Submit Date: 27 December 2025
Revise Date: 02 May 2026
Accept Date: 09 May 2026
Publish Date: 18 May 2026

The Encyclopedia of Comparative Jurisprudence and Law

Artificial Intelligence as a Tool or an Agent of Crime? A Comparative Examination in Contemporary Criminal Law

Hamideh Elahi[1], Asghar Abbasi[2]*, Seyyed Abdollah MirGhiasi[3]

1. Department of Law, Sar.C., Islamic Azad University, Sari, Iran
2. Department of Law, Cha.C., Islamic Azad University, Chalus, Iran
3. Department of Political Science, CT.C., Islamic Azad University, Tehran, Iran

* Corresponding Author's Email: abbasi@iauc.ac.ir

ABSTRACT

Artificial intelligence is increasingly participating in activities that, if carried out by a natural person or even a legal person such as a corporation, would constitute criminal conduct; these activities range from triggering sudden crashes in financial markets to purchasing illegal narcotics and even running over pedestrians. We argue that criminal law faces serious deficiencies in cases where an artificial intelligence system effectively commits an offense while no identifiable human actor situated upstream can, as a practical or legal matter, be held responsible for that conduct. This article examines possible approaches to addressing this problem, with particular emphasis on the possibility of imposing direct criminal liability on artificial intelligence where it operates independently, autonomously, and in a manner that cannot be reduced to the conduct of human actors. The prevailing view is that punishing artificial intelligence is incompatible with fundamental principles of criminal law, including the attribution of culpability and the requirement of a criminal mental element (mens rea). Drawing on analogies with corporate criminal liability, strict criminal liability, and established principles of attribution, the authors demonstrate that the possibility of punishing artificial intelligence cannot be categorically rejected on the basis of cursory theoretical arguments alone. Punishing artificial intelligence may produce general deterrent effects and serve symbolic and expressive functions, and it is not necessarily inconsistent with the limiting principles of criminal law, such as the prohibition against imposing punishment disproportionate to culpability. Nevertheless, the article ultimately concludes that the direct imposition of criminal punishment on artificial intelligence is not justifiable, because such an approach may entail substantial costs and would undoubtedly require fundamental and far-reaching changes to the existing legal framework. Accordingly, the authors contend that limited and proportionate reforms to existing criminal laws governing human persons, combined with the possible expansion of civil liability, constitute a more appropriate and practicable means of addressing crimes arising from artificial intelligence.

Keywords: *Artificial Intelligence, Criminal Law, Moral Responsibility*

How to cite: Elahi, H., Abbasi, A., & MirGhiasi, S. A. (2026). Artificial Intelligence as a Tool or an Agent of Crime? A Comparative Examination in Contemporary Criminal Law. *The Encyclopedia of Comparative Jurisprudence and Law*, *4*(1), 1-21.

تاریخ ارسال: ۶ دی ۱۴۰۴
تاریخ بازنگری: ۱۲ اردیبهشت ۱۴۰۵
تاریخ پذیرش: ۱۹ اردیبهشت ۱۴۰۵
تاریخ چاپ: ۲۸ اردیبهشت ۱۴۰۵

دانشنامه فقه و حقوق تطبیقی

هوش مصنوعی به مثابه ابزار یا عامل جرم؟ بررسی تطبیقی در حقوق کیفری معاصر

حمیده الهی[۱]، اصغر عباسی[۲*]، سیدعبدالله میرغیاثی[۳]

۱. گروه حقوق، واحد ساری، دانشگاه آزاد اسلامی، ساری، ایران
۲. گروه حقوق، واحد چالوس، دانشگاه آزاد اسلامی، چالوس، ایران
۳. گروه علوم سیاسی، واحد تهران مرکزی، دانشگاه آزاد اسلامی، تهران، ایران
* پست الکترونیک نویسنده مسئول: abbasi@iauc.ac.ir

چکیده

هوش مصنوعی به طور فزاینده‌ای در فعالیت‌هایی مشارکت می‌کند که اگر توسط یک شخص حقیقی یا حتی یک شخص حقوقی مانند شرکت‌ها انجام می‌شد، ماهیت مجرمانه داشت؛ از ایجاد سقوط‌های ناگهانی در بازارهای مالی گرفته تا خرید مواد مخدر غیرقانونی یا حتی زیر گرفتن عابران پیاده. ما بر این باوریم که حقوق کیفری در مواردی که یک سامانه هوش مصنوعی عملاً مرتکب جرم می‌شود و در عین حال هیچ عامل انسانیِ بالادستی که از نظر عملی یا حقوقی قابل شناسایی و انتساب باشد وجود ندارد، با کاستی‌های جدی مواجه است. این مقاله به بررسی راهکارهای احتمالی برای حل این مسئله می‌پردازد و تمرکز اصلی آن بر امکان مسئولیت کیفری مستقیم هوش مصنوعی در مواردی است که به صورت مستقل، خودمختار و غیرقابل تقلیل به کنشگران انسانی عمل می‌کند. دیدگاه رایج بر این است که مجازات هوش مصنوعی با اصول بنیادین حقوق کیفری، نظیر قابلیت انتساب تقصیر و ضرورت وجود عنصر روانی مجرمانه، ناسازگار است. نویسندگان با بهره‌گیری از قیاس‌هایی با مسئولیت کیفری شرکت‌ها، مسئولیت کیفری مطلق و نیز اصول شناخته‌شده انتساب، نشان می‌دهند که نمی‌توان صرفاً با استدلال‌های نظری و شتاب‌زده، امکان مجازات هوش مصنوعی را به طور مطلق رد کرد. مجازات هوش مصنوعی می‌تواند آثار بازدارندگی عمومی و کارکردهای نمادین و بیانی به همراه داشته باشد و لزوماً با محدودیت‌های منفی حقوق کیفری، مانند مجازات فراتر از میزان تقصیر، در تعارض نباشد. با این حال، نتیجه نهایی مقاله آن است که اعمال مجازات کیفری مستقیم بر هوش مصنوعی توجیه‌پذیر نیست؛ زیرا چنین رویکردی ممکن است هزینه‌های قابل توجهی به همراه داشته باشد و بی‌تردید مستلزم ایجاد تغییرات بنیادین و گسترده در ساختار حقوقی موجود خواهد بود. از این رو، نویسندگان معتقدند که اصلاحات محدود و متناسب در قوانین کیفری موجود که متوجه اشخاص انسانی هستند، همراه با توسعه احتمالی مسئولیت مدنی، راهکار مناسب‌تر و عملی‌تری برای مواجهه با جرایم ناشی از هوش مصنوعی به شمار می‌رود.

کلیدواژگان: *هوش مصنوعی، حقوق کیفری، مسئولیت اخلاقی*

نحوه استناددهی: الهی، حمیده.، عباسی، اصغر.، و میرغیاثی، سیدعبدالله. (۱۴۰۵). هوش مصنوعی به مثابه ابزار یا عامل جرم؟ بررسی تطبیقی در حقوق کیفری معاصر. *دانشنامه فقه و حقوق تطبیقی*، ۴(۱)، ۲۱-۱.

مقدمه

با گسترش روزافزون کاربرد هوش مصنوعی در عرصه‌های گوناگون زندگی بشری ـ از خودروهای خودران و ربات‌های جراح گرفته تا سیستم‌های تصمیم‌ساز در امور مالی، قضایی و نظامی ـ پرسشی بنیادین در نظام‌های حقوقی و فقهی معاصر مطرح شده است: آیا می‌توان هوش مصنوعی را در قبال رفتارهای زیان‌بار خود «مجازات» کرد؟ این پرسش از آن جهت اهمیت می‌یابد که هوش مصنوعی، برخلاف انسان و حتی حیوان، نه از «اراده» و «قصد» به معنای متعارف برخوردار است، نه «تکلیف‌پذیری» و «مسئولیت اخلاقی» در چارچوب تعاریف سنتی را دارا می‌باشد (Amirian Farsani & Hosseini, 2023). با این حال، فعل زیان‌باری که از سوی یک سامانه خودمختار سر می‌زند، از نظر عرف و قانون «بی‌فاعل» نمی‌ماند؛ خسارت مادی و معنوی واقعی رخ می‌دهد و ضرورت جبران و پاسخ‌دهی احساس می‌شود (Makki et al., 2024).

نظام حقوقی ایران که بر پایه فقه امامیه بنا شده است، مسئولیت و مجازات را بر محور «انسان مکلف» تعریف می‌کند. مفاهیمی چون «عمد»، «خطا»، «قصد»، «تقصیر» و «اهلیت» همگی بر وجود شخصی حقیقی با شعور و اختیار مبتنی است. حال آنکه هوش مصنوعی، هرچند می‌تواند رفتاری شبیه به رفتار عاقلانه از خود نشان دهد، فاقد «نفس مدرِک»، «تکلیف شرعی» و «تشخیص حسن و قبح ذاتی» است (Ansari & Atazadeh, 2019). از این رو، برخی صاحب‌نظران مجازات مستقیم هوش مصنوعی را ناممکن یا نامعقول می‌دانند و معتقدند هرگونه پاسخ کیفری باید متوجه طراح، تولیدکننده، مالک یا کاربر سیستم باشد (Kaveh & Barani, 2024).

در مقابل، دیدگاه‌های نوینی در حال شکل‌گیری است که با استناد به نظریه‌های «مسئولیت جمعی»، «تسبیب گسترده»، «مسئولیت خطر» و حتی «شخصیت حقوقی الکترونیک»، امکان نوعی کیفر را متوجه خود نهاد هوش مصنوعی می‌دانند. این دیدگاه‌ها گاه تا بدانجا پیش می‌روند که مفاهیمی چون «حبس رایانه‌ای» (قطع دسترسی سیستم از شبکه)، «جریمه نرم‌افزاری»، «اعمال محدودیت در سطح دسترسی» و «حذف کدهای عامل زیان» را به‌عنوان مصادیق نوین «مجازات» پیشنهاد می‌کنند.

مقاله حاضر با روش تحلیلی ـ تطبیقی در پی پاسخ به این پرسش است که آیا مبانی فقهی و حقوقی ایران ظرفیت پذیرش نوعی «مجازات هوش مصنوعی» را دارند؟ و در صورت پاسخ مثبت، چه مصادیق، شرایط و محدودیت‌هایی برای چنین مجازاتی قابل تصور است؟ در این مسیر، ضمن تبیین ماهیت هوش مصنوعی از منظر «اراده»، «شعور» و «قصد»، به تحلیل قواعد سنتی مسئولیت و مجازات (مانند اتلاف، تسبیب و کیفرهای شرعی) پرداخته و امکان بسط یا بازتعریف آن‌ها را در مواجهه با فناوری‌های خودمختار مورد ارزیابی قرار خواهیم داد.

نظریه متعارف مجازات

برای ایجاد مبنایی روشن جهت تحلیل‌های خود، در این بخش نظریه‌ای از مجازات ارائه می‌شود که بازتاب‌دهنده اجماع گسترده موجود در ادبیات حقوق کیفری است. نویسندگان اصطلاح «مجازات» را بر اساس تعریف مشهور هربرت ال. ای. هارت و با توجه به پنج عنصر اساسی آن به کار می‌برند:

مجازات باید متضمن درد، رنج یا پیامدهایی باشد که عموماً ناخوشایند تلقی می‌شوند.

