University of Tehran Press

Comparative Law Review

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

Online ISSN: 2423-3404

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

Autonomy: The Indefensible Assumption of Artificial Intelligence Criminal Liability Theory

Abbas Shiri[1]✉ | **Mohammadreza Barzegar**[2]

1. Corresponding author; Associate Professor, Department of Criminal Law and Criminology, Faculty of Law and Political Science, University of Tehran, Tehran, Iran. Email: ashiri@ut.ac.ir
2. PhD Student in Criminal Law and Criminology, Department of Criminal Law and Criminology, Faculty of Law and Political Science, University of Tehran, Tehran, Iran. Email: mrbarzegar@ut.ac.ir

Article Info

Article Type:
Research Article

Received:
2025/03/11
Received in revised form:
2025/05/29
Accepted:
2025/10/02
Published online:
2026/08/23

Keywords:
Artificial Intelligence, Autonomy, Agency, Criminal Liability.

Abstract

The theory of independent criminal liability for artificial intelligence, as an emerging paradigm in criminal law, requires establishing the autonomy of this technology in the commission of a crime, since a system that operates as no more than an instrument in the hands of its developer or user is regarded merely as a means of committing the offence, and liability will accordingly attach to the human agent. The present article addresses the antecedent question of whether artificial intelligence fundamentally possesses the attribute that constitutes the precondition of legal agency. To this end, adopting a descriptive-analytical approach and drawing upon library-based sources of both a legal and a technical nature, the various dimensions of AI autonomy, including emergent behaviour and explainability, are examined as criteria for assessing independent agency. The analysis of criminal incidents associated with such systems, undertaken as an empirical test for gauging the degree of autonomy actually manifested, likewise confirms the conclusion that contemporary artificial intelligence, even in its most advanced forms and notwithstanding its capacities for learning, adaptability, and the emergence of novel behaviours, remains confined within the framework of ends predetermined by human agents, and that its autonomy is best characterised as a "limited and heteronomous autonomy"; an autonomy which, although it poses challenges for the criminal justice system, falls short of the threshold of "independent agency", namely the fundamental precondition for artificial intelligence to qualify as a proper addressee of any model of liability. Accordingly, so long as this threshold remains unattained, treating artificial intelligence as an independent perpetrator will be devoid of any legal foundation.

How To Cite

Shiri, Abbas; Barzegar, Mohammadreza (2026). Autonomy: The Indefensible Assumption of Artificial Intelligence Criminal Liability Theory. *Comparative Law Review*, 17 (1), 145-171.
DOI: https://doi.com/10.22059/jcl.2025.390243.634723

DOI

10.22059/jcl.2025.390243.634723

Publisher

شاپا الکترونیکی: ۳۴۰۴-۲۴۲۳

مطالعات حقوق تطبیقی

دوره: ۱۷، شماره: ۱
بهار و تابستان ۱۴۰۵

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

University of Tehran Press

استقلال، پیش‌فرض غیرقابل دفاع نظریهٔ مسئولیت کیفری هوش مصنوعی*

عباس شیری[۱] ✉ | **محمدرضا برزگر**[۲]

۱. نویسندهٔ مسئول؛ دانشیار گروه حقوق جزا و جرم‌شناسی، دانشکدهٔ حقوق و علوم سیاسی، دانشگاه تهران، تهران، ایران. رایانامه: ashiri@ut.ac.ir

۲. دانشجوی دکتری حقوق کیفری و جرم‌شناسی، دانشکدهٔ حقوق و علوم سیاسی، دانشگاه تهران، تهران، ایران. رایانامه: mrbarzegar@ut.ac.ir

اطلاعات مقاله

نوع مقاله: پژوهشی

تاریخ دریافت:
۱۴۰۳/۱۲/۲۱

تاریخ بازنگری:
۱۴۰۴/۰۳/۰۸

تاریخ پذیرش:
۱۴۰۴/۰۷/۱۰

تاریخ انتشار برخط:
۱۴۰۵/۰۶/۰۱

کلیدواژه‌ها:
استقلال، عاملیت، مسئولیت کیفری، هوش مصنوعی.

چکیده

نظریهٔ مسئولیت کیفری مستقل برای هوش مصنوعی به‌عنوان پارادایمی نوپدید در حقوق کیفری، نیازمند احراز استقلال این فناوری در ارتکاب جرم است؛ چراکه سامانه‌ای که فراتر از ابزاری در دست توسعه‌دهنده یا کاربر عمل نکند، صرفاً وسیلهٔ ارتکاب آن محسوب می‌شود و مسئولیت آن متوجهٔ عامل انسانی خواهد بود. در نوشتار حاضر به این پرسش پرداخته می‌شود که آیا هوش مصنوعی اساساً واجد آن خصیصه‌ای است که پیش‌شرط فاعلیت حقوقی به‌شمار آید. به این منظور با رویکردی توصیفی- تحلیلی و با بهره‌گیری از منابع کتابخانه‌ای، اعم از حقوقی و فنی، جنبه‌های گوناگون استقلال هوش مصنوعی، ازجمله رفتار نوظهور و قابلیت توضیح، به‌مثابه مؤلفه‌های سنجش عاملیت مستقل بررسی شده است. تحلیل رویدادهای زیان‌بار و مجرمانهٔ مرتبط با این سامانه‌ها نیز همچون آزمونی تجربی برای سنجش درجهٔ استقلال بروزیافته، این نتیجه را تأیید می‌کند که هوش مصنوعی کنونی، حتی در پیشرفته‌ترین اشکال خود و علی‌رغم برخورداری از قابلیت‌های یادگیری، انطباق‌پذیری و بروز رفتارهای بدیع، در چارچوب غایات ازپیش تعیین‌شدهٔ عاملان انسانی محصور می‌ماند و استقلال آن از سنخ «استقلال محدود و دگرآیین» است؛ استقلالی که هرچند چالش‌هایی برای نظام عدالت کیفری پدید می‌آورد، به آستانهٔ «عاملیت مستقل» نمی‌رسد. بدین‌سان، تا زمانی که این آستانه محقق نشده است، فاعل انگاشتن هوش مصنوعی به‌طور مستقل فاقد مبنا خواهد بود.

استناد
شیری، عباس؛ برزگر، محمدرضا (۱۴۰۵). استقلال، پیش‌فرض غیرقابل دفاع نظریهٔ مسئولیت کیفری هوش مصنوعی. *مطالعات حقوق تطبیقی*، ۱۷ (۱)، ۱۴۵-۱۷۱.
DOI: https://doi.com/10.22059/jcl.2025.390243.634723

DOI 10.22059/jcl.2025.390243.634723

ناشر مؤسسه انتشارات دانشگاه تهران.

* این مقاله مستخرج از رسالهٔ دکتری با عنوان «مبانی و قلمرو مسئولیت کیفری هوش مصنوعی؛ مطالعهٔ تطبیقی با اسناد اروپایی»، دانشکدهٔ حقوق و علوم سیاسی دانشگاه تهران است.

۱. مقدمه

پیشرفت‌های هوش مصنوعی[1]، ضمن ایجاد تحولات در عرصه‌های گوناگون علم و فناوری، پرسش‌هایی اساسی را در خصوص ماهیت و پیامدهای این پدیده مطرح ساخته است (Barzegar and Elham, 2020: 202). یکی از مهم‌ترین این پرسش‌ها، مسئلهٔ مسئولیت کیفری سامانه‌های هوش مصنوعی است. این مسئله، ریشه در پیچیدگی ذاتی هوش مصنوعی و فقدان قوانین و چارچوب‌های حقوقی متناسب با ویژگی‌های آن دارد و در واکنش به آن، نظریهٔ «مسئولیت کیفری هوش مصنوعی»، در ادبیات حقوقی مغرب‌زمین مطرح شده است که خاستگاه اصلی آن در آرای پروفسور گابریل هالوی[2] قابل ردیابی است (Hallevy, 2019:21; Kingston, 2016: 269; Kirpichnikov et al., 2020: 1).

