح شده‌اند که استفادهٔ غیرمجاز این شرکت از آثار دارای کپی‌رایت برای آموزش مدل‌های خود را به‌چالش می‌کشند. این مسائل نشان می‌دهند که مدل‌های زبانی بزرگ، چارچوب‌های حقوقی سنتی را زیر سؤال برده، نیازمند رویکردهای جدیدی هستند.

گزارش دفتر کپی‌رایت امریکا در می ۲۰۲۵، دربارهٔ نگرانی‌های استفاده از آثار دارای کپی‌رایت در آموزش هوش مصنوعی، بر لزوم خلاقیت انسانی در بهره‌مندی از حمایت کپی‌رایت تأکید کرده و خواستار وضع مقرراتی روشن شده است. این گزارش که مشروعیت اصل «استفادهٔ منصفانه» را زیر سؤال برده بود، به اخراج رئیس این دفتر، شیرا پرلماتر[3]، منجر شد (U.S. Copyright Office, 2025: 32-84). این اتفاق، کشمکش میان سیاست‌گذاری‌های فناورانه، حفظ حقوق پدیدآورندگان و استقلال نهادی را در برابر تحولات هوش مصنوعی نشان می‌دهد.

۴. دعاوی حقوقی علیه مدل‌های زبانی بزرگ

دعاوی حقوقی علیه مدل‌های زبانی بزرگ، قلب تپندهٔ بحث نقض کپی‌رایت در هوش مصنوعی مولد است. این دعاوی عمدتاً بر دو محور کلیدی متمرکز شده‌اند: ۱) نقض در مرحلهٔ آموزش (استفادهٔ بدون

1. “Agere Sine-Intelligere”
2. Thaler v. Perlmutter, 2023
3. ‘Shira Perlmutter’

مجوز از آثار دارای کپی‌رایت در مجموعه‌داده‌های آموزشی)؛ ۲) نقض در مرحلهٔ خروجی (تولید محتوای مشابه یا خلاصه‌های دقیق از آثار اصلی). در مقابل، موضع دفاعی اصلی شرکت‌ها، استناد به دکترین استفادهٔ منصفانه است. در ادامه، مهم‌ترین دعاوی اخیر برای تحلیل این تمایزات و بررسی استدلال‌های طرفین مرور می‌شود.

۴. ۱. دعوای ترمبلی علیه اوپن‌اِی‌آی

نویسندگان معروف، مونا عواد و پل ترمبلی، در بهار ۲۰۲۳ در دادگاه فدرال امریکا علیه شرکت اوپن‌اِی‌آی شکایت کردند. آنها مدعی نقض در هر دو محور آموزش و خروجی بودند: اولاً ادعا کردند کتاب‌هایشان که تحت کپی‌رایت بوده، بدون اجازه برای آموزش مدل جی‌پی‌تی-۳ استفاده شده است؛ و ثانیاً مدعی شدند که این چت‌بات خلاصه‌های دقیقی از آثارشان تولید کرده است. شاکیان، اوپن‌اِی‌آی را به نقض مستقیم کپی‌رایت، نقض ناشی از فعل غیر، نقض قانون DMCA و رقابت ناعادلانه متهم کردند. در زمستان ۲۰۲۴، دادگاه بخش عمده‌ای از این ادعاها، ازجمله ادعاهای مربوط به نقض در خروجی (نقض حقوق ناشی از فعل غیر و رقابت ناعادلانه) و نقض قانون DMCA (دربارهٔ حذف یا تغییر اطلاعات مدیریت حق مؤلف) را به دلیل نبود ادلهٔ کافی رد کرد. در ادامهٔ روند تحصیل ادله، دادگاه در زمستان ۲۰۲۵ دستور داد اوپن‌اِی‌آی کلِ «مجموعه‌دادهٔ انگلیسی کولانگ»[1] را تحویل دهد تا شاکیان بتوانند بررسی کنند آیا در آموزش هوش مصنوعی از آثارشان استفاده شده است یا خیر؛ اما جزئیات فنی حساس، محرمانه باقی خواهد ماند[2]. اکنون پرونده صرفاً بر این متمرکز است که آیا استفاده از مطالب دارای کپی‌رایت در مجموعه‌داده برای آموزش هوش مصنوعی، مصداق نقض مستقیم کپی‌رایت است یا خیر.

۴. ۲. دعوای سیلورمن علیه اوپن‌اِی‌آی

در بهار ۲۰۲۳، سارا سیلورمن و دو نویسندهٔ دیگر علیه اوپن‌اِی‌آی در کالیفرنیا دعوایی دسته‌جمعی را مطرح

---

1. "The English Colang Dataset"

مجموعه داده انگلیسی کولانگ (موسوم به cc-en-colang-v2-20220131-metadata)، یک مجموعه داده عظیم متنی است که از بخش‌های فیلترشده و باکیفیت وب (عمدتاً از Common Crawl) استخراج شده است. هدف اصلی از ایجاد این مجموعه داده، تأمین حجم عظیمی از داده‌های متنی انگلیسی با کیفیت بالا و متنوع برای آموزش مدل‌های زبانی بزرگ نظیر GPT-3 و GPT-4 (توسط اوپن‌اِی‌آی) بوده است. کارکرد این مجموعه داده، ارائه ورودی خام به مدل است تا با استفاده از آن، ساختارها، الگوها، دستور زبان و محتوای گسترده زبان انگلیسی را فرا بگیرد. این مجموعه داده به دلیل گستردگی منابع، شامل مقادیر زیادی از آثار دارای کپی‌رایت است و همین موضوع، آن را به عنصری کلیدی در پرونده‌های حقوقی مربوط به نقض کپی‌رایت تبدیل کرده است.

2. Tremblay v. OpenAI, Inc., 2023 to 2205.

کردند. این شاکیان نیز مانند پروندهٔ ترمبلی، بر نقض در مرحلهٔ آموزش تأکید داشته، مدعی شدند که اوپن‌اِی‌آی بدون مجوز، از آثارشان برای آموزش مدل‌هایش استفاده کرده و داده‌ها از «کتابخانه‌های سایه» مانند Library-Genesis به‌طور غیرقانونی گردآوری شده است. اتهامات شامل نقض کپی‌رایت، سهل‌انگاری، تحصیل نامشروع و رقابت غیرمنصفانه بود. شرکت اوپن‌اِی‌آی در دفاعیات خود تأکید کرد استفاده از این آثار مشمول دکترین استفادهٔ منصفانه است و نویسندگان مفهوم کپی‌رایت و استثناهای آن را نادرست درک کرده‌اند. در زمستان ۲۰۲۴، دادگاه بخش عمده‌ای از ادعاهای حقوقی شاکیان (به‌ویژه نقض کپی‌رایت در خروجی) را رد کرد و تنها رقابت غیرمنصفانه را پذیرفت. پس از دادخواست اصلاحی شاکیان، اوپن‌اِی‌آی در تابستان ۲۰۲۴ همهٔ اتهامات را رد کرد و همچنان بر استفادهٔ منصفانه تأکید نمود. در ژانویه ۲۰۲۵، دادگاه دستور ارائهٔ مجموعه‌دادهٔ انگلیسی کولانگ را صادر کرد تا امکان بررسی ادعای نقض در مرحلهٔ آموزش فراهم شود. با این حال، درخواست شاکیان برای تعیین یک کارمند اوپن‌اِی‌آی به‌عنوان مسئول ارائهٔ اسناد در فوریه ۲۰۲۵ و نیز تلاش آنها برای مداخله در دعوای موازی نیویورک رد شد[1].

۴. ۳. دعوای نیویورک تایمز علیه مایکروسافت اوپن‌اِی‌آی

در پاییز ۲۰۲۳، روزنامهٔ نیویورک تایمز شکایتی علیه اوپن‌اِی‌آی و مایکروسافت ثبت کرد. شاکی مدعی شد این شرکت‌ها بدون کسب اجازه، میلیون‌ها مقالهٔ خبری این روزنامه را جمع‌آوری و با حذف یا تغییر اطلاعات مدیریت حقوق مؤلف (CMI)، از آنها برای آموزش هوش مصنوعی استفاده کرده‌اند که مصداق نقض مستقیم کپی‌رایت و رقابت ناعادلانه است. اوپن‌اِی‌آی در مقام دفاع، بر دکترین «استفادهٔ منصفانه» تأکید و استدلال نموده است که استفاده از این داده‌ها برای اهداف تحلیلی و «کارکرد انتقالی و تحلیلی»[2] نقض کپی‌رایت محسوب نمی‌شود. این دفاعیات استدلال می‌کنند که مدل، صرفاً الگوها را می‌آموزد و نسخهٔ اصلی را بازتولید نمی‌کند. در بهار ۲۰۲۴، دادگاه بخش عمده‌ای از ادعاهای فرعی را رد کرد، اما اجازه داد شکایت اصلی شامل نقض کپی‌رایت ادامه یابد. در پاییز ۲۰۲۴ اعلام شد که شماری از اسناد آموزشی مهم به‌اشتباه پاک شده و این مسئله مانع دسترسی کامل به شواهد گردیده است. در ادامه، پروندهٔ مشابه «مرکز گزارشگری تحقیقی (CIR)»[3] با دعوای نیویورک تایمز ادغام گردید و هم‌اکنون مراحل کشف مدارک و بررسی مباحث مربوط به «استفادهٔ منصفانه» و مدیریت حقوق مؤلف در جریان است[4].

1. Silverman v. OpenAI, Inc., 2023 to 2025.
2. “Transformative Use”
3. “The Center for Investigative Reporting”
4. The New York Times Co. v. Microsoft Corp. et al., 2023 to 2025.