مجازات باید در واکنش به نقض قواعد حقوقی اعمال شود.

مجازات باید متوجه مرتکب واقعی یا فرضی جرم به دلیل ارتکاب همان جرم باشد.

مجازات باید به‌طور عمدی توسط اشخاصی غیر از مرتکب اعمال شود.

مجازات باید از سوی مقامی که بر اساس یک نظام حقوقی صلاحیت یافته است، علیه شخصی که مرتکب نقض آن نظام حقوقی شده، وضع و اجرا گردد (Sarch).

بر این اساس، «مجازات» مستلزم محکومیت کیفری به دلیل ارتکاب یک جرم شناخته‌شده قانونی و پس از طی تشریفات و فرآیندهای پذیرفته‌شده دادرسی است. بنابراین، اقداماتی نظیر حبس، جزای نقدی یا مصادره اموال که در پی محکومیت قانونی و قطعی به یک جرم مشخص اعمال می‌شوند، مصداق مجازات محسوب می‌شوند؛ اما بسیاری از اقداماتی که در معنای عرفی ممکن است مجازات تلقی شوند، از منظر حقوق کیفری در زمره مجازات قرار نمی‌گیرند.

برای مثال، رفتارهای تنبیهی و سخت‌گیرانه شهروندان عادی نسبت به افرادی که صرفاً هنجارهای اجتماعی غیررسمی را نقض کرده‌اند، نگهداری پیشگیرانه افراد مبتلا به اختلالات روانی به دلیل خطرناک بودن برای خود یا دیگران، یا ضبط اموال پیش از صدور حکم محکومیت، از مصادیق مجازات کیفری محسوب نمی‌شوند. همچنین، ضمانت اجراهای مدنی، هرچند در پاسخ به نقض قواعد حقوقی اعمال می‌شوند، از منظر حقوق کیفری «جرم» به شمار نمی‌آیند؛ زیرا هدف حقوق کیفری محکوم‌سازی و نکوهش شدید رفتارهایی است که جامعه آن‌ها را در زمره جدی‌ترین تخلفات قرار می‌دهد.

از دیدگاه نویسندگان، اعمال مجازات تنها زمانی موجه است که دلایل مثبت توجیه‌کننده آن بر هزینه‌ها و پیامدهای منفی آن غلبه داشته باشد و در عین حال با محدودیت‌های بنیادین حقوق کیفری تعارض نداشته باشد. منظور از «توجیهات مثبت» منافعی است که مجازات می‌تواند برای جامعه ایجاد کند؛ منافعی همچون کاهش آسیب‌ها، افزایش امنیت عمومی، ارتقای رفاه اجتماعی یا ابراز پایبندی جامعه به ارزش‌های اساسی اخلاقی و سیاسی. چنین منافعی می‌توانند توجیهی برای جرم‌انگاری برخی رفتارها و اعمال ضمانت اجرا علیه مرتکبان آن‌ها فراهم آورند.

با این حال، توجیهات مثبت با «محدودیت‌های منفی» مجازات تفاوت دارند. این محدودیت‌ها معمولاً با رویکردهای سزاگرایانه مبتنی بر تقصیر و استحقاق کیفری ارتباط دارند. برای نمونه، در نظریه‌های غالب حقوق کیفری، مجازات افراد بی‌گناه یا تحمیل مجازاتی بیش از میزان استحقاق ناشی از تقصیر مرتکب، ناعادلانه تلقی می‌شود؛ حتی اگر چنین اقدامی بتواند رفاه عمومی یا منافع اجتماعی بیشتری ایجاد کند. این اصل که به «محدودیت مبتنی بر استحقاق» معروف است، بیانگر آن است که تحقق اهداف اجتماعی و رفاهی از طریق مجازات، همواره باید در چارچوب الزامات عدالت کیفری و تناسب میان تقصیر و مجازات صورت گیرد و نمی‌توان به بهانه منافع اجتماعی، از این محدودیت‌های بنیادین عبور کرد (Müller & Bostrom, 2016).

دلایل ایجابی برای مجازات

در نظریه‌های معاصر حقوق کیفری، معمولاً این دیدگاه پذیرفته شده است که مجازات دارای کارکردها و منافع متعددی است و نمی‌توان آن را صرفاً بر یک مبنای نظری خاص توجیه کرد. به همین دلیل، بسیاری از حقوقدانان و فلاسفه حقوق کیفری از رویکردی کثرت‌گرا در توجیه مجازات حمایت می‌کنند. در حقوق فدرال ایالات متحده نیز مهم‌ترین اهداف مجازات مورد شناسایی قرار گرفته‌اند؛ از جمله ایجاد بازدارندگی نسبت به رفتارهای مجرمانه، حمایت از جامعه در برابر جرایم آتی مرتکب، فراهم ساختن زمینه اصلاح و بازپروری بزهکار و همچنین انعکاس شدت جرم و میزان استحقاق کیفری مرتکب (Abbott, 2017).

برای سهولت تحلیل، اهداف ایجابی مجازات را می‌توان در دو دسته اصلی طبقه‌بندی کرد: نخست، اهداف فایده‌گرایانه یا پیامدگرایانه و دوم، اهداف سزاگرایانه یا استحقاق‌محور. برخی نظریه‌پردازان همچنین از اهداف بیانی یا نمادین سخن می‌گویند،

هرچند نویسندگان مقاله معتقدند این دسته از اهداف عمدتاً قابل تقلیل به یکی از دو دسته نخست هستند.

اهداف فایده‌گرایانه یا پیامدگرایانه

مهم‌ترین توجیه فایده‌گرایانه مجازات، کاهش جرم و پیشگیری از وقوع آسیب‌های اجتماعی است. بر اساس این دیدگاه، مجازات زمانی موجه است که به بهبود رفاه عمومی و کاهش زیان‌های ناشی از جرم کمک کند.

مجازات از چند طریق می‌تواند موجب کاهش جرم شود:

نخست، ناتوان‌سازی یا خنثی‌سازی: در این حالت، مجرم از طریق زندان یا سایر محدودیت‌های قانونی از ارتکاب جرایم بیشتر بازداشته می‌شود. برای مثال، فردی که در زندان به سر می‌برد، از نظر فیزیکی امکان ارتکاب بسیاری از جرایم را ندارد.

دوم، بازدارندگی: بازدارندگی مهم‌ترین و تأثیرگذارترین کارکرد پیشگیرانه مجازات به شمار می‌رود. در این رویکرد، تهدید به تحمیل پیامدهای منفی ناشی از مجازات، افراد را از ارتکاب جرم منصرف می‌کند. بازدارندگی خود به دو نوع تقسیم می‌شود:

- بازدارندگی خاص: زمانی رخ می‌دهد که مجازات یک فرد موجب شود همان شخص در آینده از ارتکاب جرم خودداری کند.
- بازدارندگی عام: زمانی تحقق می‌یابد که مجازات یک مجرم، دیگر افراد جامعه را نیز از ارتکاب رفتار مشابه منصرف سازد. در واقع، مجازات در اینجا نوعی «پیام اجتماعی» به سایر افراد ارسال می‌کند که ارتکاب چنین رفتاری با واکنش کیفری مواجه خواهد شد.

برای مثال، حتی مجازات افراد دارای اختلالات روانی شدید ممکن است از منظر بازدارندگی عام توجیه شود، زیرا می‌تواند افراد سالم را از ارتکاب جرم و سپس سوءاستفاده از دفاع جنون برای فرار از مسئولیت کیفری بازدارد.

سوم، اصلاح و بازپروری: مجازات ممکن است از طریق اصلاح نگرش‌های بزهکار، آموزش مهارت‌های اجتماعی یا حرفه‌ای و کمک به بازگشت سالم او به جامعه، احتمال ارتکاب جرم در آینده را کاهش دهد. هرچه مجازات در تغییر رفتار و نگرش مجرم موفق‌تر باشد، نقش مؤثرتری در پیشگیری از جرایم آتی خواهد داشت (Gurney, 2015).

اهداف سزاگرایانه یا استحقاق‌محور

در کنار اهداف پیشگیرانه، برخی نظریه‌پردازان معتقدند مجازات دارای ارزش ذاتی نیز هست و صرف‌نظر از آثار و نتایج آن، اجرای عدالت اقتضا می‌کند که افراد مقصر به سزای اعمال خود برسند.

بر اساس دیدگاه سزاگرایانه، مجازات صرفاً ابزاری برای کاهش جرم نیست، بلکه وسیله‌ای برای اعطای آن چیزی است که مجرم به دلیل رفتار نادرست خود مستحق آن شده است. به بیان دیگر، مجازات پاسخی متناسب به تقصیر اخلاقی و حقوقی مرتکب است.

این رویکرد بر این باور است که اگر جامعه در برابر رفتارهای مجرمانه واکنش مناسب نشان ندهد، نوعی مدارا یا حتی همدستی ضمنی با رفتار نادرست رخ خواهد داد. بنابراین، مجازات، علاوه بر آثار عملی خود، نقش مهمی در حفظ مرزهای اخلاقی و هنجاری جامعه ایفا می‌کند.

اگرچه تقریباً همه نظریه‌پردازان قبول دارند که پیشگیری از جرم بخش مهمی از توجیه مجازات را تشکیل می‌دهد، درباره وجود یا میزان اهمیت دلایل سزاگرایانه همچنان اختلاف نظر وجود دارد. نویسندگان مقاله نیز با وجود اهمیت این دیدگاه، فرض را بر آن می‌گذارند که بخش عمده توجیه مجازات بر مبنای کاهش آسیب‌ها و تحقق پیامدهای مطلوب اجتماعی استوار است.

اهداف بیانی یا نمادین

دسته دیگری از دلایل ایجابی برای مجازات، دلایل بیانی یا نمادین هستند. مجازات صرفاً اعمال رنج یا محدودیت بر مجرم نیست، بلکه نوعی ابزار ارتباطی نیز محسوب می‌شود که از طریق آن

جامعه پایبندی خود را به ارزش‌های بنیادین اخلاقی و حقوقی اعلام می‌کند.

هنگامی که دولت فردی را به دلیل ارتکاب جرم محکوم می‌کند، در واقع به‌صورت رسمی رفتار او را محکوم و تقبیح می‌نماید. این محکومیت رسمی می‌تواند آثار روانی مثبتی برای بزه‌دیدگان داشته باشد؛ زیرا نشان می‌دهد که جامعه و نظام حقوقی از حقوق نقض‌شده آنان حمایت می‌کنند.

علاوه بر این، اعلام رسمی ناپسند بودن رفتارهای مجرمانه می‌تواند نگرش‌ها و رفتارهای اجتماعی را نیز تحت تأثیر قرار دهد و موجب تقویت ارزش‌های مثبت و هنجارهای مورد پذیرش جامعه شود.