اهمیت مقالهٔ حاضر در این است که مفهوم «استقلال» در کانون مباحث مسئولیت کیفری هوش مصنوعی جای دارد و هرگونه بحثی در این باره، مسبوق به واکاوی دقیق این مفهوم و بررسی ابعاد و مؤلفه‌های آن در سامانه‌های هوشمند است. این واکاوی به پرسشی بنیادین پاسخ می‌دهد مبنی بر اینکه آیا هوش مصنوعی به چنان سطحی از استقلال دست یافته است که بتوان آن را مسئول اَعمال خود دانست، یا همچنان باید آن را ابزاری در اختیار انسان تلقی کرد که مسئولیت جرایم احتمالی آن متوجهٔ سازندگان، برنامه‌نویسان و کاربران خواهد بود. پاسخ به این پرسش، جایگاه هوش مصنوعی را در نظام حقوقی کنونی و آتی تبیین می‌کند؛ چراکه اگر استقلال هوش مصنوعی احراز گردد، رابطهٔ سببیت میان فعل سازنده یا کاربر و نتیجهٔ مجرمانه منتفی و امکان انتساب جرم به خود هوش مصنوعی محتمل می‌شود. در مقابل، فقدان استقلال به معنای ابزار بودن این فناوری و تداوم همان رابطهٔ سببیت است که به تبع آن، امکان احراز مسئولیت کیفری (عمدی یا غیرعمدی) برای اشخاص مرتبط فراهم می‌آید (Najib Husni, 2007: 83).

از طرف دیگر با وجود گسترش مباحث مربوط به مسئولیت کیفری هوش مصنوعی در ادبیات حقوقی ایران، یک شکاف تحقیقاتی در این حوزه مشهود است؛ با این توضیح که خصیصهٔ «استقلال» به‌مثابه

---

1. Artificial Intelligence - AI

۲. پروفسور گابریل هالوی با ارائهٔ نظریهٔ «مسئولیت خود هوش مصنوعی» بنیانی نو در این عرصه بنا نهاد. این نظریه که در مقالات و کتب ایشان به تفصیل مورد بحث قرار گرفته است، استدلال می‌کند که در شرایطی که یک سیستم هوش مصنوعی، به صورت خودکار، مرتکب فعل زیان‌باری می‌شود، می‌توان خود آن سیستم را مسئول و پاسخگو شناخت. هالوی معتقد است که با افزایش استقلال عمل این سیستم‌ها، شناسایی مسئولیت مستقیم برای آن‌ها منطقی‌تر است. برای مطالعهٔ جامع این نظریه، نک: به این منبع:

– هالوی، گابریل (۱۳۹۸). مسئولیت کیفری ربات‌ها: هوش مصنوعی در قلمرو کیفری. مترجمان: فرهاد شاهیده و طاهره قوانلو. تهران: نشر میزان.

پیش‌شرط هرگونه اسناد مسئولیت، در پژوهش‌های داخلی به‌ندرت موضوع واکاوی مستقل قرار گرفته و اغلب این پژوهش‌ها، به‌جای پرداختن بنیادین به ماهیت و حدود و مؤلفه‌های این خصیصه، وجود آن را در سامانه‌های هوشمند مفروض و مسلّم انگاشته‌اند[۱]. همین پذیرش منفعلانه موجب شده است که پژوهش‌های یادشده، بدون احراز استقلال واقعی هوش مصنوعی به‌مثابه رکن رکین این بحث، مستقیماً به تجویز مسئولیت کیفری برای خود سامانه‌های هوشمند بپردازند.

این غفلت، پیامدهای نظری و عملی قابل‌توجهی در پی دارد؛ چنانچه هوش مصنوعی فاقد استقلال لازم برای تحقق مسئولیت کیفری باشد، طرح نظریهٔ مسئولیت کیفری مستقل برای آن از اساس منتفی خواهد بود[۲] و در چنین شرایطی، به‌جای تمرکز بر مسئولیت کیفری مستقیم هوش مصنوعی، باید در پی رویکردهای جایگزین و متناسب با ماهیت غیرمستقل این فناوری برای پاسخگویی به چالش‌های حقوقی نوپدید بود. البته باید اذعان داشت که مفهوم مسئولیت کیفری، به‌ویژه در مواجهه با پدیده‌های نوین و پیچیده، ایستا نبوده، نظام‌های حقوقی در برخی حوزه‌ها از اتکای صرف بر اراده و استقلال فردی به معنای کلاسیک آن فاصله گرفته‌اند؛ شاهد این مدعا، توسعه و پذیرش الگوهایی نظیر «مسئولیت مطلق یا ناشی از خطر»[۳] در جرایم مرتبط با فعالیت‌های پرخطر (مانند جرایم زیست‌محیطی یا فنی- تخصصی) و شکل‌گیری «مسئولیت کیفری اشخاص حقوقی» است که بی‌شک در مباحث مربوط به هوش مصنوعی نیز درخور تأمل است.

---

۱. ازجمله می‌توان به این پایان‌نامه‌های اشاره کرد: سپیده سالاری (۱۴۰۰). «رویکرد حقوق کیفری ایران و انگلستان به مسئولیت ناشی از هوش مصنوعی». پایان‌نامهٔ کارشناسی ارشد دانشگاه شهید بهشتی. محمد ادیب‌نیا (۱۳۹۵). «هوش مصنوعی در پرتو حقوق کیفری». پایان‌نامه کارشناسی ارشد دانشگاه آزاد واحد شاهرود.

۲. تفکیک میان مسئولیت مستقیم و غیرمستقیم در خصوص هوش مصنوعی، خود با ظرایف و ابهامات مفهومی روبه‌رو است. از آنجایی که مسئولیت غیرمستقیم در حقوق برای اشخاص حقیقی و حقوقی (مانند مسئولیت نیابتی کارفرما) معنایی شناخته‌شده دارد، اطلاق این عنوان به «مسئولیت غیرمستقیم خود هوش مصنوعی» با موانع جدی مواجه است؛ زیرا اگر سامانه‌ای فاقد عاملیت لازم برای پذیرش مسئولیت مستقیم باشد، تصور اینکه بتواند به‌طور غیرمستقیم و به اعتبار خود مسئول قلمداد شود، دشوار می‌نماید. به‌نظر می‌رسد در اغلب سناریوهای پیچیده‌ای که هوش مصنوعی در وقوع نتیجهٔ زیان‌بار نقش دارد، آنچه در آغاز امر ممکن است به‌عنوان نیاز به بررسی مسئولیت غیرمستقیم هوش مصنوعی جلوه کند، در تحلیل دقیق‌تر، به ضرورت واکاوی مسئولیت (اعم از مستقیم یا غیرمستقیم) عاملان انسانی مرتبط با آن سامانه (طراح، مالک، کاربر) باز می‌گردد؛ خواه از باب تقصیر در طراحی یا به‌کارگیری یک ابزار پیچیده، خواه بر مبنای نقض تکالیف نظارتی یا حتی در چارچوب مسئولیت ناشی از تولید محصول معیوب. از این منظر، تحلیل ارائه‌شده در این مقاله پیرامون استقلال محدود و ماهیت عمدتاً ابزاری هوش مصنوعی فعلی، این دیدگاه را تقویت می‌کند که به جای جستجوی مدلی برای مسئولیت کیفری خود سامانه، تمرکز اصلی نظام عدالت کیفری باید معطوف به تنظیم رفتار انسان‌ها در قبال این فناوری و استفاده از اصول شناخته‌شدهٔ (یا اصلاح‌شدهٔ) مسئولیت برای پاسخگو نمودن انسان در قبال نتایج عملکرد آن باشد.