۴. ۴. پرونده بارتز علیه آنتروپیک پی‌بی‌سی

در یکی از جدیدترین و مهم‌ترین دعاوی مرتبط با نقض کپی‌رایت در حوزهٔ آموزش مدل‌های زبانی بزرگ، سه نویسنده به همراه شرکت‌های انتشاراتی وابسته به آنان، در تابستان ۲۰۲۴ علیه شرکت آنتروپیک در دادگاه اقامهٔ دعوا کردند. این شرکت متهم شد که بدون کسب مجوز، میلیون‌ها کتاب دارای کپی‌رایت را از منابع ناقض کپی‌رایت همچون Books3، LibGen و Pirate Library Mirror دانلود کرده و برای آموزش مدل‌های خود به‌کار برده است. این پرونده به‌طور ویژه‌ای بر منبع گردآوری داده‌ها و نقض در مرحلهٔ آموزش تمرکز داشت. دادگاه در بهار ۲۰۲۵ رأی داد که نگهداری نسخه‌های ناقض کپی‌رایت از این آثار در کتابخانهٔ مرکزی آنتروپیک برای استفاده‌های عمومی در آینده، مشمول استفادهٔ منصفانه نیست و نقض صریح حقوق مؤلف محسوب می‌شود. با این همه، استفاده از نسخه‌های خریداری‌شدهٔ فیزیکی که برای آموزش دیجیتال‌سازی شده بودند، تحت شرایط خاص به‌عنوان استفادهٔ منصفانه پذیرفته شد. این رأی نقطهٔ عطفی در تفکیک دقیق میان انواع استفاده از داده‌های دارای کپی‌رایت در فرایند آموزش مدل‌های زبانی محسوب می‌شود و بر ضرورت رعایت حقوق پدیدآورندگان حتی در فرایندهای فناورانه تأکید دارد[1].

بررسی این دعاوی نشان می‌دهد که بحث حقوقی پیرامون مدل‌های زبانی بزرگ حول دو محور اصلی می‌چرخد: اولاً مسئلهٔ «نقض در خروجی» (خلاصه‌سازی یا تولید محتوای مشابه) که در پرونده‌هایی مانند ترمبلی، دادگاه‌ها عمدتاً ادعای نقض مستقیم را به دلیل نبود شواهد کافی یا عدم شباهت نزدیک رد کرده‌اند. ثانیاً مسئلهٔ حیاتی‌تر «نقض در مرحلهٔ آموزش» (استفاده از آثار کپی‌رایت در مجموعه‌داده‌ها) که هنوز به نتیجهٔ قطعی نرسیده است. هستهٔ اصلی دفاع شرکت‌ها در این محور، استناد به مفهوم «استفادهٔ منصفانه» با تأکید بر ماهیت آماری و دگرگون‌کنندهٔ مدل‌های هوش مصنوعی است؛ به این معنی که مدل صرفاً الگوها را می‌آموزد و نسخهٔ کپی از اثر تولید نمی‌کند. با این حال، رأی پروندهٔ *باتلرز علیه آنتروپیک* که بین استفاده از نسخه‌های غیرقانونی (ردشده) و نسخه‌های قانونی (پذیرفته‌شده) تمایز قائل شد، نشان می‌دهد که منبع و نحوهٔ گردآوری داده‌ها برای دادگاه‌ها اهمیت فزاینده‌ای دارد. نتیجهٔ نهایی این دعاوی، به‌ویژه در پرونده‌های در جریان مانند نیویورک تایمز، چارچوب حقوقی و اقتصادی آیندهٔ هوش مصنوعی مولد را تعیین خواهد کرد.

۵. بررسی حقوقی داده‌های آموزشی هوش مصنوعی

برای رویارویی با چالش‌های کپی‌رایت در داده‌های آموزشی هوش مصنوعی، باید مفهوم استفادهٔ منصفانه

1. Bartz v. Anthropic PBC, 2024-2025.

و استثناهای مشابه آن به‌خوبی درک و تحلیل شود. در اینجا رویکردهـای حقـوقی در سـه نظـام ایـالات متحده، اتحادیهٔ اروپا و ایران به‌صورت تطبیقی بررسی می‌شود تا نقاط قوت و ضعف هر رویکرد در قبـال مقولهٔ استخراج متن و داده (TDM)[1] و انطباق این استفاده با اصول کپی‌رایت مشخص گردد. ازآنجا کـه مدل‌های یادگیری ماشین از طیف گسترده‌ای از داده‌ها اسـتفاده می‌کننـد، بررسـی میـزان انطبـاق ایـن استفاده با اصول دکترین استفادهٔ منصفانه اهمیت دارد. در این بخش، داده‌های آموزشی و تطبیق آنهـا بـا معیارهای استفادهٔ منصفانه بررسی می‌شوند.

۵. ۱. نقش داده‌ها در آموزش مدل‌های زبانی بزرگ

مدل‌های زبانی بزرگ با روش‌هایی ماننـد پیش‌آمـوزی و تنظـیم دقیـق (Roumeliotis & Tselikas, 2023: 5)، یادگیری تقویتی با بازخورد پاداش/ تنبیه انسانی و کنشگرهای انطباق‌پذیر آمـوزش می‌بیننـد که امکان یادگیری از داده‌های عظیم و انجام وظایف پردازش زبان طبیعی را فـراهم می‌کنـد (Wolfe, 2023). یادگیری خودنظارتی پایهٔ اصـلی عملکـرد آنهاسـت (Roumeliotis & Tselikas, 2023: 5). علاوه بر این، استخراج متن و داده برای شناسایی الگوهای معنـادار (Hearst, 1999: 2-4) و یـادگیری عمیق مولد (GDL)[2] برای تولید محتوای مشابه داده‌های آموزشـی بـه‌کار مـی‌رود (Franceschelli & Musolesi, 2022: 2-3). در این فرایند، آثار به واحدهای ریز زبانی موسوم به «توکن» تجزیه می‌شوند. نمونهٔ شاخص این تکنیک‌ها چت‌جی‌پی‌تی است کـه بـر معمـاری جی‌پی‌تـی اسـتوار بـوده، بـرای تولیـد محتوای دقیق و نوآورانه به داده‌های انبوه برای عملکرد مؤثر نیاز دارد.

تعمق در رأی بارتز علیه آنتروپیک نشان می‌دهد که شرکت‌های توسعه‌دهندهٔ مدل‌های زبـانی بـزرگ مانند آنتروپیک، نه‌تنها از محتوای دارای کپی‌رایت برای استخراج توکن‌ها بهره می‌گیرند، بلکه این آثار را به‌صورت فشرده در مدل ذخیره می‌کنند، به‌طوری که امکان بازتولید دقیق آثار اولیه از سوی مدل فـراهم می‌شود. این مسئله نشان‌دهندهٔ حفظ حافظه‌ای شبه‌دائمی از محتوای اصـلی در مـدل اسـت کـه از نظـر دادگاه، یکی از دلایل ردّ دفاع «استفادهٔ منصفانه» برای نسخه‌های غیرمجاز تلقی می‌شود[3].

۵. ۲. استثنای حمایت و استفادهٔ منصفانه در آموزش چت‌جی‌پی‌تی

توسعه‌دهندگان بزرگ هوش مصنوعی مانند مایکروسافت، گوگل و اوپن‌اِی‌آی، برای آموزش مدل‌هایشان به حجـم وسـیعی از داده‌هـای متنـی و تصـویری دسترسـی یافته‌انـد. بخشـی از ایـن داده‌هـا مشـمول

1. ‘Text and Data Mining’
2 .‘Generative Deep Learning’
3. ‘Bartz v. Anthropic PBC, 2025’

کپی‌رایت‌اند، و بخشی دیگر ازجمله اطلاعات عمومی یا آثار واردشده به عرصهٔ عمومی هستند (Foster, 2019: 139-140). این موضوع بحث‌هایی را دربارهٔ قانونی بودن بهره‌گیری از داده‌های دارای کپی‌رایت در آموزش مدل‌های زبانی برانگیخته است. چهار شکل اصلی استفاده از داده‌های آموزشی قابل تشخیص‌اند: الف) داده‌هایی که تحت کپی‌رایت نیستند، مانند آثار واردشده به مالکیت عمومی؛ ب) داده‌های دارای کپی‌رایت که به‌طور قانونی از صاحبان حق مجوز گرفته‌اند یا تحت مجوز بازنشر شده‌اند؛ ج) استفاده‌هایی که به بازار اثر اصلی آسیب می‌رسانند؛ و د) استفاده‌هایی که به بازار آسیبی نمی‌زنند (Sobel, 2020: 15-16).

در ایالات متحده، آثار، ۹۵ سال پس از انتشار یا ۷۰ سال پس از درگذشت نویسنده (بسته به تاریخ آفرینش)، در اتحادیهٔ اروپا ۷۰ سال پس از مرگ آخرین پدیدآورنده، و در ایران ۵۰ سال پس از فوت مؤلف، وارد مالکیت عمومی می‌شوند. اگر اثری مشمول کپی‌رایت باشد، اما مجوز انتشار دیجیتال آن محدودیتی برای تکثیر قائل نشده باشد، استفاده از آن در آموزش مدل‌های هوش مصنوعی مجاز تلقی می‌شود. اما دربارهٔ استفاده از آثاری که تحت کپی‌رایت قرار دارند، ولی به‌صورت دیجیتال در دسترس نیستند، همچنان ابهام وجود دارد (Sobel, 2020: 25). برای رفع این ابهام، مقررات مرتبط با استثنای حمایت از حقوق انحصاری پدیدآورنده و استفادهٔ منصفانه با تمرکز بر معاهدات بین‌المللی مالکیت فکری، اتحادیهٔ اروپا، ایالات متحده و ایران، تحلیل می‌شود.