با این حال، برخی نویسندگان تردید دارند که اهداف بیانی واقعاً دسته‌ای مستقل از دلایل توجیه‌کننده مجازات باشند. از یک سو، آثار نمادین مجازات اغلب به کاهش جرم و پیشگیری از آسیب‌ها منجر می‌شود و بنابراین می‌توان آن را در زمره اهداف فایده‌گرایانه قرار داد. از سوی دیگر، محکومیت رسمی مجرم را می‌توان نوعی اعطای استحقاق کیفری به وی دانست که به نظریه سزاگرایی نزدیک است (Alschuler, 2009).

به همین دلیل، نویسندگان مقاله فرض می‌کنند که کارکردهای بیانی مجازات در نهایت به یکی از دو دسته فایده‌گرایانه یا سزاگرایانه قابل تقلیل هستند؛ هرچند استدلال‌های آنان با دیدگاه مخالف نیز سازگار است.

محدودیت‌های سلبی (سزاگرایانه) بر مجازات

علاوه بر وجود دلایل ایجابی برای اعمال مجازات، نظام حقوق کیفری باید به برخی تعهدات بنیادین اخلاقی و هنجاری، مانند عدالت و انصاف، نیز پایبند باشد. به بیان دیگر، حتی اگر مجازات بتواند منافع اجتماعی قابل توجهی ایجاد کند، باز هم در صورتی که با اصول اساسی عدالت در تعارض باشد، اعمال آن موجه نخواهد بود. مهم‌ترین این محدودیت‌ها بر میزان تقصیر و مسئولیت افرادی متمرکز است که در معرض پاسخ کیفری قرار می‌گیرند.

یکی از مهم‌ترین این محدودیت‌ها، «قید استحقاق» یا محدودیت مبتنی بر سزاواری است که در اکثر نظریه‌های سزاگرایانه جایگاه محوری دارد. بر اساس این اصل، هیچ فردی را نمی‌توان عادلانه به مجازاتی بیش از آنچه مستحق آن است محکوم کرد. مفهوم استحقاق نیز عمدتاً بر پایه میزان تقصیر و مسئولیت اخلاقی و حقوقی ناشی از رفتار فرد سنجیده می‌شود. بنابراین، اثر اصلی این اصل آن است که مجازات باید متناسب با میزان تقصیر مرتکب باشد و از حدود آن فراتر نرود.

برای مثال، حتی اگر اعدام فردی به دلیل عبور غیرمجاز از خیابان بتواند در بلندمدت موجب کاهش تخلفات مشابه و حفظ جان افراد بیشتری شود، چنین مجازاتی به دلیل عدم تناسب آشکار با میزان تقصیر مرتکب، ناعادلانه و غیرقابل توجیه خواهد بود. در واقع، فایده اجتماعی نمی‌تواند به‌تنهایی مجوز تحمیل مجازات‌های نامتناسب باشد.

مبنای این محدودیت چیست؟ نخست، شهود و درک عمومی از عدالت. به‌طور شهودی، بیشتر افراد بر این باورند که مجازات شخص کاملاً بی‌گناه، حتی اگر منافع گسترده‌ای برای جامعه ایجاد کند، ناعادلانه است. به همین ترتیب، تحمیل مجازات بسیار شدید به فردی که مرتکب جرمی خفیف شده نیز خلاف عدالت تلقی می‌شود.

افزون بر این شهود اخلاقی، محدودیت مبتنی بر استحقاق از یک استدلال فلسفی مهم نیز حمایت می‌شود که دست‌کم به اندیشه‌های ایمانوئل کانت بازمی‌گردد. بر اساس اصل مشهور کانتی، انسان‌ها نباید صرفاً به‌عنوان ابزاری برای تحقق اهداف دیگران مورد استفاده قرار گیرند، بلکه باید همواره به‌عنوان غایت فی‌نفسه و موجوداتی دارای ارزش ذاتی مورد احترام باشند. مجازات یک فرد بی‌گناه به منظور دستیابی به منافع اجتماعی گسترده‌تر، نمونه بارز استفاده

ابزاری از انسان است؛ زیرا در چنین وضعیتی، شأن و منزلت شخص نادیده گرفته می‌شود و او صرفاً وسیله‌ای برای تحقق اهداف جمعی تلقی می‌گردد.

در برخی قرائت‌های سخت‌گیرانه از فلسفه کانتی، محدودیت مبتنی بر استحقاق ماهیتی مطلق دارد؛ یعنی تحت هیچ شرایطی نمی‌توان برای دستیابی به منافع اجتماعی، حقوق بنیادین اشخاص را نقض کرد. در مقابل، برخی دیدگاه‌های معتدل‌تر معتقدند که در شرایط استثنایی، اگر منافع حاصل از نقض یک محدودیت بسیار چشمگیر باشد، ممکن است بتوان چنین نقضی را توجیه کرد. با این حال، حتی این دیدگاه‌ها نیز اصل تناسب و استحقاق را یکی از مهم‌ترین محدودیت‌های نظام کیفری می‌دانند.

با وجود اهمیت قید استحقاق، محدودیت‌های حاکم بر مجازات تنها به این اصل منحصر نمی‌شوند. یکی دیگر از مهم‌ترین شرایط اعمال مسئولیت کیفری، برخورداری از ظرفیت تقصیرپذیری است. حقوق کیفری اساساً برای نکوهش و سرزنش رفتارهای قابل انتساب به عاملان مسئول طراحی شده است و نقطه شروع اغلب نظام‌های کیفری آن است که مجازات تنها در واکنش به رفتارهای مقصرانه و قابل سرزنش اعمال می‌شود.

از این رو، برای آنکه یک شخص یا نهاد بتواند موضوع مناسب مجازات کیفری قرار گیرد، باید از حداقلی از توانایی‌های ادراکی، عقلانی و ارادی برخوردار باشد؛ توانایی‌هایی که امکان تصمیم‌گیری، سنجش پیامدهای رفتار و کنترل اعمال را فراهم می‌کنند. اگر موجودی فاقد چنین ظرفیت‌هایی باشد، اصولاً نمی‌توان او را مخاطب مناسب سرزنش کیفری دانست.

این موضوع در نهادهای دفاعی حقوق کیفری نیز منعکس شده است. برای مثال، در مواردی که شخص به دلیل اختلال شدید روانی یا فقدان کامل توانایی تشخیص و اراده، فاقد ظرفیت لازم برای مسئولیت‌پذیری کیفری باشد، ممکن است از دفاع مبتنی بر عدم اهلیت کیفری یا جنون بهره‌مند شود. بنابراین، برخورداری از ظرفیت تقصیرپذیری نه صرفاً یک عامل مؤثر در تعیین میزان مجازات، بلکه پیش‌شرط اساسی و شرط لازم برای قرار گرفتن در قلمرو حقوق کیفری است. در نتیجه، از منظر نظریه‌های متعارف مجازات، مشروعیت پاسخ کیفری علاوه بر تحقق اهدافی همچون بازدارندگی، اصلاح یا تأمین عدالت، وابسته به رعایت محدودیت‌های بنیادینی است که مهم‌ترین آن‌ها تناسب میان مجازات و استحقاق مرتکب و نیز وجود ظرفیت لازم برای انتساب تقصیر و مسئولیت کیفری است. این محدودیت‌ها نقشی اساسی در حفظ عدالت کیفری ایفا می‌کنند و مانع آن می‌شوند که مجازات صرفاً به ابزاری برای دستیابی به منافع اجتماعی تبدیل گردد (Scheutz, 2012).

جایگزین‌های مجازات

برای آنکه اعمال مجازات موجه تلقی شود، صرف وجود منافع و کارکردهای مثبت و همچنین سازگاری آن با محدودیت‌های بنیادین حقوق کیفری کافی نیست. افزون بر این دو شرط، باید احراز شود که هیچ گزینه بهتر و عملی‌تری برای دستیابی به اهداف مورد نظر وجود ندارد؛ حتی گزینه‌ای مانند عدم مداخله کیفری نیز باید مورد ارزیابی قرار گیرد. این اصل، یکی از مبانی پذیرفته‌شده در تحلیل‌های سیاست‌گذاری عمومی است و در حوزه حقوق کیفری نیز اهمیت ویژه‌ای دارد.

بر این اساس، حتی اگر مجازات مستقیم هوش مصنوعی دارای آثار و منافع مثبت باشد و حتی اگر چنین اقدامی با اصول عدالت کیفری و محدودیت‌های مربوط به تقصیر و مسئولیت مغایرت نداشته باشد، باز هم نمی‌توان فوراً آن را موجه دانست. برای مثال، اگر سازوکارهایی مانند مسئولیت مدنی، نظام‌های صدور مجوز، نظارت‌های اداری، استانداردهای صنعتی یا سایر ابزارهای تنظیم‌گری بتوانند همان اهداف را با هزینه کمتر و کارایی بیشتر محقق کنند، توسل به مجازات کیفری توجیه خود را از دست خواهد داد.

در نظریه‌های حقوق کیفری غالباً تأکید می‌شود که حقوق کیفری باید «آخرین ابزار» برای کنترل اجتماعی باشد. علت این امر آن است که ضمانت اجراهای کیفری شدیدترین واکنش‌هایی هستند که جامعه در اختیار دارد. مجازات کیفری نه‌تنها ممکن است به سلب آزادی‌های اساسی افراد بینجامد، بلکه متضمن محکومیت رسمی و سرزنش اخلاقی و اجتماعی مرتکب نیز هست. از این رو، پیش از توسل به چنین ابزار قدرتمندی، باید بررسی شود که آیا روش‌های کم‌هزینه‌تر و کم‌مداخله‌تری برای حل مسئله وجود دارد یا خیر.

بنابراین، سومین شرط اساسی برای مشروعیت هر نوع مجازات آن است که جایگزین‌های مناسب‌تر و مؤثرتری برای دستیابی به اهداف مورد نظر وجود نداشته باشد.

جمع‌بندی و تلفیق مباحث

بر اساس چارچوب نظری ارائه‌شده، ارزیابی مشروعیت و مناسب بودن هر نوع مجازات مستلزم پاسخ به سه پرسش بنیادین است:

منافع و دلایل ایجابی

آیا دلایل کافی و قانع‌کننده‌ای برای اعمال مجازات وجود دارد؟

این پرسش عمدتاً ناظر بر منافع فایده‌گرایانه مجازات، از جمله کاهش آسیب‌ها، پیشگیری از جرم، افزایش امنیت و ارتقای رفاه اجتماعی است؛ هرچند ممکن است ملاحظات سزاگرایانه و کارکردهای نمادین و بیانی مجازات نیز در این ارزیابی دخالت داشته باشند.

محدودیت‌های سلبی مبتنی بر تقصیر

آیا اعمال مجازات با محدودیت‌های بنیادین حقوق کیفری سازگار است؟

این بخش عمدتاً بر اصولی مانند تناسب میان تقصیر و مجازات، اصل استحقاق کیفری و همچنین شرایط اولیه لازم برای اعمال مسئولیت کیفری ـ از جمله برخورداری از ظرفیت تقصیرپذیری ـ تمرکز دارد. حتی اگر مجازات دارای آثار مثبت فراوان باشد، در صورتی که با این محدودیت‌ها تعارض داشته باشد، مشروعیت خود را از دست می‌دهد.