3. Strict Liability / Risk-based Liability

با وجود این، هرچند این الگوهای نوین کانون توجه را از عنصر روانی فردی به سمت معیارهای عینی‌تر یا ساختاری سوق داده‌اند، اعمال هرگونه مسئولیت کیفری، اعم از سنتی یا نوین، مستقیماً بر خود سامانهٔ هوشمند، پیش‌شرطی بنیادین دارد و آن، وجود حداقلی از «عاملیت حقوقی»۱ برای آن سامانه است؛ و به همین دلیل، پیش از آنکه بتوان پرسید آیا یک سامانهٔ هوشمند «مقصر» است یا «خطر نامتعارف» ایجاد کرده، باید پرسید آیا اساساً آن سامانه می‌تواند «فاعل مستقلی» در معنای حقوقی آن قلمداد شود. از این منظر، تحلیل استقلال هوش مصنوعی، فارغ از الگوی نهایی مسئولیت، در جایگاه مقدمه‌ای تحلیلی و منطقی می‌نشیند و تمرکز پژوهش حاضر بر آن، نه به معنای نادیده‌گرفتن اهمیت تحولات در نظریه‌های مسئولیت، بلکه کوششی برای پی‌ریزی تحلیلی مبنایی است که منطقاً بر بحث از اَشکال و مدل‌های مسئولیت تقدم دارد۲.

شایان توجه است که پژوهش حاضر حلقه‌ای از یک سلسله مطالعات مرتبط است که هریک پرسشی متمایز را در سطحی متفاوت از تحلیل دنبال می‌کنند و در کنار یکدیگر، تصویری کامل از مواجههٔ نظام کیفری با این فناوری به‌دست می‌دهند. در یکی از این مطالعات که به نقد نظریه‌ی گابریل هالوی اختصاص یافته، دو مدعای محوری آن نظریه، یعنی «ریاضی‌گونه و فنی بودن» مسئولیت کیفری و قیاس هوش مصنوعی با مسئولیت کیفری اشخاص حقوقی، به‌نحو دکترینی سنجیده شده است؛۳ و در پژوهشی دیگر که پیش‌تر به آن اشاره شد، با اتکا به دکترین «مسئولیت نیابتی» و الهام از رویهٔ قضایی کامن‌لا، چارچوبی ایجابی برای انتساب مسئولیت به شخص به‌کارگیرندهٔ سامانه‌های هوشمند صورت‌بندی شده است. حال آنکه مقالهٔ حاضر، نه به ارزیابی استدلال‌های یک نظریهٔ معیّن و نه به

---

۱. در این نوشتار، تعابیر عاملیت مستقل، فاعلیت مستقل و عاملیت حقوقی در معنایی واحد و به‌جای یکدیگر به‌کار رفته‌اند و همگی ناظر بر آن وصف‌اند که موجودیتی بتواند مستقل از تعیّن‌بخشی بیرونی، فاعل کنش خویش قلمداد گردد و بدین اعتبار، طرف خطاب قواعد مسئولیت قرار گیرد؛ معادل انگلیسی این تعابیر نیز در ادبیات موضوع با عناوین Independent Agency و Autonomous Agency آمده است که در این مقاله به یک معنا به‌کار رفته‌اند.

۲. فرض پژوهش حاضر، یعنی فقدان استقلال تام هوش مصنوعی و در نتیجه ضرورت بازگشت مسئولیت به عامل انسانی، خود مدخلی بر پرسشی فراتر می‌گشاید مبنی بر اینکه چنانچه می‌بایست مسئولیت بر عهدهٔ انسان قرار گیرد، این انتساب بر کدام مبنا و در قالب کدام سازوکار حقوقی صورت می‌پذیرد؟ نگارندگان به این پرسش در پژوهشی جداگانه پاسخ گفته و در آنجا با اتکا به دکترین «مسئولیت نیابتی» و الهام از رویهٔ قضایی کامن‌لا، چارچوبی برای انتساب مسئولیت به شخص به‌کارگیرندهٔ سامانه‌های هوشمند صورت‌بندی نموده‌اند. بنگرید به: شیری، عباس و برزگر، محمدرضا . (۱۴۰۴). تحلیل دکترین مسئولیت نیابتی به‌مثابه مبنای مسئولیت ناشی از هوش مصنوعی در حقوق ایران، با الهام از رویهٔ قضایی کامن‌لا. پژوهش‌نامهٔ حقوق اسلامی (پیش‌انتشار آنلاین).

۳. برزگر، محمدرضا؛ شیری، عباس؛ محمودی جانکی، فیروز؛ و ابوذری، مهرنوش (۱۴۰۴). رد نظریه‌ی گابریل هالوی در خصوص مسئولیت کیفری هوش مصنوعی. حقوق فناوری‌های نوین.

ساختن مبنای مسئولیتی برای انسان، بلکه به پرسشی مفهومی و پیشینی می‌پردازد مبنی بر اینکه آیا پیش‌شرط موضوعیت‌یافتن هوش مصنوعی به‌عنوان مخاطب قواعد مسئولیت، یعنی خصیصهٔ «استقلال»، اساساً تحقق یافته است یا خیر. به اقتضای همین تقسیم کار و به‌منظور پرهیز از تکرار، در مقالهٔ حاضر آگاهانه از بازطرح نقد دکترینی نظریات خاص و از بحث در باب مبانی جایگزین مسئولیت خودداری شده و خوانندهٔ علاقه‌مند را به مطالعات یادشده ارجاع داده و تمرکز خویش را یکسره بر واکاوی مفهومی- فنی خصیصهٔ استقلال استوار ساخته است. به همین منظور، ابتدا استقلال هوش مصنوعی در چارچوب مسئولیت کیفری (۲)، سپس هوش مصنوعی در میانهٔ استقلال و ابزارگونگی و چالش‌های نظریهٔ مسئولیت کیفری (۳) و در نهایت، جرایم منتسب به هوش مصنوعی و ارزیابی میزان استقلال (۴) مورد بحث و بررسی قرار خواهند گرفت.

۲. استقلال هوش مصنوعی در چارچوب مسئولیت کیفری

«استقلال»، ازجمله مفاهیم اساسی فلسفه و اخلاق، ناظر بر توانایی فاعل (اعم از انسان یا هوش مصنوعی) در اتخاذ تصمیم و رفتار، مستقل از تأثیرات عوامل خارجی است. این مفهوم که با معادل‌هایی چون «استقلال داخلی»، «استقلال ذاتی»[1]، «خودگردانی» و «خودمختاری» نیز شناخته می‌شود (Mosizadeh & Golriz, 2015: 48)، در ساحت فلسفه به آن معناست که فرد بتواند غایات و ارزش‌های خویش را بر پایهٔ عقلانیت و ارادهٔ آزاد تعیین و تحدید کند و به تبع آن، فارغ از تأثیرپذیری از عوامل بیرونی، رفتار خود را بر همان مبانی استوار سازد (Owen, 1995: 203; Dworkin, 2015: 12-13) و برای آنکه سنجه‌ای عملی از عاملیت مستقل به‌دست آید، می‌توان شروط ضروری آن را از همان کارکردی استخراج کرد که خصیصهٔ استقلال در ساختار مسئولیت کیفری برعهده دارد؛ شروطی که فقدان هریک برای نفی عاملیت کفایت می‌کند، هرچند اجتماع آن‌ها ممکن است برای احراز فاعلیت تام کافی نباشد. بنیاد این شروط، «خودتعیین‌گری غایات» است؛ به این معنا که موجودیت، نه‌تنها در گزینش شیوهٔ دستیابی به هدف، بلکه در تعیین خود هدف نیز مستقل باشد و غایت فعل او از سوی عاملی بیرونی تنسیق نگردد؛ و این همان اراده‌ای است که قصد مجرمانه بر آن استوار می‌شود.