۵. ۲. ۱. معاهدات بین‌المللی مالکیت فکری

در سطح بین‌المللی، نظام حقوق مالکیت فکری ضمن اعطای حقوق انحصاری به پدیدآورندگان، ضرورت ایجاد تعادل میان منافع آنان و نیازهای جامعه برای دسترسی به دانش و نوآوری را به‌رسمیت شناخته است. این استثناها که ناظر بر مواردی مانند استفادهٔ آموزشی، پژوهشی یا کتابخانه‌ای‌اند، باید در چارچوب قواعدی دقیق اعمال شوند تا به اصل حمایت از آثار لطمه نزنند. آزمون سه‌گام[1] سنگ بنای ارزیابی استثناها و محدودیت‌های حقوق کپی‌رایت در سطح بین‌المللی است که ریشه در مادهٔ ۹ (۲) کنوانسیون برن دارد و سپس در مادهٔ ۱۳ موافقت‌نامهٔ تریپس (TRIPS) از سوی سازمان تجارت جهانی تعمیم یافت (World Trade Organization, 1994, Art. 13; Berne Convention, 1979, Art. 9(2)). این آزمون یک معیار تجمعی و اجباری برای اعضا وضع می‌کند که به‌موجب آن، محدودیت‌ها یا استثناها باید منحصراً در الف) موارد خاص معین؛ ب) بدون تعارض با بهره‌برداری عادی از اثر؛ ج) و بدون خدشهٔ ناروا به منافع مشروع پدیدآورنده مجاز شمرده شوند.

---

1. 'Three-Step Test'

در زمینهٔ داده‌های آموزشی هوش مصنوعی، مهم‌ترین چالش در انطباق با این معیارها، تعیین این است که آیا استفادهٔ انبوه از آثار دارای کپی‌رایت برای آموزش مدل‌های زبانی بزرگ، در واقع با بازار مجوزدهی و بهره‌برداری عادی از آن آثار (مانند بازار صدور لیسانس برای استخراج متن و داده) تعارض پیدا کرده، منافع اقتصادی پدیدآورنده را تضعیف می‌نماید (Koos, 2025: 12-13). اجرای گام دوم آزمون، یعنی عدم تعارض با بهره‌برداری عادی، اغلب دشوارترین مرحله در ارزیابی کپی‌رایت در آموزش هوش مصنوعی است. بهره‌برداری عادی به هرگونه استفاده‌ای اطلاق می‌شود که در چارچوب فعالیت‌های اقتصادی معمول و منطقی پدیدآورنده برای کسب سود از اثرش قرار می‌گیرد. با توسعهٔ مدل‌های زبانی بزرگ و ظهور نیاز شرکت‌های فناوری به حجم وسیعی از داده‌ها، یک بازار بالقوهٔ مجوزدهی برای استفاده از آثار دارای کپی‌رایت جهت داده‌کاوی شکل گرفته است؛ بنابراین، هرگونه استفادهٔ بدون مجوز از این آثار ممکن است به‌عنوان تعارض با این بهره‌برداری عادی نوظهور تلقی شود؛ به‌ویژه اگر هدف از آموزش، ایجاد خروجی‌هایی باشد که درنهایت جایگزین اثر اصلی در بازار شوند (Koos, 2025: 16). از سوی دیگر، گام سوم بر حفظ منافع مشروع تمرکز دارد، که فراتر از منافع اقتصادی، شامل منافع اخلاقی و توانایی پدیدآورنده در کنترل توزیع و کیفیت اثر خود نیز می‌شود. در فقدان جبران خسارت یا مکانیسم‌های کنترلی، استفادهٔ غیرمجاز در مقیاس وسیع می‌تواند مستقیماً ناقض این گام باشد (Calderón, 2024: 13-14).

۵. ۲. ۲. اتحادیهٔ اروپا

اتحادیهٔ اروپا به‌عنوان یکی از طرف‌های اصلی کنوانسیون برن و موافقت‌نامهٔ تریپس، آزمون سه‌گام را برای بررسی اعمال محدودیت بر حقوق انحصاری پدیدآورنده در مقررات خود قرار داده است. بر اساس مادهٔ ۲ دستورالعمل EC/۲۹/۲۰۰۱ اتحادیهٔ اروپا، پدیدآورندگان از حق انحصاری برای تکثیر آثار خود برخوردارند؛ چه این تکثیر مستقیم یا غیرمستقیم، موقت یا دائم، کامل یا جزئی و در هر قالب یا رسانه‌ای باشد. مادهٔ ۵ (۱) این حق را محدود می‌کند و استثناهایی برای تکثیر موقت یا اتفاقی در فرایندهای فناورانه قائل می‌شود؛ مشروط بر اینکه استفادهٔ مجاز باشد، اهمیت اقتصادی مستقل نداشته باشد و به بازار اثر اصلی آسیب نرساند (Schönberger, 2018: 16-17). این رهیافت، مشابه دکترین استفادهٔ منصفانه در امریکا است (Sobel, 2020: 25)، اما شرط «گذرا یا اتفاقی بودن» را اضافه می‌کند. این استثنا می‌تواند برای استخراج متن و داده صدق کند، زیرا داده‌ها در حین پردازش موقتاً در حافظهٔ سیستم ذخیره و پس از اتمام کار حذف می‌شوند. با این حال، نگهداری داده‌ها تا پایان پردازش– که برای تأیید، ادغام با داده‌های جدید یا تحلیل‌های تکمیلی ضروری است– مشمول این معافیت نمی‌شود (Triaille, de Meeûs d'Argenteuil, & de Francquen, 2014: 43).

محدودیت‌های مادهٔ ۵ (۱)، به‌ویژه برای شرکت‌های انتفاعی که استفادهٔ آنها اهمیت اقتصادی دارد، رقابت‌پذیری اتحادیهٔ اروپا را تضعیف می‌کرد (Geiger et al., 2018: 17). این موضوع یکی از دلایل اصلی تصویب «دستورالعمل کپی‌رایت در بازار دیجیتال واحد» در سال ۲۰۱۹ بود. مادهٔ ۳ این دستورالعمل به سازمان‌های پژوهشی و مؤسسات میراث فرهنگی اجازه می‌دهد آثار دارای کپی‌رایت را که به‌طور قانونی در دسترس‌اند، برای استخراج متن و داده تکثیر و استخراج کنند؛ بدون نیاز به موقتی بودن، به شرطی که داده‌ها به‌صورت ایمن نگهداری شوند و صرفاً برای اهداف علمی همچون اعتبارسنجی نتایج یا پشتیبانی پژوهش، مورد استفاده قرار گیرند. اما این استثنا تنها «حق تکثیر» و «حق استخراج» را شامل می‌شود و «حق عرصه یا در دسترس قرار دادن» را پوشش نمی‌دهد. بنابراین، انتشار عمومی نسخه‌های آموزشی که تأیید و تکرار فرایندها را با دشواری مواجه می‌سازد، ممنوع است؛ مگر آنکه پژوهشگران تمام جزئیات پیش‌پردازش داده‌ها یا کدهای منبع را ارائه کنند (Franceschelli & Musolesi, 2022: 6-9).

در مقابل، مادهٔ ۴ همین دستورالعمل، استثنای استخراج داده برای استفاده‌های تجاری را نیز به‌رسمیت شناخته، اما آن را به شرط اعلام صریح حق رزرو از سوی دارندگان کپی‌رایت مجاز دانسته است. این رزرو می‌تواند از طریق شرایط استفاده، متادیتا یا سایر ابزارهای مناسب صورت گیرد. با این حال، اثربخشی این مکانیسم در عمل محل تردید جدی است. بسیاری از توسعه‌دهندگان مدل‌های زبانی بزرگ از داده‌هایی استفاده می‌کنند که به‌صورت انبوه و بدون مجوز از اینترنت استخراج شده‌اند و فاقد هرگونه نشانه‌گذاری مناسب برای شناسایی رزرو حقوق هستند. در نتیجه، تلاش ناشران و دیگر صاحبان حقوق برای اِعمال رزرو معمولاً بی‌ثمر است (Kowala, 2023: 6-8).

مادهٔ ۵ (۱) برای شرکت‌های انتفاعی قابل اعمال نیست، زیرا استفاده باید صرفاً غیرتجاری باشد و این یکی از دلایل اصلی اصلاحات سال ۲۰۱۹ بود. همچنین، مادهٔ ۳ اگرچه نوآوری علمی را تسهیل می‌کند، اما چون حق «در دسترس قرار دادن» آثار را پوشش نمی‌دهد، پژوهشگران نمی‌توانند داده‌ها یا نتایج مطالعات را آزادانه با دیگران به اشتراک بگذارند و همین موضوع انجام دوباره و بررسی مستقل پژوهش‌ها را دشوار می‌سازد. مادهٔ ۴ با اعطای حق رزرو به دارندگان کپی‌رایت، تعادلی بین حقوق مالکیت فکری و نوآوری ایجاد می‌کند. از نظر اخلاقی، این حق رزرو ضروری است، زیرا به پدیدآورندگان اختیار تصمیم‌گیری دربارهٔ استفاده از آثارشان را می‌دهد. توسعه‌دهندگان خصوصی باید سه مرحله را رعایت کنند: ۱) دسترسی قانونی به آثار؛ ۲) اطمینان از عدم رزرو حق استخراج متن و داده؛ ۳) نگهداری نسخه‌ها فقط تا زمان لازم. مرحلهٔ دوم دشوار است، زیرا سازوکار مشخصی برای آگاهی از رزرو حقوق وجود ندارد. در بند ۱۸ مقدمهٔ دستورالعمل این‌گونه پیشنهاد شده است که دارندگان کپی‌رایت می‌توانند ارادهٔ خود مبنی بر رزرو حق استخراج متن و داده را در مورد آثاری که به‌صورت آنلاین در دسترس عموم قرار دارند، از طریق ابزارهای ماشین‌خوان مانند ابرداده و شروط مندرج در وب‌سایت اعلام کنند و برای

سایر موارد، از اعلامیه‌های یک‌طرفه یا توافق‌های قراردادی استفاده شود. با این حال، نبود یک سازوکار استاندارد، توسعه‌دهندگان هوش مصنوعی را در شناسایی این محدودیت‌ها با ابهام مواجه می‌سازد (Franceschelli & Musolesi, 2022: 7).