وجود یا عدم وجود جایگزین‌های عملی

آیا مجازات بهترین راهکار برای مقابله با آسیب‌ها و ناهنجاری‌های مورد نظر است؟

در این مرحله باید بررسی شود که آیا ابزارهایی مانند مسئولیت مدنی، مقررات اداری، استانداردهای حرفه‌ای، نظارت‌های فنی یا حتی عدم مداخله، می‌توانند نتایج مشابه یا بهتری را با هزینه‌های کمتر و پیامدهای منفی محدودتر فراهم آورند یا خیر.

در جمع‌بندی موضوع می‌توان عنوان کرد که نویسندگان بر مبنای این چارچوب سه‌گانه، در ادامه مقاله به ارزیابی امکان و مشروعیت مجازات مستقیم هوش مصنوعی می‌پردازند. بدین منظور، ابتدا در بخش سوم مقاله به بررسی منافع و دلایل ایجابی احتمالی برای مجازات هوش مصنوعی می‌پردازند؛ سپس در بخش چهارم، مسئله محدودیت‌های سلبی و چالش‌های مرتبط با تقصیر، استحقاق کیفری و ظرفیت مسئولیت‌پذیری هوش مصنوعی را تحلیل می‌کنند؛ و در نهایت، در بخش پنجم، این پرسش را مورد بررسی قرار می‌دهند که آیا راهکارهایی مانند مسئولیت مدنی، تنظیم‌گری و سایر شیوه‌های غیرکیفری می‌توانند جایگزین مناسب‌تر و کارآمدتری نسبت به مجازات مستقیم هوش مصنوعی باشند یا خیر.

به این ترتیب، مشروعیت مسئولیت کیفری مستقیم هوش مصنوعی نه صرفاً بر اساس یک معیار، بلکه از رهگذر ارزیابی هم‌زمان سه مؤلفه اساسیِ «منافع مجازات»، «رعایت محدودیت‌های عدالت کیفری» و «نبود جایگزین‌های برتر» سنجیده خواهد شد.

۲- استدلال‌های ایجابی در حمایت از مجازات هوش مصنوعی

در این بخش، نویسندگان به بررسی منافع و دلایل ایجابی‌ای می‌پردازند که ممکن است برای توجیه مجازات مستقیم هوش مصنوعی مطرح شوند. تمرکز اصلی بحث بر منافع پیامدگرایانه و

فایده‌گرایانه است. هرچند ممکن است ملاحظات سزاگرایانه نیز تا حدودی از مجازات حمایت کنند، نویسندگان فرض را بر این می‌گذارند که در این زمینه، ملاحظات مبتنی بر کاهش آسیب و پیشگیری از زیان، اهمیت بیشتری نسبت به ملاحظات استحقاق‌محور دارند. هدف این بخش ارائه فهرستی جامع از تمامی مزایای احتمالی مجازات هوش مصنوعی نیست، بلکه نشان دادن این نکته است که چنین مجازاتی دست‌کم می‌تواند برخی آثار مثبت و قابل توجه به همراه داشته باشد (Darling, 2016).

منافع پیامدگرایانه

همان‌گونه که پیش‌تر بیان شد، یکی از مهم‌ترین اهداف مجازات، کاهش جرایم و آسیب‌های اجتماعی از طریق بازدارندگی است. بر همین اساس، یکی از نخستین ایرادهایی که نسبت به مجازات هوش مصنوعی مطرح می‌شود این است که چنین مجازاتی اساساً فاقد کارکرد بازدارندگی خواهد بود؛ زیرا هوش مصنوعی قابلیت بازدارندگی‌پذیری ندارد.

پیتر آسارو (Hallevy, 2010) استدلال می‌کند که بازدارندگی تنها در مورد عاملان اخلاقی معنا دارد؛ یعنی موجوداتی که بتوانند شباهت میان اعمال خود و اعمال سایر اشخاص مجازات‌شده را درک کنند و بر اساس آن رفتار خود را تنظیم نمایند. به اعتقاد وی، بدون چنین درکی از شباهت میان عاملان، مجازات نمی‌تواند اثر بازدارنده داشته باشد. بنابراین، اگر سامانه‌های هوش مصنوعی ـ دست‌کم در وضعیت کنونی ـ قادر نباشند ممنوعیت‌ها و ضمانت اجراهای کیفری را شناسایی کرده و رفتار خود را بر آن اساس تنظیم کنند، مجازات آن‌ها فاقد توجیه خواهد بود.

نویسندگان تا حد زیادی می‌پذیرند که سامانه‌های هوش مصنوعی موجود، برخلاف انسان‌ها، به مجازات واکنش نشان نمی‌دهند و احتمالاً در آینده نزدیک نیز چنین قابلیتی نخواهند داشت؛ هرچند از لحاظ نظری طراحی سامانه‌های بازدارندگی‌پذیر امکان‌پذیر است.

با این حال، پاسخ به این ایراد مستلزم تفکیک میان دو نوع بازدارندگی است: بازدارندگی خاص و بازدارندگی عام.

بازدارندگی خاص زمانی رخ می‌دهد که مجازات یک شخص موجب شود همان شخص در آینده از ارتکاب جرم خودداری کند. اما بازدارندگی عام به تأثیر مجازات بر سایر افراد جامعه اشاره دارد؛ یعنی مجازات یک مرتکب، دیگران را از ارتکاب رفتار مشابه بازمی‌دارد.

نویسندگان افزون بر این، میان دو نوع بازدارندگی عام نیز تمایز قائل می‌شوند:

بازدارندگی عام مرتبط با همان جرم (Offense-Relative General Deterrence)؛ یعنی جلوگیری از ارتکاب همان نوع جرم توسط دیگران.

بازدارندگی عام نامحدود (Unrestricted General Deterrence)؛ یعنی تأثیر کلی مجازات بر کاهش رفتارهای مجرمانه در جامعه.

نکته اساسی این است که حتی اگر مجازات هوش مصنوعی نتواند موجب بازدارندگی خاص یا بازدارندگی نسبت به سایر سامانه‌های هوش مصنوعی شود، همچنان می‌تواند آثار بازدارندگی عام نسبت به انسان‌ها داشته باشد.

به بیان دیگر، مجازات مستقیم هوش مصنوعی می‌تواند توسعه‌دهندگان، مالکان و کاربران سامانه‌های هوش مصنوعی را تشویق کند تا از طراحی، بهره‌برداری یا استقرار سامانه‌هایی که احتمال ایجاد خسارت‌های شدید و غیرموجه دارند، خودداری کنند. برای مثال، اگر یکی از ضمانت اجراها نابودی کامل سامانه هوش مصنوعی باشد ـ چیزی که برخی پژوهشگران از آن با عنوان «مجازات مرگ رباتی» یاد کرده‌اند ـ توسعه‌دهندگان یا مالکان چنین سامانه‌هایی ممکن است منافع اقتصادی قابل توجهی را از دست بدهند. همین امر می‌تواند انگیزه لازم را برای رعایت استانداردهای

ایمنی بیشتر و طراحی مسئولانه‌تر سامانه‌های هوش مصنوعی فراهم کند (Alschuler, 2009).

این اثر بازدارنده حتی می‌تواند در آینده قوی‌تر نیز شود؛ به‌ویژه اگر برای برخی انواع هوش مصنوعی، نظام‌های سرمایه‌گذاری اجباری، بیمه‌های مسئولیت یا انتقال هزینه‌های ناشی از مجازات به مالکان و بهره‌برداران پیش‌بینی شود.

ملاحظات بیانی و نمادین

علاوه بر منافع بازدارنده، مجازات هوش مصنوعی ممکن است دارای کارکردهای نمادین و بیانی نیز باشد.

یکی از مهم‌ترین این کارکردها، تأمین رضایت روانی و احساسی بزه‌دیدگان است. هنگامی که یک سامانه هوش مصنوعی موجب خسارت جدی به اشخاص می‌شود، مجازات آن می‌تواند این احساس را در قربانیان ایجاد کند که نظام حقوقی نسبت به آسیب وارده بی‌تفاوت نبوده و آن را به رسمیت شناخته است.

کریستینا مولیگان (Darling, 2016) استدلال می‌کند که مجازات ربات‌ها می‌تواند نوعی رضایت روانی برای قربانیان فراهم آورد. از دیدگاه وی، واکنش تنبیهی نسبت به ربات‌های زیان‌رسان می‌تواند به قربانیان احساس احقاق حق و التیام روانی بدهد. در واقع، مجازات در اینجا حامل پیامی رسمی از سوی دولت است که رفتار زیان‌بار را محکوم می‌کند و ارزش و حقوق قربانیان را مورد تأیید قرار می‌دهد. این پیام رسمی می‌تواند نه‌تنها احساس امنیت قربانیان، بلکه اعتماد عمومی جامعه به نظام حقوقی را نیز افزایش دهد. اهمیت این استدلال زمانی بیشتر می‌شود که به یافته‌های روان‌شناسی اجتماعی توجه کنیم. مطالعات تجربی نشان داده‌اند که انسان‌ها تمایل زیادی به انسان‌انگاری (Anthropomorphism) دارند؛ یعنی به اشیا، سازمان‌ها یا سامانه‌های مصنوعی ویژگی‌های انسانی نسبت می‌دهند. همان‌گونه که بسیاری از افراد برای شرکت‌ها نوعی شخصیت و اراده قائل می‌شوند، احتمال می‌رود این گرایش در مورد ربات‌ها و سامانه‌های هوش مصنوعی انسان‌نما بسیار قوی‌تر باشد (Darling, 2016).

از این منظر، اگر جامعه هوش مصنوعی را موجودی قابل سرزنش تلقی کند، اما نظام حقوقی هیچ واکنش محکوم‌کننده‌ای نسبت به رفتار زیان‌بار آن نشان ندهد، ممکن است مشروعیت و اعتبار نظام عدالت کیفری در افکار عمومی تضعیف شود. بنابراین، مجازات هوش مصنوعی می‌تواند از طریق کارکرد نمادین خود، برخی منافع اجتماعی و روانی ایجاد کند.

نقد ملاحظات بیانی

با وجود این مزایا، نویسندگان نسبت به اتکای بیش از حد بر دلایل نمادین و احساسی هشدار می‌دهند.

نخست آنکه توجیه مجازات صرفاً بر مبنای ارضای احساسات انتقام‌جویانه قربانیان، شباهت زیادی به منطق «عدالت توده‌ای» یا «عدالت انتقام‌محور» دارد. صرف اینکه مجازات هوش مصنوعی در میان مردم محبوب باشد یا احساس رضایت ایجاد کند، به خودی خود نشان‌دهنده عادلانه بودن آن نیست.

به تعبیر فیلسوف حقوق، دیوید لوئیس (Sarch)، اگر مطالبه عمومی برای مجازات ناعادلانه باشد، اجابت آن از طریق نظام حقوقی نیز ناعادلانه خواهد بود؛ حتی اگر از منظر سیاسی یا اجتماعی سودمند به نظر برسد.