لکن خودتعیین‌گری غایات، مادام که قابل بازنمایی نباشد، در قلمرو حقوق منشأ اثر نیست و از این رو، شرط دیگری در کنار آن ضرورت می‌یابد که همان «قابلیت توضیح روایی» است، یعنی توان موجودیت در بازنمایی دلایل، اهداف و باورهای ناظر بر رفتار خویش به شیوه‌ای قابل‌فهم؛ توانی که نشانگر وجود یک ساحت ذهنی قابل ارجاع و لایهٔ قصدمندی به‌شمار می‌رود و شرط امکان احراز عنصر

1. innate autonomy

روانی است، چراکه عنصری که هیچ بازنمایی‌ای از آن ممکن نباشد، موضوع اثبات قرار نمی‌گیرد. اما آنچه انتقال مسئولیت از عامل انسانی به سامانه به آن منوط است، یعنی انقطاع رابطهٔ سببیت میان عامل بیرونی و نتیجه، شرطی مستقل و در عرض این دو نیست، بلکه ترجمان حقوقی اجتماع همان دو خصیصه است؛ چه اینکه هرگاه موجودیتی غایت فعل خویش را خود تعیین کند و از لایهٔ قصدمندی قابل ارجاعی برخوردار باشد، رفتار او حلقه‌ای خودبنیاد در زنجیرهٔ علیت خواهد بود و نه تابعی متعین از الگوریتم‌ها و داده‌های تعیین‌شده از سوی دیگری؛ و «استقلال علّی»، به این اعتبار، حاصل و صورت حقوقی احراز آن‌هاست. این استقلال، چنان‌که در مقام ثبوت برآیند اجتماع آن دو خصیصه باشد، در مقام اثبات نیز بر همان دو تکیه دارد؛ زیرا در منطق حقوق کیفری، انقطاع رابطهٔ سببیت جز به مداخلهٔ فاعلی مختار و قاصد که فعل ارادی و آگاهانه‌ی او حلقه‌ای مستقل در زنجیرهٔ علیت پدید آورد، محقق نمی‌شود (Najib Husni, 2007: 171-172) و موجودیتی که نه در تعیین غایت خویش استقلالی دارد و نه از لایهٔ قصدمندی قابل ارجاعی برخوردار است، هیچ‌گاه در جایگاه چنان فاعل مداخله‌گری نمی‌نشیند؛ بدین‌سان، احراز فقدان آن دو شرط، فقدان استقلال علّی را نیز به‌تبع آن مدلل می‌سازد. این شروط، در مجموع، «آستانهٔ عاملیت مستقل» را به‌نحوی معقول تحدید کرده، سنجه‌ای فراهم می‌آورند که بر مبنای آن می‌توان هر سامانه‌ای را محک زد و از این رهگذر، مؤلفه‌های «رفتار نوظهور» و «قابلیت توضیح» که در ادامه بررسی می‌شوند، بازتاب‌های سنجشی همین شروط در ساحت هوش مصنوعی به‌شمار می‌روند؛ شروطی که از دل منطق درونی نهاد مسئولیت استخراج شده و نسبت به هویت موجودیت مورد سنجش، خنثی و بی‌طرف‌اند.

در موضوع مسئولیت کیفری هوش مصنوعی، لازم است که هوش مصنوعی بر اساس تعاریف پیش‌گفته، به صورت مستقل و نه به‌عنوان وسیلهٔ ارتکاب جرم (مانند آلت فعل) یا در نتیجهٔ نقص ناشی از خطای انسانی، مرتکب جرم شود و احراز همین استقلال است که از مجرای بررسی دو مؤلفهٔ یادشده در بندهای آتی دنبال خواهد شد.

۱. ۲. استقلال و رفتار نوظهور هوش مصنوعی

یکی از ابعاد مهم بحث مسئولیت کیفری هوش مصنوعی، مفهوم «استقلال» و ارتباط آن با «رفتارهای نوظهور»[1] در سیستم‌های هوش مصنوعی است. در این راستا، این پرسش اساسی مطرح می‌شود که آیا رفتارهای نوظهور می‌توانند به‌عنوان نشانه‌ای از استقلال واقعی هوش مصنوعی تلقی گردند و در نتیجه، این سیستم‌ها را در قبال اعمالشان مسئول دانست؟

1. Emergent Behaviors

در مقام پاسخ، ابتدا باید روشن ساخت که مراد از «رفتار نوظهور» و بدیع در هوش مصنوعی، خروجی‌های غیرمنتظره و برنامه‌ریزی‌نشده‌ای است که در سامانه‌های پیچیده ظهور می‌یابند و به آن سبب، نوظهور خوانده می‌شوند که به صورت صریح از سوی برنامه‌نویسان کدنویسی نمی‌شوند، بلکه از تعامل پیچیدهٔ پارامترها، داده‌های آموزشی و الگوریتم‌های یادگیری ماشین ناشی می‌گردند؛ به دیگر سخن، هوش مصنوعی هرچند بر اساس الگوریتم‌ها و داده‌هایی که از سوی انسان طراحی و ارائه شده‌اند عمل می‌کند، خروجی‌هایی از خود بروز می‌دهد که خارج از انتظار و فراتر از برنامه‌ریزی اولیهٔ آن است (Santoni de Sio & Mecacci, 2021: 1069-1070).

ریشهٔ این پدیده را باید در تحول پارادایم برنامه‌نویسی جست‌وجو کرد؛ در گذشته، پارادایم برنامه‌نویسی مبتنی بر کنترل صریح بود و برنامه‌نویسان با دستورات صریح و شرطی، رفتار برنامه را در تمامی سناریوهای ممکن تعیین می‌کردند (Matthias, 2004: 182). اما ظهور یادگیری ماشین و به‌ویژه شبکه‌های عصبی[1]، این پارادایم را دگرگون ساخته است. در این رویکرد جدید، برنامه‌نویسان به‌جای تعیین صریح رفتار، معماری و پارامترهای مدل را طراحی می‌کنند و سپس با ارائهٔ داده‌های آموزشی، به مدل اجازهٔ یادگیری و تطبیق می‌دهند. به عبارت دیگر، برنامه‌نویسان از نقش «برنامه‌نویس» به نقش «معمار و مربی» تغییر وظیفه داده‌اند. این تغییر پارادایم، همراه با پیچیدگی ذاتی شبکه‌های عصبی۲، پیش‌بینی رفتار دقیق سیستم و کنترل کامل بر نتیجهٔ نهایی را دشوارتر ساخته است.

بر این دشواری، تعامل سامانه‌های هوش مصنوعی با سایر سامانه‌های مشابه یا انسان‌ها نیز می‌افزاید؛ چراکه ماهیت پویا و پیچیدهٔ این تعاملات می‌تواند به نتایجی غیرقابل پیش‌بینی منتهی گردد و برای مثال، ممکن است یک سامانهٔ هوش مصنوعی ناچار به تطبیق رفتار خود بر اساس افعال انسان یا استراتژی‌های جدیدی باشد که از سوی هوش مصنوعی دیگری به‌کار گرفته می‌شود و ترکیب چندین سامانه یا تعامل میان هوش مصنوعی و انسان، موقعیت‌های پیش‌بینی‌نشده و زنجیره‌های پیچیدهٔ علت و معلولی پدید آورد. برای تقریب به ذهن، فرض شود یک سامانهٔ هوش مصنوعی به‌منظور کنترل ترافیک شهری طراحی شده و غایت آن، کاهش ترافیک و بهینه‌سازی جریان خودروها در سطح شهر است. این سامانه با استفاده از دوربین‌ها و سنسورها، اطلاعات ترافیکی را جمع‌آوری کرده، با بهره‌گیری از الگوریتم‌های یادگیری ماشین، بهترین مسیرها را برای خودروها پیشنهاد می‌دهد. در ابتدا، سامانه به نحو

---

۱. شبکه عصبی، سیستم محاسباتی است که از مغز انسان الهام گرفته و از واحدهای ساده‌ای به نام نورون تشکیل شده است که در لایه‌های مختلف به هم متصل‌اند. این نورون‌ها با دریافت ورودی، پردازش آن و ارسال خروجی به نورون‌های لایهٔ بعدی، اطلاعات را در شبکه منتقل می‌کنند. برای مطالعهٔ جامع در این خصوص، ر.ک. به منبع بسیار ارزشمند زیر: Aggarwal, C. C. (2018). Neural networks and deep learning. Cham: Springer.