دستورالعمل ۲۰۱۹ راه‌حلی متعادل بین حمایت از حقوق پدیدآورندگان و ترویج نوآوری ارائه می‌دهد، اما چالش‌های عملی دارد. مادهٔ ۳ تحقیقات علمی را تقویت می‌کند، اما محدودیت‌هایش تأیید نتایج را دشوار می‌سازد. مادهٔ ۴ خلاقیت بخش خصوصی را حمایت کرده، وابستگی به عدم رزرو حقوق، اجرا را پیچیده می‌نماید. این مسائل باید پیش از اجرای کامل استخراج متن و داده رفع شوند تا توسعه‌دهندگان بتوانند با اطمینان عمل کنند و توازن میان حقوق فکری و پیشرفت فناورانه حفظ گردد. عدم وضوح در شناسایی رزرو حقوق و الزام به روش‌های ماشینی، پرسش‌هایی است که پاسخ به آنها برای کارایی این چارچوب ضروری است (Rosati. 2019: 23).

۵. ۲. ۳. امریکا

طبق قانون ایالات متحده، استفاده از آثار دارای کپی‌رایت ممکن است تحت دکترین «استفادهٔ منصفانه» مجاز باشد (17 U.S.C. § 107). این قانون چهار معیار را برای ارزیابیِ منصفانه بودنِ استفاده تعیین کرده است: الف) هدف و ماهیت استفاده؛ ب) ماهیت اثر دارای کپی‌رایت؛ پ) میزان و اهمیت بخش استفاده‌شده؛ ت) تأثیر استفاده بر بازار اثر اصلی. این معیارها امکان تصمیم‌گیری موردی را برای دادگاه‌ها فراهم می‌سازد. با این حال، این رویکرد مورد انتقاد بوده، زیرا نیازمند بررسی پرونده‌به‌پرونده است؛ چه‌بسا یک معیار بیش از حد برجسته شود، درحالی که ارزیابی باید با توجه به تمامی عوامل انجام گیرد (Nimmer, 2003: 263-287). با وجود این، دکترین استفادهٔ منصفانه مزایایی دارد؛ ازجمله تعادلی که بین تشویق خلق آثار جدید و افزایش دسترسی عمومی به آثار موجود برقرار می‌کند[1]. این دکترین مواردی را که انحصار کپی‌رایت سودی برای جامعه ندارد، از شمول آن خارج می‌کند (Lunney, 2002: 975). با وجود غیرقابل پیش‌بینی به‌نظر آمدن پرونده‌های استفادهٔ منصفانه، این پرونده‌ها معمولاً از انسجام لازم برخوردارند و در برخی موارد، به شاخصی برای تصمیم‌گیری دادگاه‌ها بدل می‌شوند (Samuelson, 2008: 68-73).

برای بررسی اینکه آیا آموزش داده در هوش مصنوعی می‌تواند تحت استفادهٔ منصفانه قرار گیرد، معیارها باید تحلیل شوند. در معیار «هدف و ماهیت استفاده»، این موضوع سنجیده می‌شود که آیا کاربرد موردنظر جنبهٔ تحول‌آفرین دارد یا نه؛ یعنی آیا عنصر جدیدی به اثر اصلی اضافه می‌کند یا خیر. اگر استفاده برای هدفی متفاوت با هدف اصلی اثر باشد (مانند اهداف بیانی جدید)، احتمال پذیرش آن

1. "Sony Corp. v. Universal, 1984"

به‌عنوان استفادهٔ منصفانه بیشتر است (Asay, Sloan, & Sobczak, 2020: 14-18). در معیار «ماهیت اثر»، باید دید اثر تا چه حد خلاقانه یا واقعی است، زیرا آثار خلاقانه‌تر معمولاً حفاظت بیشتری دارند (Asay, Sloan, & Sobczak, 2020: 14-18). معیار «میزان و اهمیت استفاده» در یادگیری ماشین متفاوت عمل می‌کند. اگرچه یادگیری عمیق مولد ممکن است از کل اثر استفاده کند (مثلاً با تقسیم رمان‌ها یا موسیقی به بخش‌ها)، اما این استفاده بیشتر بر مفاهیم، حقایق و الگوهای موجود در داده‌ها متمرکز است، نه بیان اصلی اثر (Margoni & Kretschmer, 2022: 5-6). ازآنجا که کپی‌رایت ناظر بر شیوهٔ بیان است و نه بر مفاهیم یا روش‌ها، استخراج داده در اغلب موارد نقض این حقوق به‌شمار نمی‌رود. حقوق‌دانان معتقدند استفاده‌های غیربیانی (مانند استخراج اطلاعات سطح بالا) باید منصفانه تلقی شوند، زیرا بیان اصلی اثر را هدف قرار نمی‌دهند (Sag, 2018: 47-48).

در معیار «تأثیر بر بازار»، آموزشِ داده معمولاً تأثیر کمی بر بازار اثر اصلی دارد. در این فرایند، هر اثر در کنار مجموعه‌ای از آثار دیگر به‌کار می‌رود و خروجی‌ها اغلب با آثار خاص شباهت چندانی ندارند، بلکه ویژگی‌های منحصربه‌فرد خود را نمایان می‌سازند (Franceschelli & Musolesi, 2022: 4-5). این امر به‌ویژه در مجموعه‌های آموزشی ناهمگن صادق است؛ اما اگر مجموعه‌داده از آثار محدودی از یک نویسنده تشکیل شود، این استدلال ضعیف‌تر می‌شود. برقراری ارتباط بین خروجی و آثار آموزشی دشوار است، لذا تأثیر بر بازار معمولاً کم‌اهمیت تلقی می‌شود.

دکترین استفادهٔ منصفانه، علی‌رغم ظرفیت بالقوهٔ خود برای حمایت از نوآوری، در حوزهٔ یادگیری ماشین با چالش‌های فزاینده‌ای روبه‌رو شده است. پیش‌تر، ماهیت تحول‌آفرین فرایند استخراج متن و داده موجب شد که این دکترین در پرونده‌هایی مانند *Google Books* و *HathiTrust* به نفع پروژه‌های فناورانه تفسیر شود[1]. اما در زمینهٔ مدل‌های زبانی مولد T به‌ویژه زمانی که خروجی‌ها دارای ویژگی‌های بیانی و مشابه با آثار دارای کپی‌رایت هستند، وضعیت حقوقی پیچیده‌تر می‌شود. استفاده‌هایی که از انتخاب‌های خلاقانهٔ مؤلفان در آموزش مدل بهره می‌برند، ممکن است تحت بررسی سخت‌گیرانه‌تری قرار گیرند. رأی اخیر در پروندهٔ *باترز علیه آنتروپیک* نشان می‌دهد که دادگاه‌ها در حال بازتعریف حدود استفادهٔ منصفانه در زمینهٔ هوش مصنوعی هستند و استفاده از آثار دزدیده‌شده، حتی با اهداف تحقیقاتی را غیرقابل‌توجیه می‌دانند. بنابراین، هرچند استفادهٔ منصفانه می‌تواند بستری برای نوآوری باشد، اما نیازمند تفسیر دقیق، شفاف‌سازی قضایی و رعایت حقوق پدیدآورندگان است (Bonadio & McDonagh, 2020: 13-14)[2].

رأی اخیر صادره از دادگاه امریکایی که سابقاً اشاره شد، نمونه‌ای عملی از به‌کارگیری این چهار معیار به‌شمار می‌رود. دادگاه به‌صراحت اعلام کرد که هدف از ذخیره‌سازی دائمی داده‌ها در کتابخانهٔ دیجیتال

1. "Authors Guild v. Google, 2015; Authors Guild v. HathiTrust, 2014"
2. "Bartz v. Anthropic PBC, 2025"

شرکت آنتروپیک، فاقد ویژگی تغییر ماهیت است و بنابراین ذخیرهٔ آثار دارای کپی‌رایت بدون رضایت مؤلف، حتی با اهداف پژوهشی یا توسعه‌ای، توجیه‌پذیر نیست[1]. این رأی تأکیدی است بر اینکه هرچند آموزش مدل‌های هوش مصنوعی می‌تواند مصداق استفادهٔ منصفانه باشد، اما این توجیه نمی‌تواند نقض سیستماتیک حقوق مؤلف را موجه جلوه دهد.