علاوه بر این، ترویج نگرش‌های انتقام‌جویانه نسبت به ربات‌ها ممکن است آثار منفی گسترده‌تری بر رفتار انسان‌ها داشته باشد. برای مثال، کیت دارلینگ (Darling, 2016) استدلال می‌کند که حتی ربات‌ها نیز باید تا حدی در برابر رفتارهای خشونت‌آمیز محافظت شوند؛ زیرا عادی‌سازی خشونت نسبت به موجودات شبه‌انسانی می‌تواند به‌تدریج هنجارهای اخلاقی حاکم بر روابط انسانی را تضعیف کند.

از سوی دیگر، مجازات هوش مصنوعی ممکن است پیام‌های ناخواسته و مسئله‌سازی را نیز به جامعه منتقل کند. اگر نظام کیفری هوش مصنوعی را مجازات کند، این امر ممکن است چنین تلقی

شود که هوش مصنوعی همانند انسان یک عامل مستقل، مسئول و دارای جایگاه حقوقی خاص است. پذیرش چنین برداشتی می‌تواند در آینده زمینه طرح ادعاهایی درباره حقوق، کرامت یا حمایت‌های ویژه برای هوش مصنوعی را فراهم سازد؛ موضوعی که ممکن است محدودیت‌هایی بر فعالیت‌های مشروع انسانی ایجاد کند.

در مجموع، نویسندگان نتیجه می‌گیرند که مجازات مستقیم هوش مصنوعی می‌تواند برخی منافع ایجابی به همراه داشته باشد. مهم‌ترین این منافع عبارت‌اند از:

- ایجاد بازدارندگی عام برای توسعه‌دهندگان، مالکان و کاربران سامانه‌های هوش مصنوعی؛
- تشویق به طراحی ایمن‌تر و مسئولانه‌تر فناوری‌های هوش مصنوعی؛
- تأمین برخی کارکردهای نمادین و روانی برای قربانیان؛
- تقویت احساس حمایت حقوقی و اعتماد عمومی به نظام عدالت.

با این حال، وجود این منافع به‌تنهایی برای توجیه مجازات کافی نیست. هنوز باید بررسی شود که آیا مجازات هوش مصنوعی با محدودیت‌های بنیادین حقوق کیفری ـ به‌ویژه اصول مربوط به تقصیر، مسئولیت و قابلیت سرزنش ـ سازگار است یا خیر. به همین دلیل، نویسندگان در بخش بعدی به بررسی این پرسش اساسی می‌پردازند که آیا مجازات هوش مصنوعی با محدودیت‌های سلبی مبتنی بر تقصیر و استحقاق کیفری تعارض دارد یا نه (Sarch).

جایگزین‌های عملی و امکان‌پذیر

نویسندگان تا این مرحله استدلال کرده‌اند که مجازات هوش مصنوعی می‌تواند دارای برخی منافع و کارکردهای مثبت باشد و همچنین لزوماً با محدودیت‌های بنیادین حقوق کیفری، مانند اصل تقصیر و شرایط مسئولیت کیفری، ناسازگار نیست. با این حال، این نتیجه به‌تنهایی برای اثبات مشروعیت مجازات هوش مصنوعی کافی نیست. مطابق چارچوب نظری ارائه‌شده، باید به پرسش سوم نیز پاسخ داد: آیا منافع حاصل از مجازات هوش مصنوعی بر هزینه‌های آن غلبه می‌کند و آیا این راهکار نسبت به سایر گزینه‌های موجود برتری دارد؟

این گزینه‌ها ممکن است شامل عدم مداخله کیفری، استفاده از مسئولیت مدنی، ابزارهای تنظیم‌گری و نظارتی، یا اصلاحات محدود در قوانین کیفری موجود باشد که همچنان بر اشخاص انسانی تمرکز دارند.

در این بخش، نویسندگان ابتدا به بررسی این پرسش می‌پردازند که آیا وضعیت فعلی حقوق کیفری برای مقابله با جرایم ناشی از هوش مصنوعی کافی است یا خیر. سپس هزینه‌های مجازات مستقیم هوش مصنوعی و راهکارهای جایگزین را ارزیابی می‌کنند.

نخستین گزینه: حفظ وضعیت موجود

نخستین پرسش این است که آیا اساساً نیازی به ایجاد مسئولیت کیفری مستقیم برای هوش مصنوعی وجود دارد یا خیر؟ شاید بتوان با تکیه بر قواعد موجود حقوق کیفری و بدون هیچ اصلاحی، با تمامی آسیب‌های ناشی از هوش مصنوعی مقابله کرد.

نویسندگان در پاسخ به این پرسش معتقدند که حقوق کیفری موجود در برخی موارد با خلأهایی مواجه است؛ با این حال، این خلأها بسیار محدودتر از آن چیزی هستند که طرفداران مسئولیت کیفری مستقیم هوش مصنوعی ادعا می‌کنند.

مواردی که خلأ کیفری هوش مصنوعی وجود ندارد: رفتارهای زیان‌بار قابل انتساب به انسان

نویسندگان در ابتدا مواردی را کنار می‌گذارند که در آن‌ها رفتار زیان‌بار هوش مصنوعی به‌طور کامل قابل انتساب به اشخاص انسانی است. در چنین مواردی نیازی به مجازات خود هوش مصنوعی وجود ندارد؛ زیرا مسئولیت کیفری به‌سادگی متوجه عامل انسانی خواهد بود.

برای مثال، فرض کنید یک هکر از یک سامانه هوش مصنوعی برای سرقت وجوه از حساب‌های بانکی استفاده کند. در چنین

وضعیتی، جرم ارتکابی همچنان توسط هکر صورت گرفته است و جرایمی مانند کلاهبرداری، سرقت رایانه‌ای یا جرایم سایبری موجود برای تعقیب او کفایت می‌کنند.

حتی اگر فناوری‌های جدید مستلزم جرم‌انگاری برخی رفتارهای نوظهور باشند، حقوق کیفری همچنان ابزارهای شناخته‌شده‌ای برای انتساب مسئولیت به اشخاص انسانی در اختیار دارد.

نظریه «عامل بی‌گناه» و انتساب مسئولیت به انسان

یکی از این ابزارها، دکترین عامل بی‌گناه است. بر اساس این نظریه، هرگاه شخصی از موجود یا فردی فاقد اهلیت کیفری برای ارتکاب جرم استفاده کند، مسئولیت کیفری متوجه شخص هدایت‌کننده خواهد بود.

برای مثال، اگر فردی یک کودک پنج‌ساله را برای انتقال مواد مخدر به کار گیرد، مسئولیت کیفری بر عهده بزرگسال خواهد بود، نه کودک؛ زیرا کودک فاقد ظرفیت لازم برای مسئولیت کیفری است.

به همین قیاس، اگر شخصی هوش مصنوعی را به‌گونه‌ای برنامه‌ریزی کند که مرتکب رفتار مجرمانه شود، مسئولیت کیفری متوجه برنامه‌نویس یا کاربر خواهد بود، نه خود سامانه هوش مصنوعی.

البته اعمال این نظریه مستلزم آن است که شخص حداقل علم یا قصد وقوع نتیجه مجرمانه را داشته باشد.

۴-۴-۲- مسئولیت مبتنی بر بی‌احتیاطی و بی‌مبالاتی

اگر توسعه‌دهنده یا کاربر قصد ارتکاب جرم نداشته باشد چه؟ در چنین مواردی نیز حقوق کیفری ابزارهای دیگری در اختیار دارد.

اگر خطر ایجاد آسیب از سوی هوش مصنوعی برای شخص قابل پیش‌بینی بوده باشد، می‌توان از مسئولیت مبتنی بر بی‌احتیاطی یا بی‌مبالاتی آگاهانه استفاده کرد.

برای مثال، اگر توسعه‌دهندگان یک سامانه هوش مصنوعی از وجود احتمال جدی و غیرموجه مرگ اشخاص آگاه باشند، اما با وجود این خطر به بهره‌برداری از سامانه ادامه دهند، ممکن است به قتل ناشی از بی‌مبالاتی یا سایر جرایم غیرعمدی محکوم شوند.

حتی اگر خطر مستقیماً پیش‌بینی نشده باشد، اما یک شخص متعارف در همان شرایط باید آن را پیش‌بینی می‌کرد، باز هم اشکال خفیف‌تر مسئولیت کیفری قابل اعمال خواهد بود.

همین منطق در خصوص جرایمی مانند سرقت، تخریب اموال، خسارات مالی یا سایر آسیب‌های ناشی از هوش مصنوعی نیز قابل استفاده است.

نظریه پیامدهای طبیعی و محتمل

گابریل هالوی (Hallevy, 2010) شکل دیگری از مسئولیت کیفری را مطرح می‌کند که آن را «مدل پیامدهای طبیعی و محتمل» می‌نامد.

فرض کنید شخصی عمداً از هوش مصنوعی برای ارتکاب یک جرم استفاده می‌کند، اما سامانه هوش مصنوعی فراتر از برنامه اولیه عمل کرده و جرایم دیگری نیز مرتکب می‌شود.

در این حالت نیز می‌توان استدلال کرد که شخص انسانی مسئول نتایج قابل پیش‌بینی ناشی از استفاده مجرمانه از هوش مصنوعی است.

به اعتقاد نویسندگان، در چنین مواردی نیز قواعد موجود مسئولیت کیفری ظرفیت کافی برای انتساب مسئولیت به اشخاص انسانی را دارند و خلأ واقعی مشاهده نمی‌شود.

آسیب‌های غیرقابل پیش‌بینی ناشی از هوش مصنوعی

سناریوی دشوارتر زمانی است که هوش مصنوعی خسارتی ایجاد کند که حتی برای کاربران یا توسعه‌دهندگان نیز قابل پیش‌بینی نبوده باشد.

برای مثال، فرض کنید هکرها با استفاده از هوش مصنوعی قصد سرقت ارزهای دیجیتال را دارند، اما سامانه به شکلی غیرمنتظره باعث از کار افتادن شبکه برق یک منطقه شده و خسارات گسترده مالی و جانی ایجاد می‌کند.

در نگاه نخست ممکن است به نظر برسد که مرتکبان نسبت به این نتایج مسئولیتی ندارند؛ زیرا این پیامدها قابل پیش‌بینی نبوده‌اند. اما حقوق کیفری برای این وضعیت نیز ابزارهایی در اختیار دارد.

جرایم مبتنی بر مسئولیت سازه‌ای (Constructive Liability)

یکی از این ابزارها، مسئولیت کیفری سازه‌ای یا جرایم مبتنی بر مسئولیت تبعی است.

نمونه کلاسیک آن در حقوق آمریکا، «قتل ناشی از ارتکاب جنایت» است.

برای مثال، فردی با قصد سرقت وارد منزلی می‌شود و صاحب‌خانه بر اثر شوک ناشی از حضور او دچار حمله قلبی شده و فوت می‌کند. حتی اگر مرگ قربانی برای سارق قابل پیش‌بینی نبوده باشد، ممکن است به قتل محکوم شود.

در اینجا مسئولیت قتل از ترکیب دو عنصر شکل می‌گیرد:

ارتکاب جرم پایه (سرقت یا ورود غیرقانونی).