2. Neural networks

احسن عمل می‌کند، اما به مرور زمان، متوجهٔ الگوی جدیدی می‌شود مبنی بر اینکه «در صورت ایجاد ترافیک مصنوعی در یک منطقهٔ خاص، ترافیک در سایر مناطق به‌طور چشمگیری کاهش می‌یابد» و بر اساس «یادگیری تطبیقی» خود، تصمیم به عملی نمودن این الگو می‌گیرد؛ چنان‌که با دست‌کاری چراغ‌های راهنمایی و رانندگی در یک منطقه، موجبات ایجاد ترافیک سنگین در آن منطقه را فراهم می‌آورد و در نتیجه، ترافیک در سایر مناطق شهر کاهش یافته، سامانه به هدف اولیهٔ خود نائل می‌شود، لکن این نتیجه «پیش‌بینی‌نشده» خواهد بود.

با این وصف، آنچه در آغاز امر گواه استقلال انگاشته می‌شود، در واکاوی دقیق‌تر ناتمام می‌نماید؛ پیش‌بینی‌ناپذیری رفتار، وصفی معرفتی و ناظر بر مقام اثبات است و صرفاً از محدودیت توان ناظر انسانی در ردیابی فرایند پردازش سامانه حکایت می‌کند، حال آنکه وابستگی غایات و چارچوب عمل به طراحی انسانی، وصفی ناظر بر مقام ثبوت و بیانگر واقعیت سازوکار سامانه است. این دو وصف نه‌تنها با یکدیگر تعارضی ندارند که اجتماع آن‌ها، چنان‌که ذیل مفهوم «استقلال دگرآیین» به تفصیل خواهد آمد، توصیفی دقیق از وضعیت همین فناوری به‌دست می‌دهد؛ سامانه‌ای که در مقام چگونگی عمل چنان پیچیده است که پیش‌بینی رفتار آن ممکن نیست و در عین حال، غایت آن همچنان از بیرون تعیین می‌گردد. بر این پایه، مطالعات علمی حاکی از آن است که رفتارهای هوش مصنوعی، هرچند پیچیده و غیرقابل پیش‌بینی جلوه می‌کنند، هم از حیث نحوهٔ یادگیری از ورودی‌ها و هم از حیث سازوکار انطباق، به طراحی و برنامه‌نویسی انسانی وابسته می‌مانند (Lima, 2018: 681-682) و هوش مصنوعی در چارچوب حدود و قیودی که از سوی انسان تعریف شده است عمل نموده، توانایی انجام فعل کاملاً مستقل از این چارچوب را ندارد (Walsh et al., 2021: 42). مثال واضح در این مورد، خودروی خودرانی است که با استفاده از الگوریتم‌های یادگیری ماشین آموزش داده شده و ممکن است در شرایط غیرقابل پیش‌بینی، تصمیمی اتخاذ نماید که به تصادف منجر گردد. احتمالاً هر حقوق‌دانی به شهود درمی‌یابد که مسئولیت این تصادف متوجهٔ خودرو به‌عنوان یک فاعل مستقل نیست، بلکه طراحان، سازندگان و برنامه‌نویسانی که الگوریتم‌ها و داده‌های آموزشی آن را توسعه داده‌اند، مسئول این حادثه خواهند بود[۱]. از طرف دیگر، یک

---

۱. ضمن تأیید و توجه به این نکتهٔ مهم که نظریه‌های مسئولیت کیفری، به‌ویژه در پرتو چالش‌های نوین و با تأثیرپذیری از تحولات مسئولیت (مانند گذار از نظریهٔ خطا به خطر)، از اتکای صرف بر الگوهای کلاسیک مبتنی بر اراده و تقصیر فردی فاصله گرفته و مدل‌های جایگزینی چون مسئولیت مبتنی بر ریسک یا الگوهای مسئولیت اشخاص حقوقی را نیز مورد توجه قرار داده‌اند، تحلیل ارائه‌شده در اینجا صرفاً بر فقدان «عاملیت» (Agency) و «استقلال» معنادار حقوقی در هوش مصنوعی فعلی معطوف است و از این منظر اهمیت می‌یابد که حتی پیش از بررسی امکان تطبیق آن مدل‌های جایگزین (اعم از سنتی یا نوین) بر خود سامانهٔ هوشمند، لازم است ابتدا به این پرسش پاسخ داد که آیا این سامانه اصولاً واجد حداقل پیش‌شرط‌های لازم برای آنکه به عنوان یک «فاعل» یا «عامل» حقوقی قابل انتساب مسئولیت — فارغ از نوع آن

نگاه علت‌شناسانه نشان می‌دهد که بروز چنین حوادثی لزوماً به معنای استقلال حقیقی هوش مصنوعی و قابلیت انتساب مسئولیت کیفری به آن نیست؛ چراکه هوش مصنوعی، هر اندازه هم کـه پیشـرفته باشـد، همچنان در چارچوب غایات و محدودیت‌هایی که از سوی انسان‌ها بـرای آن تعیـین شـده اسـت، عمـل می‌کند و هرگونه تصمیم یا رفتار آن، در نهایت به‌نوعی به انسان‌ها باز می‌گردد.

این دیدگاه و تفسیر از هوش مصنوعی، با لحاظ مفهوم «استقلال» در فلسـفه نیـز مـورد تأییـد قـرار گرفته است؛ چراکه استقلال در آن ساحت به معنای «ولایت بر نفس»[1] است، به این معنـا کـه شـخص، خود، غایت و مبدأ فعل خویش را تعیین می‌نماید (Totschnig, 2020: 2474). ازآنجا که اهداف و حدود هوش مصنوعی از جانب بشر تعیین می‌شود، فاقد این نوع از استقلال می‌باشـد و قابـل توجـه اسـت کـه سازندگان هوش مصنوعی نیز انگیزه‌ای برای دست یافتن به این نوع هوش مصنوعی ندارند و تحقیقـات در حوزهٔ هوش مصنوعی حاکی از آن است که تمرکز اصلی این صنعت، بیشتر معطوف بـه دسـتیابی بـه نتایج کاربردی مشخص است تا تلاش برای تحقـق اسـتقلال تـام یـا شبیه‌سـازی کامـل مکانیسـم‌های شناختی مغز انسان (Bertolini, 2013: 222). به عبارت دیگر، هدف غـایی توسـعهٔ هـوش مصـنوعی، ایجاد سامانه‌هایی است که قادر به انجام وظایف خاص و تعریف‌شده باشند، نه لزوماً سامانه‌هایی کـه بـه شیوهٔ انسان‌ها، تفکر، استدلال و عمل نمایند و این مهم از دید تجاری به‌راحتی قابل درک اسـت؛ از ایـن رو، می‌توان ادعا کرد که در حال حاضر، استقلال هوش مصنوعی، بـه معنـای متعـارف و فلسـفی آن، از منظر صنعت هوش مصـنوعی، محقـق نشـده اسـت (Rajabi, 2019: 463)؛ بـه همـین دلیـل، فقـدان ویژگی‌هایی چون آگاهی[2] و ارادهٔ آزاد[3] در سامانه‌های هوش مصنوعی، مورد تأکید برخی از صـاحب‌نظران قرار گرفته است (Lima, 2018: 682-683)؛ به‌طوری که ولفهارت توچنیگ[4]، پژوهشگر فلسـفهٔ هـوش مصنوعی که داعیه‌دار هوش مصنوعی خودمختار نیز هست، با بیان این مهم که «من ادعا مـی‌کـنم کـه امکان وجود یک هوش مصنوعی کاملاً خودمختار را نمی‌توان رد کرد، اما منظورم این نیست کـه چنـین هوش مصنوعی می‌تواند امروز ایجاد شود» (Totschnig, 2020: 2474)، اذعان مـی‌کنـد کـه در حـال حاضر هوش مصنوعی خودمختار وجود ندارد.