در تحولی جدید، در ژوئیه ۲۰۲۵، دو نمایندهٔ مجلس سنای امریکا، لایحهٔ دوحزبی «قانون مسئولیت‌پذیری هوش مصنوعی و حفاظت از داده‌های شخصی» را ارائه کردند (Hawley & Blumenthal, 2025). هدف اصلی این لایحه، ممنوعیت استفادهٔ بدون مجوز از آثار دارای کپی‌رایت و داده‌های شخصی در فرایند آموزش مدل‌های هوش مصنوعی و ایجاد حق شکایت در دادگاه‌های فدرال برای افرادی است که آثار خلاقانه یا اطلاعاتشان بدون کسب رضایت و مجوز صریح به‌کار گرفته شده است. این قانون، با تمرکز بر حمایت از پدیدآورندگان آثار خلاقانه و ادبی و هنری، در راستای ترسیم چارچوب شرایط استفادهٔ منصفانه، شرکت‌های فناوری بزرگ را ملزم به شفاف‌سازی در خصوص اشخاص ثالث دارای دسترسی به داده‌ها می‌کند و سازوکارهایی چون شکایت فردی و گروهی، دستورهای موقت، و مجازات‌های مالی سنگین را برای مقابله با نقض کپی‌رایت و سوءاستفاده از داده‌ها پیش‌بینی کرده است (Hawley & Blumenthal, 2025). هرچند به دلیل وجود اختلافات گسترده در دیدگاه‌ها و به‌ویژه تأکیداتی که بر توسعهٔ پرشتاب فناوری هوش مصنوعی می‌شود، نمی‌توان سرنوشت لایحه را از هم‌اکنون پیش‌بینی کرد، اما سند یادشده از یک سو، نشان‌دهندهٔ رهیافت‌های جدی است که به‌دنبال حفظ حقوق پدیدآورندگان آثار است و از سوی دیگر قابل کتمان نیست که می‌تواند مانعی بر سر توسعهٔ هوش مصنوعی باشد.

۵. ۲. ۴. ایران

در قوانین ایران، مفهومی تحت عنوان «استفادهٔ منصفانه» مانند آنچه در نظام‌های کامن‌لا، به‌ویژه امریکا، پذیرفته شده، به‌صورت صریح تعریف نشده است. با این حال، تحولات اخیر، به‌خصوص «سند ملی هوش مصنوعی جمهوری اسلامی ایران» مصوب ۱۴۰۳ و «سند ستاد توسعهٔ فناوری و کاربرد هوش مصنوعی» مصوب ۱۴۰۴، به‌طور ضمنی بر خلأ قانونی موجود در این حوزه صحه گذاشته و اصلاح قوانین مالکیت فکری و تسهیل دسترسی به داده‌ها برای آموزش مدل‌ها را به‌عنوان راهبردهای کلیدی تعیین کرده‌اند (مواد ۳ و ۵ سند ملی). قانون حمایت از حقوق مؤلفان، مصنفان و هنرمندان مصوب ۱۳۴۸، و قانون ترجمه و تکثیر کتب و نشریات و آثار صوتی مصوب ۱۳۵۲، عمدتاً بر حمایت از حقوق صاحبان اثر تأکید دارند و صرفاً موارد معدودی از استفادهٔ بدون رضایت صاحب اثر را مجاز شمرده‌اند. در قانون حمایت از

1. Ibid.

حقوق مؤلفان، مصنفان و هنرمندان، در مواد ۱ و ۲ به ترتیب، پدیدآورنده را تعریف کرده، دامنهٔ آثار مشمول حمایت را شامل کتاب، رساله، موسیقی، نقاشی و دیگر آثار ادبی و هنری می‌داند؛ و در مادهٔ ۳ حقوق انحصاری پدیدآورندگان در زمینه‌هایی مانند تکثیر، اجرا و عرضهٔ عمومی را تعیین کرده است. با این حال، استثناهایی محدود در مواد ۷ تا ۱۱ ذکر شده‌اند، مانند نقل قول کوتاه با ذکر منبع (مادهٔ ۷) یا استفادهٔ شخصی (مادهٔ ۱۱) که عمدتاً برای اهداف غیرتجاری و خاص مانند آموزش یا نقد است. این استثناها مشابه دکترین «استفادهٔ منصفانه» در ایالات متحده یا استثناهای استخراج متن و داده در اتحادیهٔ اروپا نیستند، زیرا به استفادهٔ گسترده از آثار برای آموزش مدل‌های هوش مصنوعی اشاره‌ای ندارند و دامنهٔ آنها بسیار محدودتر است.

برای بررسی امکان شمول داده‌های آموزشی هوش مصنوعی در استثنای «استفادهٔ منصفانه» در ایران، ابتدا باید به خلأ قانونی موجود توجه کرد. برخلاف ایالات متحده که معیارهایی چون هدف، ماهیت، میزان و اثر بر بازار را در مادهٔ ۱۰۷ قانون کپی‌رایت بررسی می‌کند، یا اتحادیهٔ اروپا که در مادهٔ ۳ دستورالعمل ۲۰۱۹ استثنای خاصی برای استخراج متن و داده پیش‌بینی کرده است، قوانین ایران فاقد چنین معیارهایی هستند. استفاده از آثار دارای کپی‌رایت برای آموزش مدل‌ها- که غالباً مستلزم تکثیر کامل آثار و استخراج الگوها بدون تولید بیان جدید است- با مادهٔ ۳ قانون ۱۳۴۸ دربارهٔ «حق انحصاری تکثیر» در تضاد قرار می‌گیرد. استثنای مادهٔ ۱۱ دربارهٔ «استفادهٔ شخصی» نیز قابل اعمال نیست، زیرا آموزش مدل‌ها ماهیتی تجاری یا پژوهشی دارد. مادهٔ ۷ که استفادهٔ آموزشی را مجاز می‌داند، ناظر به مقاصد عادی علمی، ادبی و فنی است و همچنین به شرایطی مانند ذکر منبع و عدم آسیب به بهره‌برداری عادی اثر محدود می‌شود، که با مقیاس وسیع داده‌های لازم برای آموزش هوش مصنوعی همخوانی ندارد (Shakeri & Esmaili, 2017: 9-10). به‌علاوه، ابهام در مفهوم «حدود متعارف» استفادهٔ غیرانتفاعی، موجب تردید در شمول آموزش مدل‌های هوش مصنوعی می‌شود.

با وجود فقدان مقررات صریح در ایران دربارهٔ استفاده از آثار دارای کپی‌رایت در فرایند آموزش مدل‌های هوش مصنوعی، دو دیدگاه حقوقی متعارض قابل طرح است: یک دیدگاه می‌تواند استدلال کند که با توجه به ماهیت دگرگون‌کنندهٔ آموزش مدل‌ها و هدف پژوهشی/ غیرتجاری توسعهٔ مدل‌های بومی، این فعالیت‌ها باید با الهام از دکترین‌های حقوق در نظام‌های حقوقی پیشرفته (مانند استفادهٔ منصفانه) و با تکیه بر تعادل میان حقوق پدیدآورندگان و منافع عمومی در دسترسی به نوآوری، مورد حمایت قرار گیرند. دیدگاه حقوقی دیگر، بر اصل انحصاری بودن حقوق پدیدآورنده تأکید دارد. فقدان تعریف «استفادهٔ منصفانه» و تأکید صریح قانون بر حقوق انحصاری تکثیر (مادهٔ ۳) و جرم بودن نقض کپی‌رایت (مادهٔ ۲۳)، نشان می‌دهد که استفاده از آثار دارای کپی‌رایت برای آموزش هوش مصنوعی بدون مجوز، به احتمال بسیار زیاد در تضاد با قوانین موجود است. این چالش‌ها محدود به مسئلهٔ داده‌های آموزشی نبوده،

در ابعاد گسترده‌تری، ازجمله حمایت از فناوری‌های ناملموس هوش مصنوعی مانند الگوریتم‌ها، ذیل قوانین حق اختراع و حق مؤلف ایران نیز دیده می‌شوند (Dehghanpour & Rahbar, 2022: 14-16). برای رفع این ابهام و جلوگیری از تضعیف حقوق مؤلفان، اصلاح قوانین یا تدوین مقررات خاص، یک ضرورت حقوقی و فناورانه محسوب می‌شود. این امر نه‌تنها برای برقراری توازن میان نوآوری در هوش مصنوعی و حقوق مالکیت فکری حیاتی است (Shakeri, 2025: 53-54)، بلکه مستقیماً با اهداف سند ملی هوش مصنوعی ۱۴۰۳ برای تبدیل ایران به قطب تحقیقاتی منطقه و رشد تولیدات فکری همخوانی دارد (Sadeghi & Javadi Parsa, 2025: 18-22). بنابراین، توسعهٔ مدل‌های بومی هوش مصنوعی بدون تسهیل قانونی فرایند آموزش آنها و تعیین حدود رعایت کپی‌رایت، با چالش‌های جدی حقوقی روبه‌رو خواهد بود.

۶. راهکارهای ممکن برای توازن میان حمایت از کپی‌رایت و پیشبرد هوش مصنوعی

این قسمت به راهکارهایی برای ایجاد توازن میان حمایت از حقوق کپی‌رایت و توسعهٔ هوش مصنوعی می‌پردازد. همان‌طور که گفته شد، با پیچیده‌تر شدن مدل‌های هوش مصنوعی، تقاضا برای حجم عظیمی از داده‌های آموزشی که اغلب شامل محتوای دارای کپی‌رایت هستند، افزایش می‌یابد. پیشرفت هوش مصنوعی مولد مستلزم شفافیت و عدالت بیشتر در دسترسی به داده‌هاست. منظور از این مهم، برقراری توازن میان دو هدف است: ۱) تضمین جبران خسارت منصفانه و حق کنترل برای صاحبان حقوق مالکیت فکری؛ ۲) تسهیل دسترسی پژوهشی و نوآورانه به داده‌ها برای جامعه و توسعه‌دهندگان. این امر نیازمند راهکارهایی است که هم به تقویت نوآوری بینجامد و هم از حقوق مالکیت فکری صیانت کند. در ادامه، راه‌حل‌های پیشنهادی بررسی می‌شوند.