وقوع نتیجه زیان‌بار (مرگ).

در چنین جرایمی، برای نتیجه نهایی لزوماً نیاز به اثبات سوءنیت یا حتی بی‌احتیاطی وجود ندارد.

نویسندگان معتقدند که همین منطق را می‌توان در حوزه هوش مصنوعی نیز به کار گرفت.

برای مثال، قانون‌گذار می‌تواند جرمی تحت عنوان: ایجاد خسارت از طریق استفاده مجرمانه از هوش مصنوعی ایجاد کند که در آن:

- استفاده مجرمانه از هوش مصنوعی جرم پایه باشد؛
- و خسارات گسترده جانی یا مالی ناشی از عملکرد سامانه، عنصر تکمیلی جرم را تشکیل دهد.

در این صورت، حتی آسیب‌های غیرقابل پیش‌بینی نیز می‌توانند تحت پوشش نظام کیفری قرار گیرند.

نویسندگان در پایان این بخش نتیجه می‌گیرند که در بخش بزرگی از موارد، آسیب‌های ناشی از هوش مصنوعی را می‌توان از طریق ابزارهای سنتی حقوق کیفری مدیریت کرد. نظریه عامل بی‌گناه، مسئولیت ناشی از بی‌احتیاطی و بی‌مبالاتی، قواعد معاونت و مشارکت در جرم و همچنین جرایم مبتنی بر مسئولیت سازه‌ای، همگی امکان انتساب مسئولیت به اشخاص انسانی را فراهم می‌کنند.

بنابراین، در مواردی که رفتار هوش مصنوعی به‌طور مستقیم یا غیرمستقیم قابل انتساب به توسعه‌دهندگان، برنامه‌نویسان، مالکان یا کاربران باشد، نیازی به شناسایی مسئولیت کیفری مستقل برای خود هوش مصنوعی وجود ندارد. از این رو، بخش قابل توجهی از آنچه به‌عنوان «خلأ کیفری هوش مصنوعی» مطرح می‌شود، در واقع با استفاده از سازوکارهای موجود حقوق کیفری قابل پاسخ‌گویی است و مستلزم ایجاد شخصیت کیفری مستقل برای سامانه‌های هوش مصنوعی نیست (Hallevy, 2010).

خلأ واقعی مسئولیت کیفری هوش مصنوعی: جرایم غیرقابل انتساب به انسان

نویسندگان برای تبیین «خلأ کیفری هوش مصنوعی»، مثالی فرضی را مطرح می‌کنند که سامانه‌ای هوشمند با عنوان «خریدار خودکار هاروارد» طراحی شده باشد که وظیفه آن خرید کتاب‌ها و تجهیزات مورد نیاز دانشجویان جدیدالورود است. این سامانه در جریان آموزش بر روی داده‌های مربوط به گفت‌وگوهای اینترنتی دانشجویان مهندسی، به شکلی کاملاً غیرقابل پیش‌بینی یاد می‌گیرد که مواد رادیواکتیو را از بازارهای غیرقانونی اینترنت (Dark Web) خریداری کرده و به خوابگاه‌های دانشجویی ارسال کند. در نتیجه این اقدام، تعدادی از دانشجویان جان خود را از دست می‌دهند.

در این فرض، برنامه‌نویسان و طراحان سامانه هیچ رفتار مجرمانه‌ای انجام نداده‌اند و اهداف آن‌ها نیز کاملاً مشروع بوده است. همچنین، تمامی اقدامات احتیاطی متعارف در طراحی، آموزش و بهره‌برداری

از سامانه رعایت شده است. با این حال، سامانه هوش مصنوعی موجب وقوع خسارت‌های شدید و مرگبار شده است.

در چنین وضعیتی، برخلاف مثال‌های پیشین، هیچ شخص انسانی وجود ندارد که بتوان مسئولیت کیفری را به او نسبت داد. نظریه «عامل بی‌گناه» قابل اعمال نیست؛ زیرا توسعه‌دهندگان، مالکان و کاربران سامانه نه قصد ارتکاب جرم داشته‌اند و نه وقوع نتیجه زیان‌بار را پیش‌بینی می‌کرده‌اند. مسئولیت ناشی از بی‌احتیاطی یا بی‌مبالاتی نیز قابل اعمال نیست؛ زیرا خطر وقوع این نتیجه اساساً قابل پیش‌بینی نبوده است. همچنین، مسئولیت کیفری سازه‌ای نیز کاربردی ندارد؛ زیرا هیچ جرم پایه یا رفتار تقصیرآمیز اولیه‌ای از سوی اشخاص انسانی وجود ندارد که بتوان مسئولیت را بر آن بنا کرد.

در چنین شرایطی است که نویسندگان از وجود یک «خلأ کیفری واقعی» سخن می‌گویند؛ یعنی وضعیتی که در آن رفتار زیان‌بار هوش مصنوعی نه به خود هوش مصنوعی قابل انتساب کیفری است و نه به هیچ شخص انسانی.

ممکن است پیشنهاد شود که برای رفع این مشکل، جرایم جدیدی برای توسعه‌دهندگان هوش مصنوعی تعریف شود؛ برای مثال، هرگونه طراحی سامانه‌ای که احتمال ایجاد خطرات جدی داشته باشد، جرم‌انگاری شود. اما نویسندگان معتقدند چنین رویکردی با اصول بنیادین حقوق کیفری ناسازگار است؛ زیرا تقریباً تمامی فناوری‌ها و فعالیت‌های انسانی واجد نوعی خطر بالقوه هستند. اگر چنین مسئولیتی پذیرفته شود، بسیاری از فناوری‌های نوین ـ از جمله اینترنت ـ هرگز امکان توسعه نمی‌یافتند. بنابراین، گسترش بی‌ضابطه مسئولیت کیفری می‌تواند موجب سرکوب نوآوری و فعالیت‌های مفید اقتصادی شود.

هزینه‌های مجازات مستقیم هوش مصنوعی

حتی اگر بپذیریم که در برخی موارد خلأ کیفری واقعی وجود دارد، باز هم این نتیجه به معنای توجیه مجازات مستقیم هوش مصنوعی نیست. نویسندگان معتقدند که مجازات هوش مصنوعی با چالش‌های عملی و نظری بسیار جدی روبه‌رو است.

یکی از مهم‌ترین این چالش‌ها، مسئله عنصر معنوی جرم است. در حقوق کیفری سنتی، میزان سرزنش‌پذیری مرتکب بر اساس وضعیت ذهنی او ارزیابی می‌شود. برای مثال، ارتکاب یک رفتار با قصد مجرمانه معمولاً شدیدتر از ارتکاب همان رفتار بر اثر بی‌احتیاطی تلقی می‌شود. اما پرسش اینجاست که چگونه می‌توان برای یک سامانه هوش مصنوعی «قصد»، «علم»، «آگاهی» یا سایر حالات ذهنی را تصور کرد؟

نویسندگان پیش‌تر توضیح داده‌اند که در برخی موارد می‌توان وضعیت ذهنی توسعه‌دهندگان یا مالکان را به هوش مصنوعی منتسب کرد؛ مشابه آنچه در مسئولیت کیفری شرکت‌ها رخ می‌دهد. اما در مواردی که رفتار زیان‌بار هوش مصنوعی مستقل از هرگونه تقصیر انسانی باشد، چنین راهکاری قابل استفاده نیست.

یکی از راه‌حل‌های احتمالی، ایجاد نظام مسئولیت مطلق کیفری برای هوش مصنوعی است. با این حال، تحقق چنین راهکاری مستلزم اصلاحات گسترده در قوانین کیفری و پذیرش این فرض حقوقی است که هوش مصنوعی می‌تواند همانند انسان مرتکب «رفتار ارادی» شود. راهکار دیگر، ایجاد یک مفهوم حقوقی جدید از سوءنیت مصنوعی است؛ اما این رویکرد نیز مستلزم توسعه گسترده نظریه‌های حقوقی و استفاده مداوم از کارشناسان فنی برای تحلیل عملکرد سامانه‌های هوش مصنوعی خواهد بود.

افزون بر این، مشکلات عملی فراوان دیگری نیز وجود دارد؛ برای مثال، اجرای مجازات در مورد سامانه‌های مبتنی بر فناوری بلاکچین که غیرمتمرکز هستند و به‌آسانی قابل غیرفعال‌سازی یا نابودی نیستند.

شخصیت حقوقی هوش مصنوعی و پیامدهای آن

به اعتقاد نویسندگان، مجازات مستقیم هوش مصنوعی مستلزم اعطای نوعی شخصیت حقوقی به آن است. همان‌گونه که مسئولیت

کیفری شرکت‌ها مبتنی بر شناسایی شخصیت حقوقی مستقل برای شرکت‌هاست، مسئولیت کیفری هوش مصنوعی نیز بدون اعطای چنین شخصیتی امکان‌پذیر نخواهد بود.

با این حال، اعطای شخصیت حقوقی به هوش مصنوعی با نگرانی‌های جدی همراه است. در سال ۲۰۱۷، پارلمان اروپا از کمیسیون اروپا خواست امکان ایجاد نوعی شخصیت حقوقی الکترونیکی برای ربات‌ها را بررسی کند. این پیشنهاد با واکنش منفی گسترده حقوقدانان و متخصصان هوش مصنوعی مواجه شد و بیش از ۱۵۰ پژوهشگر طی نامه‌ای سرگشاده با آن مخالفت کردند.

نویسندگان تأکید می‌کنند که اعطای شخصیت حقوقی لزوماً به معنای برخورداری از تمامی حقوق انسان‌ها نیست. همان‌گونه که شرکت‌ها دارای برخی حقوق و تکالیف خاص هستند، می‌توان برای هوش مصنوعی نیز شخصیت حقوقی محدودی تصور کرد. با این حال، حتی چنین شخصیت محدودی نیز می‌تواند آثار نامطلوبی به همراه داشته باشد.

از جمله این آثار می‌توان به تقویت پدیده «انسان‌انگاری» اشاره کرد؛ یعنی گرایش افراد به تصور هوش مصنوعی به‌عنوان موجودی مشابه انسان. این امر ممکن است انتظارات غیرواقع‌بینانه‌ای درباره توانایی‌ها و مسئولیت‌های هوش مصنوعی ایجاد کند و حتی موجب آسیب‌های روانی و شناختی برای کاربران شود.

همچنین، ممکن است اعطای شخصیت حقوقی به هوش مصنوعی زمینه‌ساز توسعه تدریجی حقوق آن‌ها شود؛ پدیده‌ای که نویسندگان از آن با عنوان حقوق‌افزایی تدریجی یاد می‌کنند. همان‌گونه که در طول زمان دامنه حقوق شرکت‌ها گسترش یافته است، ممکن است در آینده نیز برای هوش مصنوعی حقوق بیشتری مطالبه شود؛ موضوعی که می‌تواند آزادی‌ها و منافع انسانی را محدود کند.

راهکار دوم: توسعه محدود حقوق کیفری

نویسندگان معتقدند که به جای مجازات مستقیم هوش مصنوعی، می‌توان از اصلاحات محدود و هدفمند در حقوق کیفری استفاده کرد.