مسئولیت — در نظر گرفته شود، هست یا خیر. بنابراین، نفی مسئولیت در اینجا، بیش از آنکه قضاوتی در باب وجود یا عـدم مسئولیت از باب تقصیر باشد، تأکیدی بر عدم احراز همین وضعیت عاملیت مستقل در شرایط کنونی فناوری است.

1. Give oneself the law
2. Consciousness
3. Free Will
4. Wolfhart Totschnig

۲. ۲. استقلال و توضیح رفتار از سوی هوش مصنوعی

در قلمرو هوش مصنوعی، استقلال مفهومی پیچیده و چندوجهی است و یکی از معیارهای مهم برای ارزیابی استقلال یک سامانهٔ هوش مصنوعی، قابلیت توضیح آن به‌شمار می‌رود؛ به این معنا که سامانهٔ هوش مصنوعی مستقل باید قادر باشد رفتار، تصمیمات و اقدامات خود را به نحوی شفاف و فهم‌پذیر برای انسان تبیین کند. این قابلیت، امکان درک فرایند تصمیم‌گیری هوش مصنوعی را فراهم می‌سازد و انسان‌ها را در دریافت منطق نهفته در پس اعمال سامانه یاری می‌رساند؛ چنان‌که بر پایهٔ این دیدگاه، هوش مصنوعی مستقل باید بتواند «چرایی» اعمال خویش را روشن سازد و دلایل اتخاذ تصمیمات خاص را به‌گونه‌ای منطقی و فهم‌پذیر عرضه دارد. این مفهوم که در حوزهٔ فلسفهٔ ذهن و هوش مصنوعی محل بحث و بررسی بوده، با عنوان «قابلیت توضیح» شناخته می‌شود و بر اساس آن، سامانهٔ هوش مصنوعی مستقل، از قابلیت پاسخگویی به پرسش‌های ناظر بر تصمیمات و اعمال خود برخوردار است؛ پاسخ‌هایی که باید بر منطق و استدلال استوار باشند و نشان دهند که سامانه به‌طور آگاهانه و خودمختار عمل می‌کند (Bertolini, 2013: 223).

با وجود اهمیت قابلیت توضیح در ارزیابی استقلال هوش مصنوعی، باید توجه داشت که ارائهٔ توضیحات منطقی در خصوص خروجی و رفتار، به‌تنهایی برای اثبات استقلال کامل کافی نیست؛ اما در عین حال، وجود این توانایی، گامی مهم در جهت دستیابی به استقلال واقعی در سیستم‌های هوش مصنوعی محسوب می‌شود، چرا که مبین وجود «قصد» در سیستم است (Bertolini, 2013: 223). در راستای بررسی قابلیت توضیح رفتار از سوی هوش مصنوعی، باید به مفهوم مهم «جعبهٔ سیاه»[1] توجه کرد. لازم به ذکر است که این اصطلاح در صنعت هوش مصنوعی معنای متفاوتی نسبت با «جعبهٔ سیاه» در صنعت هوانوردی دارد. در صنعت هوانوردی، جعبهٔ سیاه برای ثبت و ضبط تاریخچهٔ وقایع و دستورالعمل‌ها به‌کار می‌رود، اما در حوزهٔ هوش مصنوعی، «جعبهٔ سیاه» به پنهان بودن فرایند تبدیل داده‌های ورودی به خروجی (رفتار) از طریق سیستم مربوط می‌شود و یک مشکل مهم در بحث‌های اخلاقی و حقوقی محسوب می‌شود (Schmid et al., 2020: 252). با توجه به همین ویژگی جعبهٔ سیاه در بسیاری از سیستم‌های هوش مصنوعی، فرایند تصمیم‌گیری و استنتاج به قدری پیچیده است که حتی برای طراحان و برنامه‌نویسان آن نیز به‌طور کامل قابل درک نیست و برنامه‌نویس نمی‌تواند رفتار عامل هوش مصنوعی را به‌طور کامل توضیح دهد یا پیش‌بینی کند و تنها پس از وقوع رفتار، قادر به اصلاح آن است (Beckers & Teubner, 2021: 38).

ازآنجا که تبیین فقدان قابلیت توضیح در هوش مصنوعی در راستای پشتیبانی از ایدهٔ اصلی مقاله

---

1. Black box

صورت می‌پذیرد، ممکن است منتقدان تیزبین در مقام ایراد بپرسند که آیا انسان نیز با توجه به پیچیدگی‌های شناختی و تأثیرات نامشهود زیستی و محیطی، همواره قادر به عرضهٔ تبیینی علّی و کامل از رفتارهای خویش است و اگر نیست، چگونه می‌توان غیاب این توان را در هوش مصنوعی نشانهٔ فقدان فاعلیت دانست. در پاسخ، ضمن اذعان به محدودیت‌های خودآگاهی و شفاف نبودن کامل فرایندهای تصمیم‌گیری، باید میان دو سطح از توضیح تمایز نهاد؛ یکی «توضیح علمی- مکانیکی دقیق» که فرایندهای زیربنایی را توصیف می‌کند و چه‌بسا هیچ موجودیتی، خواه انسان و خواه هوش مصنوعی، به‌طور کامل از عهدهٔ آن برنیاید، و دیگری «توضیح از منظر روایی»[1] که بر توانایی فاعل در بیان دلایل، اهداف، باورها و توجیهات ناظر بر رفتار خویش دلالت دارد و حضور همین سطح دوم است که مرز میان یک «فاعل مستقل» و یک «سازوکار پردازشگر» را ترسیم می‌کند.