الف) پایگاه‌های دادهٔ منبع‌باز. ایجاد پایگاه‌های دادهٔ متن‌باز برای آموزش هوش مصنوعی، خلاقیت و همکاری را تقویت می‌کند. این پایگاه‌ها با دسترسی آزاد به آثار بدون کپی‌رایت، به پژوهشگران، برنامه‌نویسان و کسب‌وکارها اجازه می‌دهند بدون نگرانی از نقض حقوق، از دانش جمعی بهره ببرند. این رویکرد برای کسب‌وکارهای کوچک و پدیدآورندگان مقرون‌به‌صرفه است و شفافیت، قطعیت حقوقی و توسعه را به ارمغان می‌آورد. پروژه‌هایی مانند ویکی‌پدیا، کریتیو کامنز، ویکی‌مدیا، فری‌پیک و پیکسابای[1] نشان‌دهندهٔ موفقیت این مدل‌اند. این پایگاه‌ها با فراهم کردن دسترسی به مجموعه‌ای متنوع از آثار، می‌توانند منابعی ارزشمند برای توسعه‌دهندگان باشند. با این حال، چالش‌هایی مانند مدیریت به‌روزرسانی وضعیت کپی‌رایت، تأیید اصالت آثار و حفظ تنوع پایگاه داده وجود دارد. این پایگاه‌ها جایگزین کامل

1. "Wikipedia, Creative Commons, Wikimedia Commons, Freepik, and Pixabay"

مجوزهای تجاری نیستند، بلکه مکمل آنها برای توانمندسازی ذی‌نفعان‌اند و نیازمند خط‌مشی‌های روشن و سازوکارهای مؤثر قابل اتکا خواهند بود (OECD, 2019: 9).

ب) توافق‌نامه‌های اشتراک داده. انعقاد قراردادهای جامع بین توسعه‌دهندگان هوش مصنوعی و ارائه‌دهندگان داده ضروری است. این توافق‌نامه‌ها می‌توانند استفادهٔ مجاز و محدودیت‌ها را مشخص کنند. با تعریف شفاف مالکیت، کسب رضایت صریح و بازنگری مستمر برای تطابق با قوانین، این قراردادها را به ابزاری برای حفظ تعادل حقوقی تبدیل می‌کند. بهترین روش‌ها شامل محدود کردن استفاده به اهداف مشخص، اجرای تدابیر محرمانگی و تضمین مذاکرات حسب منافع طرفین است. این رویکرد انطباق با مقررات کپی‌رایت را تقویت و از سوءاستفاده جلوگیری می‌کند، ضمن اینکه همکاری بین طرفین را تسهیل می‌نماید (Leistner & Antoine, 2022: 31-33).

ج) جبران خسارت برای پدیدآورندگان. سازوکارهای جبران خسارت، ارزش محتوای دارای کپی‌رایت را حفظ و برجسته می‌کند. در این راستا، مدل‌های جبران مالی نظیر پرداخت حق امتیاز، تقسیم درآمد و سایر مکانیزم‌های مالی می‌توانند پیاده‌سازی شوند تا پدیدآورندگان آثار دارای کپی‌رایت بتوانند سهم منصفانه‌ای از منافع اقتصادی حاصل از کاربرد آثارشان در آموزش هوش مصنوعی دریافت نمایند. این مدل‌ها می‌توانند انگیزه‌ای برای تولیدکنندگان محتوا ایجاد کنند تا آثار خود را در اختیار توسعه‌دهندگان هوش مصنوعی قرار دهند و اعتماد میان طرفین تقویت شود (Nyambura, 2023).

د) بیمه کردن سیستم‌های هوش مصنوعی. بیمه در برابر نقض کپی‌رایت، ریسک‌های مالی را کاهش می‌دهد. این ابزار با تشویق رفتار مسئولانه، زمینهٔ ادغام ایمن هوش مصنوعی را فراهم کرده، زیان‌های احتمالی را پوشش می‌دهد. عواملی مانند نوع هوش مصنوعی، ریسک‌ها و هویت بیمه‌گذار (توسعه‌دهنده یا کاربر) در ترسیم سازوکارها مهم‌اند. بیمه به‌عنوان شبکهٔ ایمنی، خاطر توسعه‌دهندگان را از نگرانی‌های حقوقی آسوده کرده، از پدیدآورندگان محافظت می‌نماید. در این راستا، پیشنهاد شده است که واحدهایی تخصصی برای پوشش بیمه در شرکت‌های توسعه‌دهندهٔ هوش مصنوعی راه‌اندازی شود (Lior, 2021: 63-64). گزارش اتحادیهٔ اروپا بیمهٔ اجباری برای ربات‌ها را مشابه بیمهٔ خودرو پیشنهاد کرده است که می‌تواند الگویی برای هوش مصنوعی باشد. این رویکرد می‌تواند توسعهٔ فناوری را بدون تهدید حقوق مالکیت فکری تضمین کند (European Parliament, 2017).

ه) مراکز مجوزدهی تخصصی. مراکز یا پلتفرم‌های متمرکز می‌توانند فرایند أخذ مجوز برای استفاده از محتوای دارای کپی‌رایت را ساده‌تر کرده، امکان مذاکره بهتر میان پدیدآورندگان آثار و توسعه‌دهندگان هوش مصنوعی را فراهم سازند. چنین مراکزی می‌توانند شفافیت بیشتری در نحوهٔ استفاده از داده‌های آموزشی ایجاد کرده، حل اختلافات را تسهیل و اطمینان حاصل نمایند که حقوق پدیدآورندگان آثار رعایت می‌شود (Kop, 2020: 5).

و) قوانین اخلاقی و استانداردهای صنعتی. تدوین استانداردهای دقیق برای آموزش هوش مصنوعی، توسعهٔ مسئولانه را ترویج می‌کند. این قواعد باید بر جبران عادلانه، رضایت آگاهانه، شفافیت جمع‌آوری داده و فرایند آموزش تمرکز کند. پایبندی به این اصول، ضمن اینکه چارچوبی اخلاقی برای صنعت فراهم می‌آورد، هم به جلب اعتماد کمک می‌کند و هم از حقوق پدیدآورندگان پشتیبانی می‌نماید (European Commission, 2021).

با وجود ظرفیت‌های راهکارهای پیشنهادی، اجرای آنها در مقیاس جهانی با چالش‌های بنیادین، ازجمله حجم عظیم داده‌های آموزشی، تعدد و ناشناخته بودن صاحبان حقوق و عدم تقارن قدرت میان شرکت‌های بزرگ فناوری و پدیدآورندگان فردی روبه‌رو است. هیچ‌یک از این روش‌ها به‌تنهایی کافی نیست؛ برای مثال، توافق‌نامه‌های فردی در برابر میلیون‌ها صاحب حق، کارایی ندارند و مدل‌های جبران خسارت بدون ردیابی دقیق استفاده از محتوا، عملی نمی‌شوند. موفقیت در گرو اجرای ترکیبی و نهادی است، ازجمله مراکز مجوزدهی تخصصی برای تجمیع حقوق و تسهیل مذاکره؛ استانداردهای شفافیت صنعتی برای ردیابی منبع داده؛ نظام‌های جبران مالی خودکار، مانند «حق‌الامتیاز داده» یا صندوق مشترک با مشارکت اجباری شرکت‌های بزرگ فناوری. این چارچوب‌های چندلایه‌ای، عدالت در دسترسی به داده‌ها، یعنی جبران مالی متناسب، دسترسی برابر و شفافیت را عملیاتی می‌کنند و توازن میان نوآوری هوش مصنوعی و حمایت از حقوق مالکیت فکری را بدون حذف یکی از دو قطب، تضمین می‌نمایند.

۷. نتیجه‌گیری

مدل‌های زبان بزرگ مولد، مانند چت‌جی‌پی‌تی که با حجم وسیعی از داده‌های دارای کپی‌رایت آموزش می‌بینند، چالش‌های حقوقی نوظهور و بنیادینی را در عرصهٔ مالکیت فکری ایجاد کرده‌اند. توانایی این مدل‌ها در استخراج الگوها و بازتولید محتوای مشابه، ضرورت بازتعریف حدود حقوق انحصاری پدیدآورندگان را برجسته کرده است. تحلیل تطبیقی این پژوهش نشان داد که دو رویکرد اصلی حقوقی در سطح جهان برای ارزیابی قانونی بودن این استفاده وجود دارد: دکترین استفادهٔ منصفانه در ایالات متحده و آزمون سه‌گام. آزمون سه‌گام ریشه در مادهٔ ۹ (۲) کنوانسیون برن دارد و سپس از طریق مادهٔ ۱۳ موافقت‌نامهٔ تریپس تعمیم یافت و مبنای قوانین اتحادیهٔ اروپا قرار گرفت. نوآوری این پژوهش آن است که در پرتو این تفاوت‌های ماهوی، نظام حقوقی ایران باید بین مدل انعطاف‌پذیر استفادهٔ منصفانه (حامی نوآوری) و مدل قاعده‌محور آزمون سه‌گام (حامی حقوق مؤلف) دست به یک انتخاب استراتژیک بزند.