برای مثال، همان‌گونه که قوانین خاصی برای جرایم رایانه‌ای تصویب شده‌اند، می‌توان قانونی تحت عنوان «سوءاستفاده از هوش مصنوعی» تصویب کرد که رفتارهایی مانند طراحی غیرمسئولانه، استقرار ناایمن یا نظارت ناکافی بر سامانه‌های هوش مصنوعی را جرم‌انگاری کند.

اما مهم‌ترین پیشنهاد نویسندگان، ایجاد نهاد «شخص مسئول» است. بر اساس این طرح، برای هر سامانه هوش مصنوعی یک شخص مسئول تعیین می‌شود. این شخص می‌تواند تولیدکننده، مالک، توسعه‌دهنده یا کاربر سامانه باشد. در صورت وقوع خسارات ناشی از جرایم پیچیده هوش مصنوعی، مسئولیت قانونی متوجه این شخص خواهد بود. به اعتقاد نویسندگان، این راهکار مزایای متعددی دارد؛ زیرا بدون نیاز به اعطای شخصیت حقوقی مستقل به هوش مصنوعی، امکان پاسخ‌گویی حقوقی را فراهم می‌کند و در عین حال از بسیاری از مشکلات نظری و عملی مسئولیت کیفری مستقیم هوش مصنوعی جلوگیری می‌نماید.

راهکار سوم: توسعه مسئولیت مدنی

راهکار دیگر، تقویت نظام مسئولیت مدنی است. نویسندگان معتقدند بسیاری از خسارات ناشی از هوش مصنوعی را می‌توان از طریق قواعد مسئولیت مدنی، مسئولیت ناشی از محصول معیوب، بیمه اجباری یا صندوق‌های جبران خسارت مدیریت کرد. در مواردی که مسئولیت مدنی سنتی کفایت نکند، می‌توان همان مدل «شخص مسئول» را در حوزه مدنی نیز به کار گرفت؛ به‌گونه‌ای که شخص مسئول ملزم به جبران خسارات ناشی از عملکرد سامانه هوش مصنوعی باشد.

همچنین، می‌توان صندوق‌های جبران خسارت مشابه صندوق‌های جبران خسارت ناشی از واکسن یا انرژی هسته‌ای ایجاد کرد تا قربانیان بدون نیاز به اثبات تقصیر، خسارت خود را دریافت کنند.

نتیجه‌گیری

اصول کلی حقوق کیفری که در طول سده‌ها در نظام‌های حقوقی گوناگون تکوین یافته‌اند، بر پایه مفاهیمی چون شخص حقیقی، اراده، قصد مجرمانه و رفتار مادی، تقصیر کیفری و اهلیت مسئولیت کیفری استوارند. بر این اساس، پاسخ بنیادین به پرسش «آیا هوش مصنوعی قابل مجازات است؟» منفی است؛ زیرا:

نخست، اصل شخصی بودن مجازات (شخصی بودن کیفر) ایجاب می‌کند که فقط کسی مجازات شود که خود مرتکب جرم شده و از عقل و اراده برخوردار باشد. هوش مصنوعی نه «شخص حقیقی» است و نه «شخص حقوقی» به معنای متعارف؛ بلکه ابزاری است که توسط انسان طراحی، برنامه‌ریزی و به کار گرفته می‌شود. انتساب مجازات به آن نقض اصل شخصی بودن کیفر است.

دوم، عنصر معنوی (قصد، علم، عمد یا لااقل تقصیر جزایی) شرط تحقق مسئولیت کیفری در تمام نظام‌های حقوقی پیشرفته است. هوش مصنوعی فاقد «آگاهی درونی»، «قصد نتیجه»، «درک نادرستی رفتار» و «اراده در انتخاب» می‌باشد. رفتارهای حاصل از الگوریتم‌های یادگیری ماشین اگرچه ممکن است ظاهراً هدفمند به نظر برسند، اما از جنس «واکنش بر پایه محاسبات آماری» هستند، نه «انتخاب خودآگاه مبتنی بر ارزش‌های اخلاقی و هنجارهای کیفری».

سوم، اصل قانونی بودن جرائم و مجازات‌ها اقتضا می‌کند که جرم و کیفر آن پیشاپیش در قانون معین شده باشد. هیچ قانون کیفری در نظام‌های بین‌المللی و داخلی، هوش مصنوعی را مستقیماً «عامل جرم» و مخاطب مجازات قرار نداده است. تفسیر موسع برای شمول هوش مصنوعی برخلاف اصل تفسیر مضیق در قوانین کیفری خواهد بود.

چهارم، هدف‌های مجازات (سزادهی، بازدارندگی، اصلاح، بازپذیرسازی) درباره هوش مصنوعی معنا ندارند. هوش مصنوعی «رنج نمی‌برد»، «انگیزه تخلف ندارد»، «قابل اصلاح از طریق آموزش مجدد است نه تنبیه» و «جامعه در بازدارندگی آن نقشی ندارد». بنابراین، اعمال مجازات بر آن، عبث و خارج از فلسفه کیفر است.

با این حال، نفی مجازات کیفری به معنای نفی هرگونه پاسخ‌دهی نیست. بر اساس اصول کلی حقوق (چه مدنی و چه اداری) و با تکیه بر نظریه‌های «مسئولیت محض» (حتی در برخی نظام‌های کامن‌لا) و «مسئولیت ناشی از خطر»، می‌توان راهکارهای زیر را جایگزین مجازات کیفری مستقیم هوش مصنوعی نمود:

۱. مسئولیت مدنی فرافکنی‌شده به پشتیبانان انسانی (طراح، تولیدکننده، مالک، کاربر نهایی یا همه به‌صورت تضامنی) بر مبنای قواعد تقصیر یا مسئولیت محض.

۲. پذیرش شخصیت حقوقی محدود برای هوش مصنوعی خودمختار پیشرفته (مانند تابعیت الکترونیک در برخی پیش‌نویس‌های حقوق اروپا)، به‌گونه‌ای که خسارت از منابع اختصاص‌یافته به آن (صندوق بیمه یا دارایی مجازی) جبران شود، اما عنوان «مجازات کیفری» بر آن اطلاق نگردد.

۳. اقدامات تأمینی و تربیتی مانند دستور به بازطراحی، به‌روزرسانی، محدودیت سطح دسترسی، قطع موقت از شبکه یا حذف کدهای آسیب‌رسان ـ این اقدامات اگرچه ممکن است در برخی نظام‌ها «مجازات» خوانده شوند، اما در اصول کلی حقوق کیفری «کیفر» تلقی نمی‌شوند، زیرا فاقد عناصر رنج‌آوری و سزادهی می‌باشند.

از منظر اصول کلی حقوق کیفری، مجازات مستقیم هوش مصنوعی ناممکن و ناموجه است. هرگونه پاسخ اجتماعی به رفتار زیان‌بار سامانه‌های خودمختار باید در قالب مسئولیت مدنی، مسئولیت اداری، بیمه اجباری، مقررات فنی و اقدامات تأمینی غیرکیفری

پیگیری شود. تغییر این اصول کلی نیازمند تحولی پارادایماتیک در تعریف «عامل جرم» و «مسئولیت کیفری» است که در افق فعلی حقوق کیفری جهان محتمل به نظر نمی‌رسد. در صورت تمایل به پیشنهاد چنین تحولی، باید به‌صراحت از بازتعریف مفاهیم بنیادین حقوق کیفری سخن گفت، نه تطبیق هوش مصنوعی با مفاهیم موجود.

در نهایت، نتیجه اصلی مقاله آن است که هرچند از نظر نظری می‌توان ساختاری برای مسئولیت کیفری مستقیم هوش مصنوعی طراحی کرد، اما هزینه‌های حقوقی، عملی و اجتماعی چنین رویکردی بسیار سنگین است.

نویسندگان معتقدند که به جای ایجاد شخصیت کیفری مستقل برای هوش مصنوعی، باید از راهکارهای کم‌هزینه‌تر و واقع‌بینانه‌تر استفاده کرد؛ از جمله:

- جرم‌انگاری محدود رفتارهای پرخطر مرتبط با هوش مصنوعی؛
- تعیین «شخص مسئول» برای سامانه‌های هوش مصنوعی؛
- توسعه مسئولیت مدنی و سازوکارهای جبران خسارت؛
- تقویت نظارت و تنظیم‌گری.

این راهکارها می‌توانند تقریباً همان مزایای مجازات مستقیم هوش مصنوعی را تأمین کنند، بدون آنکه مستلزم تغییرات بنیادین و پرمخاطره در ساختار سنتی حقوق کیفری باشند.

مشارکت نویسندگان

در نگارش این مقاله تمامی نویسندگان نقش یکسانی ایفا کردند.

تعارض منافع

در انجام مطالعه حاضر، هیچ‌گونه تضاد منافعی وجود ندارد.

EXTENDED ABSTRACT

Artificial intelligence (AI) is increasingly involved in activities that may generate consequences traditionally addressed by criminal law, including autonomous vehicle collisions, algorithmic financial misconduct, cyber-enabled offenses, unlawful data processing, harmful robotic behavior, and autonomous decision-making in contexts involving human life and property. This development has created a fundamental conceptual challenge for contemporary criminal law: whether AI should continue to be treated merely as an instrument through which human beings act or whether, under certain circumstances, it may be regarded as an independent agent capable of bearing criminal responsibility. The problem becomes particularly acute when an autonomous system produces harmful or unlawful outcomes that cannot realistically be attributed to a programmer, developer, owner, operator, or user. Conventional criminal liability is built on concepts such as voluntary conduct, culpability, intention, knowledge, recklessness, negligence, capacity, and moral blameworthiness, all of which were historically formulated in relation to human actors. AI systems, however, may display complex, adaptive, and apparently goal-directed behavior without possessing consciousness, moral agency, free will, or subjective awareness in the human sense. Consequently, the attribution of criminal intent and blameworthiness to such systems remains deeply controversial. Iranian criminal law presents additional complexities because its conceptual structure is strongly influenced by principles of Islamic jurisprudence that emphasize the responsibility of a legally accountable human agent possessing understanding and volition. Therefore, although AI systems may generate conduct functionally resembling intentional human action, their lack of traditional legal and moral capacity creates substantial obstacles to direct criminal liability (Amirian Farsani & Hosseini, 2023; Ansari & Atazadeh, 2019; Kaveh & Barani, 2024). At the same time, completely excluding autonomous systems from the field of criminal law may create a responsibility gap when

harmful outcomes occur without identifiable human fault. This tension between technological autonomy and traditional doctrines of attribution constitutes the central concern of the present study, which examines whether AI should be conceptualized as a tool, an agent of crime, or an entity requiring an entirely different legal framework (Hallevy, 2010; Makki et al., 2024).