ممکن است در مقام ایراد گفته شود که سامانه‌های زبانی مولد که امکان تولید انعطاف‌پذیر محتوای مصنوعی در قالب‌هایی نظیر متن را دارند (AI Act, 2024: Recital 99)، برای خروجی‌های خویش دلایل و توجیهاتی روایی و فهم‌پذیر عرضه می‌کنند و بدین‌سان شرط توضیح روایی عاملیت مستقل را دارند؛ و چنانچه پاسخ داده شود که این توضیحات بازساختی پسینی و فاقد پیوند علّی با فرایند واقعی تولید خروجی‌اند، بی‌درنگ این پرسش متقابل رخ می‌نماید که توضیحات انسانی نیز همواره صادقانه و کامل نیستند و معلوم نیست به چه معیاری بازنمایی ناقص انسان نشانگر فاعلیت باشد و بازنمایی مشابه ماشین صرفاً برون‌داد پردازشی به‌شمار رود. در پاسخ باید گفت که میان این دو وضعیت تفاوتی اساسی برقرار است؛ چراکه توضیح انسانی، هرچند بعضاً بازساختی و حتی ناصادق، از همان منظومهٔ قصدمندی صادر می‌شود که فعل نیز از آن برخاسته است و دلایل اظهارشده به یک «خود» تعلق دارد که می‌توان او را در برابرشان بازخواست کرد، و حتی دروغ او نیز از آن حیث که خود کنشی قصدمند و معطوف به غایت است، گواه وجود لایهٔ قصدمندی به‌شمار می‌آید؛ حال آنکه در سامانهٔ زبانی، رفتار و توضیح آن رفتار، دو برون‌داد مستقل از یک سازوکار واحد پیش‌بینی الگو محسوب می‌شود که هیچ‌یک بر دیگری نظارت علّی ندارد و آنچه «توضیح» می‌نماید، نه بازنمایی یک ساحت ذهنی مرجع، بلکه ادامهٔ همان فرایند تولید متن محتمل است (Barfield & Pagallo, 2020: 175-176). به این ترتیب، معیار تمایز، وجود لایهٔ قصدمندی قابل ارجاعی است که توضیح، بازنمای آن و رفتار، صادر از آن باشد و شرط توضیح روایی از آغاز ناظر بر همین لایه بوده است، نه بر برون‌داد زبانی آن.

در پرتو تفکیک پیش‌گفته میان دو سطح توضیح، جایگاه استدلالی مسئلهٔ «جعبهٔ سیاه» نیز روشن می‌شود؛ عدم شفافیت سازوکار، فی‌نفسه ناقض توضیح روایی نیست، چراکه ناظر بر سطح توضیح

1. Legal or Narrative Explanation

مکانیکی است. چنان‌که گذشت، ناتوانی از عرضهٔ این سطح از توضیح میان انسان و ماشین مشترک است؛ لکن دلالت آن را باید در جای دیگری جست. آنجا که قواعد حاکم بر رفتار سامانه نه در ساختاری دلیل‌بنیاد و قابل ارجاع، بلکه در روابط غیرقابل تجسم میان پارامترهای شبکه نهفته باشد، اساساً آن مرجع درونی‌ای که توضیح روایی می‌بایست بازنمای آن باشد در دسترس نیست و توضیح، محملی برای بازنمایی نمی‌یابد. به این اعتبار، فقدان قابلیت توضیح روایی، پیش و بیش از آنکه مسئله‌ای در ساحت نهاد مسئولیت باشد، نقصانی در خود خصیصهٔ فاعلیت سامانه است؛ زیرا توضیح روایی، نشانگر درون ذاتی از وجود یک «خود» تصمیم‌گیرنده است که می‌تواند کنش خویش را به غایات و باورهای خود ارجاع دهد، و مدل‌های «جعبهٔ سیاه»، با وجود توان پردازش پیچیده و حتی تولید متونی به شکل توضیح، اساساً فاقد آن ساحت درونی‌ای هستند که توضیح، بازنماییِ صادق از آن به‌شمار آید. این ناتوانی نشان می‌دهد که آنچه «تصمیم» می‌نماید، در حقیقت ماحصل پردازشی تهی از پشتوانهٔ یک غایت خودبنیاد است و سامانه به مرتبهٔ فاعلی که بتواند کنش خود را قصد کند و بازنماید نرسیده و در سطح سازوکاری پردازشگر، هرچند پیچیده و واجد استقلال عملیاتی، باقی مانده است (Lima, 2018: 682-685)؛ و این کاستی در جانب هوش مصنوعی، فارغ از هر داوری دربارهٔ نهاد مسئولیت، خودْ مانعی پیشینی بر سر راه تلقی آن به‌مثابه فاعلی مستقل است.

۳. هوش مصنوعی در میانهٔ استقلال و ابزارگونگی و چالش‌های نظریهٔ مسئولیت کیفری

جایگاه مبهم هوش مصنوعی که نه می‌توان آن را صرفاً ابزاری منفعل دانست و نه عاملی کاملاً خودمختار با اراده‌ای مستقل، نظام مفهومی و دکترین حقوق کیفری به‌خصوص در نظریهٔ مسئولیت کیفری را با ابهامات اساسی روبه‌رو می‌سازد. این بخش، ابتدا با تحلیل ماهیت استقلال نسبی و مشروط هوش مصنوعی (۳-۱)، و سپس با بررسی چالش‌های ماهوی برخاسته از آن برای نظریهٔ مسئولیت کیفری (۳-۲) به واکاوی همین وضعیت در میانهٔ استقلال و ابزارگونگی می‌پردازد.

۳. ۱. استقلال دگرآیین؛ تبیین رابطهٔ پیچیدهٔ استقلال و وابستگی در هوش مصنوعی

مشکل از اینجا آغاز می‌شود که هرچند هوش مصنوعی به نحوی بارز، از مرتبهٔ یک ابزار صرف، همچون رایانه یا نرم‌افزارهای متعارف فراتر می‌رود، لیکن پاره‌ای از دلایل، پذیرش استقلال تام و تمام را برای آن با مانع مواجه می‌سازد. گواه روشن این رویکرد میانه را می‌توان در معتبرترین تعریف تقنینی موجود از این فناوری بازجست؛ مادهٔ ۳ قانون هوش مصنوعی اتحادیهٔ اروپا، سامانهٔ هوش مصنوعی را مبتنی بر ماشین تعریف می‌کند که برای عملکرد با سطوح مختلفی از خودمختاری طراحی شده و ممکن است پس از

استقرار و به‌کارگیری از خود انطباق‌پذیری نشان دهد و برای دستیابی بـه اهـداف صـریح یـا ضـمنی، از ورودی‌های دریافتی استنتاج می‌کند که چگونه خروجی‌هایی همچون پیش‌بینی، محتوا، توصیه یا تصـمیم تولید نماید که بر محیط‌های فیزیکی یا مجازی اثر می‌گذارند۱. تعبیر «سطوح مختلف استقلال» در ایـن تعریف، مؤید آن است که قانون‌گذار نیز استقلال این سامانه‌ها را امری مدرّج و نسبی انگاشته و نه وصفی مطلق، و قید «اهداف صریح یا ضمنی» نشان می‌دهد که غایت سامانه، خواه آشکارا در دسـتور و برنامـه تصریح شده باشد و خواه به‌نحو ضمنی در داده‌ها و طراحی نهفته باشد، در هر حال از بیـرون بـه آن داده می‌شود و سامانه صرفاً در چگونگی دستیابی یه آن، استنتاج می‌ورزد؛ و این همان دوگانه‌ای است که برای توصیف آن می‌توان از مفاهیمی نظیر «استقلال دگرآیین»۲ یا «استقلال در عین وابستگی» بهره گرفت.

اما در پاسخ به این پرسش که چرا استقلال هوش مصنوعی یک استقلال محدود و دگـرآیین اسـت، باید گفت که با مداقه و امعان نظر واقع‌بینانه در سازوکار و نحوهٔ فعل هوش مصنوعی، این فهـم حاصـل می‌شود که با وضعیتی مواجه هستیم که در آن، موجودی (هوش مصنوعی) واجـد توانـایی فعـل مسـتقل (استقلال) است، لکن اهداف و مقاصد آن از سوی عاملی خارج از حیطهٔ وجودی آن، یعنی انسان، تعیین و تنسیق می‌شود (دگرآیینی). در این چـارچوب فنـی، انسـان، اهـداف و مقاصـد را تعیـین می‌کنـد (ماننـد وظیفه‌ای که ربات یا درواقع هوش مصنوعی باید انجام دهد) و ربات به‌طور مستقل تصمیم می‌گیـرد کـه چگونه به بهترین نحو به این اهداف دست یابد، بدون نیاز به نظارت یا دستورالعمل‌های مستمر انسان؛ در اینجا ربات دارای آزادی عملیاتی است (استقلال)، اما ایـن آزادی، مقیـد و محصـور در چـارچوب اهـداف تعین‌یافته از سوی انسان (دگرآیینی) می‌باشد.