در ایالات متحده، رویکرد انعطاف‌پذیر استفادهٔ منصفانه، در پرونده‌هایی مانند *Google Books*، با محوریت ماهیت تحول‌آفرین، یک مشوق قوی برای نوآوری ارائه می‌دهد، اما عدم قطعیت قضایی و

ریسک بالا (همانند پروندهٔ *بارتز علیه آنتروپیک*) از چالش‌های اصلی آن برای توسعه‌دهندگان مدل‌های زبان بزرگ است. در مقابل، اتحادیهٔ اروپا رویکردی قاعده‌محور بر اساس آزمون سه‌گام را اتخاذ کرده است. مادهٔ ۴ دستورالعمل DSM با سازوکار حق رزرو حق استخراج متن و داده، حفاظت از منافع مشروع (گام سوم آزمون سه‌گام) را هدف قرار داده و امنیت حقوقی بالاتری برای مؤلفان فراهم می‌آورد، اما در عین حال، محدودیت‌های اجرایی و شفافیتی آن، توسعهٔ مدل‌های زبان بزرگ را با موانع کندکنندهٔ نوآوری مواجه ساخته است.

در نظام حقوقی ایران، با توجه به خلأ صریح قانونی و نبود استثنایی برای آموزش هوش مصنوعی، رفع این نقیصه نیازمند صرف اصلاح قانون نیست، بلکه مستلزم یک انتخاب استراتژیک ملی است: اگر هدف غایی سیاست‌گذاران، توسعهٔ شتابان و رقابت بین‌المللی در حوزهٔ هوش مصنوعی مولد است، اتخاذ یک مدل انعطاف‌پذیر مشابه استفادهٔ منصفانه (با تمرکز بر تحول‌آفرینی) توجیه پیدا می‌کند؛ اما اگر اولویت، حفظ کامل حقوق پدیدآورندگان و توسعهٔ یک بازار مجوزدهی داخلی است، اتخاذ مدل قاعده‌محور آزمون سه‌گام و تعریف سازوکار حق رزرو ضروری خواهد بود. این دو رویکرد، نتایج متفاوتی را برای زیست‌بوم فناوری و محتوای ایران رقم خواهند زد.

در نهایت، برای مدیریت این چالش‌ها و پیاده‌سازی مدل منتخب، استفاده از راهکارهای ترکیبی نظیر ایجاد پایگاه‌های داده تحت مجوزهای باز، توافق‌نامه‌های اشتراک داده با لیسانس‌های منصفانه و توسعهٔ مدل‌های بیمهٔ ریسک حقوقی، می‌تواند زمینه‌ساز نوآوری مسئولانه و توسعهٔ پایدار در حوزهٔ هوش مصنوعی باشد.

References

- **Books**

1. Foster, D. (2019). *Generative Deep Learning. Teaching Machines to Paint, Write, Compose and Play.* Beijing-Boston-Farnham-Sebastopol-Tokyo, OREILLY.
2. Karapapa, S. (2023). *The Press Publishers' Right under EU Law: Rewarding Investment through Intellectual Property*. Edward Elgar Publishing.
3. Netanel, N. W. (2008). *Copyright's paradox*. Oxford University Press.

- **Journal Articles and Legal Reviews**

4. Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5185–5198). Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/2020.acl-main.463.
5. Bonadio, E., & McDonagh, L. (2020). Artificial intelligence as producer and consumer of copyright works: Evaluating the consequences of algorithmic creativity. *Intellectual Property Quarterly, 2*(2), 112–137. SSRN: https://ssrn.com/abstract=3617197.
6. Bonadio, E., Dinev, P., & McDonagh, L. (2022). Can artificial intelligence infringe

copyright? Some reflections. In *Research handbook on intellectual property and artificial intelligence* (pp. 245-257). Edward Elgar Publishing. http://dx.doi.org/10.4337/9781800881907.00019.

7. Calderón, R. (2024). AI Training Through Copyrighted Works As Infringement: Perspectives Under The Berne Three-step Test and The Pane Fair Use Test and Plausible Solutions. *IDEA-The Law Review of the Franklin Pierce Center for IP*, 84-103.
8. Carroll, M. W. (2019). Copyright and the progress of science: Why text and data mining is lawful. *UC Davis Law Review, 53* (2), 893–940. SSRN: https://ssrn.com/abstract=3531231
9. Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., & Wang, Y. (2023). A survey on evaluation of large language models. *arXiv preprint,* arXiv:2307.03109. http://dx.doi.org/10.1145/3641289.
10. Chiou, T. (2021). Copyright lessons on machine learning: What impact on algorithmic art? *JIPITEC,* (12), 398–427. https://nbn-resolving.de/urn:nbn:de:0009-29-50250.
11. Craig, C. J. (2022). The AI-copyright challenge: Tech-neutrality, authorship, and the public interest. In *Research handbook on intellectual property and artificial intelligence* (pp. 134-155). Edward Elgar Publishing. http://dx.doi.org/10.4337/9781800881907.00013.
12. Dehghanpour Farashah, S. and Rahbar, N. (2022). Intellectual Property Protection for Intangible Technologies in Autonomous Vehicles in Iran, US, and EU: A Comparative Study with the Focus on Algorithms. Comparative Law Review, 13 (2), 531-551. doi: 10.22059/jcl.2022.336887.634298 (In Persian).
13. Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Dwivedi, R., & Balakrishnan, J. (2023). "So what if ChatGPT wrote it?": Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management, 71*, 102642. https://doi.org/10.1016/j.ijinfomgt.2023.102642.
14. Elgammal, A., Liu, K., Elgammal, N., & Elgammal, K. (2017). CAN: Creative adversarial networks, generating "art" by learning about styles and deviating from style norms. *arXiv preprint*, arXiv:1706.07068. http://dx.doi.org/10.48550/arXiv.1706.07068.
15. Floridi, L. (2023). AI as agency without intelligence: On ChatGPT, large language models, and other generative models. *Philosophy & Technology, 36*(1), 15. http://dx.doi.org/10.1007/s13347-023-00621-y.
16. Franceschelli, G., & Musolesi, M. (2022). Copyright in generative deep learning. *Data & Policy, 4*, e17. http://dx.doi.org/10.48550/arXiv.2105.09266.
17. Geiger, C., Luth, A., Rosati, E., & Peukert, A. (2018). The exception for Text and Data Mining (TDM) in the Proposed Directive on Copyright in the Digital Single Market–Legal Aspects. *Centre for International Intellectual Property Studies (CEIPI)*, Research Paper 2018-02. http://dx.doi.org/10.2139/ssrn.3160586.
18. Ginsburg, J. C., & Budiardjo, L. A. (2019). Authors and machines. *Berkeley Technology Law Journal, 34* (2), 343–404. http://dx.doi.org/10.2139/ssrn.3233885.
19. Guadamuz, A. (2017). Do androids dream of electric copyright? Comparative analysis of originality in artificial intelligence generated works. *Intellectual Property Quarterly*,

(3), 209–241. http://dx.doi.org/10.1093/oso/9780198870944.003.0008.
20. Hadi, M. U., Qureshi, R., Shah, A., Irfan, M., Zafar, A., Shaikh, M. B., ... & Mirjalili, S. (2023). Large language models: a comprehensive survey of its applications, challenges, limitations, and future prospects. *Authorea preprints*, *1* (3), 1-26. http://dx.doi.org/10.36227/techrxiv.23589741.v7.
21. Haleem, A., Javaid, M., & Singh, R. P. (2022). An era of ChatGPT as a significant futuristic support tool: A study on features, abilities, and challenges. *BenchCouncil Transactions on Benchmarks, Standards and Evaluations, 2* (4), 100089. http://dx.doi.org/10.1016/j.tbench.2023.100089.
22. Hearst, M. A. (1999). Untangling text data mining. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics* (pp. 3–10). Association for Computational Linguistics. http://dx.doi.org/10.3115/1034678.1034679.
23. Hilty, R. M., & Sutterer, M. (2017). Position Statement of the Max Planck Institute for Innovation and Competition on the Proposed Modernisation of the European Copyright Rules, Part H, Content Circulation in Europe. *Max Planck Institute for Innovation & Competition*. Research Paper 17-02. http://dx.doi.org/10.1093/grurint/ikae058.
24. Koos, S. (2025). AI Model Training and Copyright Law: Legal Challenges and Regulatory Approaches in the EU, US, and Asia. *Indonesian Civil Law Journal*, 1 (2), 55-78. https://hukumperdata.id/ojs/index.php/iclj/article/view/22
25. Kop, M. (2020). Machine learning & EU data sharing practices. *Stanford-Vienna Transatlantic Technology Law Forum, Transatlantic Antitrust and IPR Developments*. SSRN: https://ssrn.com/abstract=3409712.
26. Kowala, M. (2024). Protection of Press Publishers in the Age of Generative AI–In Search of Legal Remedies to Adapt to the Pace of Technology. *IIC-International Review of Intellectual Property and Competition Law, 55* (1), 1-20. http://dx.doi.org/10.1007/s40319-024-01515-y.
27. Leistner, M., & Antoine, L. (2022). IPR and the use of open data and data sharing initiatives by public and private actors. *European Parliament's Policy Department for Citizens' Rights and Constitutional Affairs*. http://dx.doi.org/10.2139/ssrn.4125503.
28. Lemley, M. A., & Casey, B. (2020). Fair learning. *Texas Law Review, (99)*, 743.
29. Lennartz, J., & Kraetzig, V. (2024). Forbidden Fruits? Artistic Creation in the AI Copyright War. *IIC-International Review of Intellectual Property and Competition Law, 55* (1), 1-5. http://dx.doi.org/10.1007/s40319-024-01551-8.
30. Lior, A. (2021). Insuring AI: The role of insurance in artificial intelligence regulation. *Harvard Journal of Law & Technology, 35*(1), 63–64. SSRN: https://ssrn.com/abstract=4266259.
31. Lunney Jr, G. S. (2002). Fair use and market failure: Sony revisited. *Boston University Law Review, 82*(5), 975–1043. SSRN: https://ssrn.com/abstract=301389.
32. Margoni, T., & Kretschmer, M. (2022). A deeper look into the EU text and data mining exceptions: Harmonisation, data ownership, and the future of technology. *GRUR International, 71*(8), 685–701. http://dx.doi.org/10.1093/grurint/ikac054.
33. Netanel, N. W. (2011). Making sense of fair use. *Lewis & Clark Law Review, 15* (3), 715–783. SSRN: https://ssrn.com/abstract=1874778.
34. Nimmer, D. (2003). "Fairest of them all" and other fairy tales of fair use. *Law and Contemporary Problems,* 66 (1/2), 263–287. http://dx.doi.org/10.2307/20059179.

35. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1 (8), 9.
36. Rosati, E. (2018). *The exception for text and data mining (TDM) in the proposed Directive on Copyright in the Digital Single Market: Technical aspects*. European Parliament. https://data.europa.eu/doi/10.2861/480649.
37. Rosati, E. (2019). Copyright as an obstacle or an enabler? A European perspective on text and data mining and its role in the development of AI creativity. *Asia Pacific Law Review, 27* (2), 198–217. http://dx.doi.org/10.1080/10192557.2019.1705525.
38. Roumeliotis, K. I., & Tselikas, N. D. (2023). ChatGPT and Open-AI Models: A Preliminary Review. *Future Internet, 15* (6), 192. http://dx.doi.org/10.3390/fi15060192.
39. Sadeghi,M. and Javadi Parsa,K. (2025). Legal Policymaking in the field of AI in European Union; Suitable Model for Iranian Law. *Journal of Research and Development in Comparative Law*, e729602 doi: 10.22034/law.2025.2065565.1673 (In Persian).
40. Sag, M. (2018). The new legal landscape for text mining and machine learning. *Journal of the Copyright Society of the USA, 66* (2), 291–312. http://dx.doi.org/10.2139/ssrn.3331606.
41. Samuelson, P. (2008). Unbundling fair uses. *Fordham Law Review, 77* (5), 2537–2604. Available at: https://ir.lawnet.fordham.edu/flr/vol77/iss5/16.
42. Savrai, P. (2025). The Challenges Facing the Human-Centered Copyright Regime with the Evolution of Artificial Intelligence. *Economic and Commercial Law Researches*, 2 (3), 11-40. doi: 10.48308/eclr.2025.236827.1078 (In Persian).
43. Schönberger, D. (2018). Deep copyright: Up-and downstream questions related to artificial intelligence (AI) and machine learning (ML). *Droit d'auteur*, (4), 145–173. http://dx.doi.org/10.1628/zge-2018-0003.
44. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal Policy Optimization Algorithms. *arXiv preprint* arXiv:1707.06347. http://dx.doi.org/10.48550/arXiv.1707.06347.
45. Senftleben, M. (2023). Generative AI and author remuneration. *IIC-International Review of Intellectual Property and Competition Law*, *54* (10), 1535-1560. http://dx.doi.org/10.1007/s40319-023-01399-4.
46. Shakeri, Z. (2025). The Literary and Artistic Property Law System in the Age of Artificial Intelligence; Considerations for Policymaking in Future Governance. *Iranian Journal of Public Policy*, 11 (1), 41-55. doi: 10.22059/jppolicy.2025.101189 (In Persian).
47. Shakeri, Z. and Esmaili Yadaki, M. (2017). A Comparative Analysis of the Exception of Educational Use in Literacy and Artistic Property Rights. *Civil Law Knowledge*, *6* (1), 19-30. https://dor.isc.ac/dor/20.1001.1.23221712.1396.6.1.3.0 (In Persian).
48. Sobel, B. (2020). A taxonomy of training data: Disentangling the mismatched rights, remedies, and rationales for restricting machine learning. In R. Hilty, J.-A. Lee, & K.-C. Liu (Eds.), *Artificial Intelligence and Intellectual Property*. Oxford University Press. http://dx.doi.org/10.1093/oso/9780198870944.003.0011.
49. Sobel, B. L. W. (2017). Artificial intelligence's fair use crisis. *Columbia Journal of Law & the Arts, 41* (1), 45. SSRN: https://ssrn.com/abstract=3032076.
50. Strowel, A. (2023). ChatGPT and Generative AI Tools: Theft of Intellectual Labor?.

IIC-International Review of Intellectual Property and Competition Law, 54 (4), 491–494. http://dx.doi.org/10.1007/s40319-023-01321-y.

51. Triaille, J.-P., de Meeûs d'Argenteuil, J., & de Francquen, A. (2014). Study on the legal framework of text and data mining (TDM). *European Union Studies* KM-03-13-426-EN-N. https://data.europa.eu/doi/10.2780/1475.
52. Zhao, W. X., Zhang, H., Ding, L., Wang, Y., Gu, Y., Chen, W., Zhang, B., Xu, K., Cai, W., & Chen, H. (2023). A survey of large language models. arXiv preprint arXiv:2303.18223. http://dx.doi.org/10.48550/arXiv.2303.18223.

- **Legal Cases**

53. Authors Guild v. Google, Inc., No. 13-4829 (2d Cir. 2015); https://law.justia.com/cases/federal/appellate-courts/ca2/13-4829/13-4829-2015-10-16.html.
54. Authors Guild, Inc. v. HathiTrust, No. 12-4547 (2d Cir. 2014); https://law.justia.com/cases/federal/appellate-courts/ca2/12-4547/12-4547-2014-06-10.html.
55. Bartz v. Anthropic PBC, No. C 24-05417 WHA, (N.D. Cal., June 23, 2025); https://storage.courtlistener.com/recap/gov.uscourts.cand.434709/gov.uscourts.cand.434709.231.0.pdf.
56. Silverman v. OpenAI, Inc., 3:23-cv-03416, (N.D. Cal.); https://dockets.justia.com/docket/california/candce/3:2023cv03416/415174; https://www.courtlistener.com/docket/67538258/tremblay-v-openai-inc/.
57. Sony Corp. of America v. Universal City Studios, Inc., 464 U.S. 417 (1984); https://supreme.justia.com/cases/federal/us/464/417/.
58. THALER v. PERLMUTTER et al, Civil Action 22-1564 (BAH); https://s3.documentcloud.org/documents/23919666/thalervperlmutter.pdf.
59. The New York Times Company v. Microsoft Corporation et al., No. 1:23-cv-11195 (S.D.N.Y. Apr. 4, 2025). Retrieved from https://law.justia.com/cases/federal/district-courts/new-york/nysdce/1:2023cv11195/612697/514/.
60. Tremblay v. OpenAI, Inc., 3:23-cv-03223, (N.D. Cal.); https://dockets.justia.com/docket/california/candce/3:2023cv03223/414822; https://docs.justia.com/cases/federal/district-courts/california/candce/3%3A2023cv03223/414822/358?utm_source=chatgpt.com.

- **Legal and Technical Websites**

61. Gertner, Jon, Wikipedia's Moment of Truth Can the online encyclopedia help teach A.I. chatbots to get their facts right — without destroying itself in the process?, July 18, 2023; https://www.nytimes.com/2023/07/18/magazine/wikipedia-ai-chatgpt.html. (Last visited: 8/17/2025)
62. Guadamuz, Andres, Artificial intelligence and copyright, oct 2017; https://www.wipo.int/wipo_magazine/en/2017/05/article_0003.html. (Last visited: 8/17/2025)
63. Hu, Krystal, ChatGPT sets record for fastest-growing user base, Feb 2, 2023; https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/. (Last visited: 8/17/2025)

64. Keary, Tim, 12 Practical Large Language Model (LLM) Applications, October 2, 2023; https://www.techopedia.com/12-practical-large-language-model-llm-applications. (Last visited: 8/17/2025)
65. Lock, Samantha, What is AI chatbot phenomenon ChatGPT and could it replace humans?, Dec 5, 2022; https://www.theguardian.com/technology/2022/dec/05/what-is-ai-chatbot-phenomenon-chatgpt-and-could-it-replace-humans. (Last visited: 8/17/2025)
66. Nyambura, Kiarie, Understanding Training Data in Contracts with AI Vendors, November 2023; https://contractnerds.com/understanding-training-data-in-contracts-with-ai-vendors/. (Last visited: 8/17/2025)
67. Ortiz, Sabrina, What is ChatGPT and why does it matter? Here's what you need to know, Sep 15, 2023; https://www.zdnet.com/article/what-is-chatgpt-and-why-does-it-matter-heres-everything-you-need-to-know/. (Last visited: 8/17/2025)
68. Self-Supervised Learning, Aug 2023; https://www.dremio.com/wiki/self-supervised-learning/#:~:text=In%20self%2Dsupervised%20learning%2C%20the,to%20create%20its%20own%20labels. (Last visited: 8/17/2025)
69. Uszkoreit, Jakob, Transformer: A Novel Neural Network Architecture for Language Understanding, Aug 31, 2017; https://blog.research.google/2017/08/transformer-novel-neural-network.html. (Last visited: 8/17/2025)
70. Wolfe, Cameron R., Ph.d., Data is the Foundation of Language Models, Jul 17, 2023; https://cameronrwolfe.substack.com/p/data-is-the-foundation-of-language. (Last visited: 8/17/2025)