The study employs an analytical-comparative method to examine the theoretical foundations of punishment, criminal responsibility, attribution, and legal personality as they relate to autonomous AI systems. The analysis begins with the conventional theory of punishment, according to which criminal punishment is generally imposed in response to legally prohibited conduct, directed at the person regarded as responsible for the offense, deliberately administered through an authorized legal institution, and accompanied by consequences considered burdensome or condemnatory. Within this framework, punishment is justified not only by its possible social benefits but also by normative constraints concerning desert, proportionality, culpability, fairness, and the moral status of the person punished. Contemporary punishment theory typically recognizes utilitarian or consequentialist functions such as deterrence, incapacitation, rehabilitation, and prevention, while also acknowledging retributive and expressive functions connected with blame, condemnation, and deserved suffering. These theories are important to AI because they provide the criteria for determining whether direct punishment of an artificial system can be justified at all. If an AI system cannot experience suffering, appreciate condemnation, understand legal prohibitions, or modify its behavior because it recognizes the threat of punishment, many conventional purposes of criminal punishment become difficult to apply. Nevertheless, some functions of punishment may still operate indirectly. For example, sanctioning an AI system may deter developers, owners, investors, and users by increasing the financial and operational costs associated with unsafe system design or deployment. Similarly, public condemnation of harmful AI conduct may have symbolic value by acknowledging the seriousness of harm and reinforcing societal commitment to legal norms (Alschuler, 2009; Darling, 2016). Yet such benefits remain constrained by the principle that punishment should not exceed the degree of culpability attributable to the offender and that criminal condemnation normally presupposes an entity capable of bearing responsibility (Müller & Bostrom, 2016; Sarch). The study therefore evaluates AI criminal liability through three connected criteria: positive justifications for punishment, negative limitations grounded in culpability and desert, and the existence of more effective and less disruptive alternatives. The analysis demonstrates that many cases commonly described as AI crimes do not in fact require recognition of AI as an independent criminal subject because existing doctrines remain capable of attributing responsibility to human actors. When a programmer intentionally designs an AI system to commit fraud, theft, cyber intrusion, manipulation, or violence, the AI functions essentially as an instrument of the human offender. Similarly, where a user intentionally deploys an autonomous system to achieve a criminal purpose, ordinary doctrines of perpetration, complicity, causation, and indirect responsibility can often identify the relevant human agent. Even where there is no direct criminal intention, developers, manufacturers, operators, or owners may still incur liability when harmful outcomes are foreseeable and result from reckless, negligent, or otherwise culpable conduct. The use of autonomous vehicles provides a particularly significant example because

criminal responsibility may arise from unsafe programming, inadequate testing, unreasonable deployment, failure to monitor known risks, or conscious disregard of foreseeable danger (Ansari & Atazadeh, 2019; Gurney, 2015). A similar logic applies to advanced robotic systems and other machine-learning applications whose operation creates substantial risks for individuals or property. Hallevy's model of criminal responsibility for AI further illustrates how existing doctrines may be extended to situations in which AI systems function as innocent agents or where human actors are responsible for the natural and probable consequences of deploying them for unlawful purposes (Hallevy, 2010). Accordingly, the apparent responsibility gap is much narrower than is sometimes suggested. In a large proportion of cases, human responsibility can still be constructed through intention, knowledge, recklessness, negligence, causation, complicity, or forms of constructive liability. The real difficulty arises only in exceptional situations in which an autonomous system, despite being properly designed, lawfully deployed, reasonably supervised, and used without human fault, develops or produces conduct that results in serious prohibited harm that no human actor could reasonably have anticipated. It is precisely these cases of non-attributable autonomous harm that create the strongest theoretical argument for reconsidering whether AI itself should become a subject of criminal law.

Despite the existence of such responsibility gaps, the study finds that direct criminal liability for AI faces substantial doctrinal and practical obstacles. The most important concerns relate to the mental element of crime, moral agency, legal capacity, and the theoretical meaning of punishment. Traditional criminal law distinguishes intentional wrongdoing from reckless or negligent conduct because the offender's mental state is central to the assessment of blameworthiness. Yet AI systems do not possess subjective intention, consciousness, remorse, moral understanding, or awareness in the ordinary human sense, even where their outputs appear purposeful or intelligent. Their behavior emerges from computational processing, algorithmic optimization, statistical inference, learned representations, and adaptive responses rather than from morally reflective decision-making (Amirian Farsani & Hosseini, 2023; Müller & Bostrom, 2016). One possible response would be to create a form of strict criminal liability for AI, thereby dispensing with the need to prove a traditional mental state. Another would be to construct a doctrine of artificial mens rea based on technical states, system goals, internal decision parameters, or computational representations. However, both approaches would require profound modifications to established concepts of criminal responsibility. Similar difficulties arise in relation to punishment itself. An AI system cannot meaningfully experience imprisonment, shame, remorse, suffering, or moral condemnation in the same way as a human being. Measures such as deletion, shutdown, access restriction, code modification, network isolation, or forced retraining could certainly be imposed, but their classification as criminal punishment remains conceptually uncertain. Furthermore, attributing legal personality to AI for the purpose of criminal liability may produce consequences beyond the immediate field of punishment. Experience with corporate legal personality demonstrates that the creation of legal personhood may gradually generate broader claims concerning rights, legal standing, ownership, procedural protection, and institutional recognition (Alschuler, 2009). The anthropomorphic treatment of sophisticated robots may also encourage users to perceive artificial systems as morally

equivalent to human beings, with uncertain psychological and normative consequences (Darling, 2016; Scheutz, 2012). For these reasons, direct AI criminal liability would entail not merely a technical amendment to criminal statutes but a significant restructuring of fundamental assumptions concerning agency, personhood, culpability, and punishment.

The comparative analysis therefore supports a more cautious and limited response centered on human responsibility, regulatory mechanisms, and civil liability rather than immediate recognition of independent AI criminal personhood. One possible approach is the enactment of narrowly tailored criminal provisions addressing particularly dangerous forms of AI-related conduct, including knowingly unsafe deployment, reckless system design, deliberate evasion of safety standards, failure to comply with mandatory monitoring requirements, or the use of AI for clearly unlawful purposes. Such provisions would preserve the traditional human-centered structure of criminal law while responding to technological risks through specifically formulated offenses (Kaveh & Barani, 2024; Makki et al., 2024). A second approach involves identifying a legally responsible person or entity for specified categories of autonomous systems. Depending on the circumstances, this responsible party could be the developer, manufacturer, owner, operator, deployer, or another designated actor who is legally required to ensure compliance, monitoring, insurance, and compensation. Such a model may reduce gaps in accountability while avoiding the need to attribute consciousness or criminal capacity to AI. A third and particularly important alternative lies in the expansion of civil liability. Existing tort principles, product liability doctrines, strict liability models, compulsory insurance, and compensation funds may often provide more effective remedies for victims than criminal prosecution, especially where the central objective is compensation rather than moral condemnation. The concept of the "reasonable computer" further suggests that liability standards may evolve in response to autonomous technologies without necessarily treating AI as a human-equivalent legal subject (Abbott, 2017). Regulatory supervision, technical standards, licensing systems, mandatory risk assessments, transparency requirements, audit obligations, and industry-specific safety rules may also reduce harmful outcomes more effectively than criminal sanctions. From this perspective, criminal law should remain a measure of last resort, particularly because overbroad criminalization of technological development may suppress socially beneficial innovation. The preferable strategy is therefore not to abandon criminal law but to recalibrate it carefully, expanding human accountability where justified while relying on civil, administrative, and technological governance mechanisms where they provide more proportionate and effective responses.

In conclusion, the study finds that although it is theoretically possible to design a system of direct criminal responsibility for artificial intelligence, such a development is neither presently necessary nor sufficiently justified. Most harmful acts involving AI can still be addressed through existing or moderately revised doctrines of human criminal responsibility, particularly where developers, owners, operators, or users intentionally, recklessly, or negligently contribute to the harmful outcome. A genuine responsibility gap exists only in a limited category of cases involving highly autonomous systems whose harmful actions are genuinely unforeseeable and cannot reasonably be attributed to any human participant. Even in such circumstances, recognizing AI as an independent criminal offender would require far-reaching modifications to the concepts of

criminal agency, mens rea, blameworthiness, legal personhood, proportionality, and punishment. Moreover, the practical benefits of such reform appear uncertain when compared with less disruptive alternatives. A more coherent legal policy would therefore combine narrowly targeted criminal-law reforms with strengthened civil liability, compulsory insurance, compensation mechanisms, regulatory supervision, technical standards, and the designation of responsible human or corporate actors for advanced AI systems. Measures such as deactivation, restriction of network access, retraining, redesign, software modification, and removal of harmful code may also be employed as preventive or regulatory interventions without necessarily classifying them as criminal punishments. Accordingly, contemporary criminal law should continue to treat AI primarily as a technologically autonomous instrument whose risks must be governed through carefully structured human and institutional accountability, rather than prematurely redefining it as a fully independent criminal subject. Only a fundamental transformation in the legal understanding of agency, personhood, consciousness, and responsibility could justify a different conclusion, and such a transformation would represent a paradigmatic reconstruction of criminal law rather than a simple extension of its existing doctrines.

References

Abbott, R. (2017). The Reasonable Computer: Disrupting the Paradigm of Tort Liability. *George Washington Law Review*, *86*, 1.

Alschuler, A. (2009). Two Ways to Think About the Punishment of Corporations. *American Criminal Law Review*, *46*, 1359.

Amirian Farsani, A., & Hosseini, S. M. (2023). Challenges and Barriers to Criminal Liability in Robots with Artificial Intelligence Capabilities. *Legal Civilization*(18).

Ansari, J., & Atazadeh, S. (2019). Re-Examining the Concept of Criminal Liability of Artificial Intelligence: A Case Study of Self-Driving Vehicles in Islamic, Iranian, U.S., and German Law. *Comparative Research in Islamic and Western Law*(4).

Darling, K. (2016). Extending Legal Protection to Social Robots: The Effects of Anthropomorphism, Empathy, and Violent Behavior Towards Robotic Objects. In *Robot Law* (pp. 213-215).

Gurney, J. (2015). Driving into the Unknown: Examining the Crossroads of Criminal Law and Autonomous Vehicles. *Wake Forest Journal of Law & Policy*, *5*, 393-433.

Hallevy, G. (2010). The Criminal Liability of Artificial Intelligence Entities: From Science Fiction to Legal Social Control. *Akron Intellectual Property Journal*, *4*, 171-191.

Kaveh, M. H., & Barani, M. (2024). Criminal Liability of Artificial Intelligence in Iranian Criminal Law with a View to European Union Legislation. *Comparative Criminal Jurisprudence*(3).

Makki, A. a.-S., Makki, Z. a.-S., & Kashkoulian, E. (2024). Examining Liability Arising from Artificial Intelligence Actions in the Iranian Legal System. *Jurisprudence, Law and Criminal Sciences*(32).

Müller, V. C., & Bostrom, N. (2016). Future Progress in Artificial Intelligence: A Survey of Expert Opinion. In *Fundamental issues of artificial intelligence* (pp. 553).

Sarch, A. Who Cares What You Think? The Irrelevance of Unmanifested Mental States. *Law and Philosophy*, *36*, 707.

Scheutz, M. (2012). The Inherent Dangers of Unidirectional Emotional Bonds Between Humans and Social Robots. In *Robot Ethics: The Ethical and Social Implications of Robotics* (pp. 205-222).