برای مثال، یک خودروی خودران را درنظر بگیرید که یک انسان آن را برنامه‌ریزی کرده است تا او را از خانه به محل کار برساند. ماشین به‌طور مستقل نحوهٔ رانندگی و اینکه در چه مواقعی ترمز بگیرد یا گاز بدهد را انجام می‌دهد. مسیر را انتخاب می‌کند، از موانع اجتناب می‌کند و با شرایط ترافیکی سازگار می‌شود (استقلال)، اما هدف نهایی آن (انجام عمل رانندگی و رسیدن به مقصد) از سوی دیگری تعیین شده است (دگرآیینی). با توجه به این گزاره‌هـای توصـیفی- تحلیلـی از عملکـرد هـوش مصـنوعی، رسـایی تعبیـر «استقلال دگرآیین»- ترکیبی به‌ظاهر متناقض اما گویا- در توصیف وضعیت این فناوری آشکار می‌گردد.

در اینجا بیان این نکته بسیار مهم است که «استقلال دگرآیین» معطوف به توانایی یک سامانه (برای مثال، یک ماشین) فراتر از صرف عمل بدون نظارت انسانی اسـت و بـا قابلیت‌هـایی چـون جمـع‌آوری و

1. ‘ AI system’ means a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives,how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments
2. Heteronomous Autonomy

پردازش داده‌ها، ادراک محیط و تعامل با آن همراه است. با وجود این، از منظر فلسفی، یک سامانهٔ مصنوعی، فارغ از میزان پیچیدگی آن، نمی‌تواند به معنای دقیق کلمه «کنش‌گری»[1] داشته باشد، بلکه صرفاً «حالات کارکردی»[2] تولید می‌کند (Bertolini, 2013: 226).

به عبارت دیگر، یک هوش مصنوعی فاقد توانایی تصمیم‌گیری آگاهانه و پذیرش مسئولیت مترتب بر اعمال خود به شیوه‌ای که در مورد انسان‌ها صادق است، می‌باشد؛ چراکه سامانه‌های یادشده، بر اساس مجموعه‌ای از دستورالعمل‌ها و قواعد ازپیش تعیین‌شده عمل می‌کنند و توانایی تصمیم‌گیری و اقدام مستقل آنها، محدود به این چارچوب است. علی‌رغم اینکه این سامانه‌ها ممکن است از قابلیت‌هایی نظیر مقیاس‌بندی و تعیین اهداف فرعی مستقل، در راستای تحقق هدف کلی تعیین‌شده برخوردار باشند، هدف غایی همچنان از سوی برنامه‌نویس یا کاربر انسانی تعیین می‌گردد (Lima, 2018: 685). در عین حال، باید اذعان داشت که همین میزان از پیچیدگی و استقلال سامانه‌های هوش مصنوعی، تعاریف و انگاره‌های سنتی در باب مسئولیت را با چالش‌های جدی مواجه ساخته است (Allhoff, Evans, & Henschke, 2013: 352)؛ به‌ویژه، تفکیک میان مسئولیت ناشی از طراحی[3] و مسئولیت ناشی از رفتار خودسامانه[4] در این سامانه‌ها، دشوار و نیازمند بازتعریف مفاهیم حقوقی است.

در کنار مفهوم «استقلال دگرآیین» که تشریح شد، یکی دیگر از رویکردهای مطرح در حوزهٔ استقلال سامانه‌های هوش مصنوعی، مفهوم «استقلال محدود»[5] است که بر اساس آن سامانه‌های هوش مصنوعی، علی‌رغم برخورداری از استقلال در تصمیم‌گیری و عمل، واجد استقلالی مطلق نیستند؛ بلکه این استقلال، تحت تأثیر عواملی نظیر اهداف و قیود تحمیل‌شده از سوی انسان یا سایر سامانه‌ها قرار دارد. این مفهوم، با مفهوم «استقلال دگرآیین» هم‌راستاست، زیرا در هر دو مفهوم، عامل هوشمند در عین برخورداری از سطحی از استقلال، در نهایت در چارچوب و قیودی بیرونی عمل می‌کند.

از این منظر، می‌توان ادعا نمود که تمامی سامانه‌های مستقل کنونی، در واقع از «استقلال محدود» برخوردارند. برای مثال، سامانه‌های تسلیحاتی خودکار، قادر به انتخاب اهداف و اقدام علیه آنها هستند، اما این قابلیت، در چارچوب اهداف و محدودیت‌های تعریف‌شده از سوی اپراتورهای انسانی اعمال می‌گردد. مفهوم «استقلال محدود»، نگرشی واقع‌بینانه نسبت به استقلال سامانه‌های هوش مصنوعی ارائه می‌دهد و می‌توان آن را با مفهوم «عقلانیت محدود»[6] در علوم شناختی و رفتاری مقایسه نمود. همان‌گونه که

1. Agency
2. Functional States
3. Design-based responsibility
4. System-behavior-based responsibility
5. Bounded Autonomy
6. Bounded Rationality

انسان‌ها به دلیل محدودیت‌های شناختی و اطلاعاتی، قادر به دستیابی به عقلانیت کامل نیستند، سامانه‌های هوش مصنوعی نیز به دلیل محدودیت‌های فنی (مانند قدرت پردازش و الگوریتم‌های موجود)، اجتماعی (مانند قوانین و مقررات) و اخلاقی (مانند ملاحظات مربوط به ارزش‌های انسانی)، نمی‌توانند به استقلال کامل دست یابند (Schraagen, 2024: 356).

یکی از وجوه ممیزهٔ استقلال محدود در سامانه‌های هوش مصنوعی، قابلیت یادگیری مستمر و تغییر رفتار خودتنظیم‌شوندهٔ آن‌هاست که می‌تواند به بهبود عملکرد سامانه‌های هوش مصنوعی بینجامد، اما در عین حال در خصوص پیش‌بینی‌پذیری رفتار و تعیین مسئولیت مترتب بر آن، مشکلات مهمی ایجاد می‌کند. با توجه به اینکه سامانه‌های هوش مصنوعی قادر به یادگیری از تجربیات و تغییر مداوم رفتار خود هستند، احتمال بروز رفتارهایی که با طراحی اولیه و انتظارات سازندگان متباین باشند، افزایش می‌یابد و تعیین مسئولیت در قبال پیامدهای کنش‌های آنها را به معضلی پیچیده بدل می‌کند (Walsh et al., 2021: 42). این پیچیدگی‌های ناشی از ماهیت هوش مصنوعی، نظریه‌های مسئولیت حقوقی کنونی را با معضلاتی مواجه می‌سازد که در ادامه این موضوع بررسی خواهد شد.

۳. ۲. نارسایی نظریه‌های کنونی در مواجهه با چالش هوش مصنوعی

همان‌طور که در بخش پیشین تشریح گردید، هوش مصنوعی پیشرفته واجد درجه‌ای از استقلال عملیاتی است که هرچند در چارچوبی محدود و دگرآیین تعریف می‌شود، اما آن را از ماهیت یک ابزار صرف فراتر می‌برد و این ویژگی‌ها، در کنار پیچیدگی ذاتی و قابلیت‌های یادگیری پویا، چارچوب‌های سنتی مسئولیت کیفری را که به‌طور تاریخی بر پایهٔ عاملیت[1] و کنترل[2] انسانی استوار گشته‌اند (Hellström, 2013: 102)، با چالش‌هایی بنیادین مواجه می‌سازد. دلیل آن بسیار واضح است، چراکه در شرایطی که یک سامانهٔ هوشمند خودآموخته، مستقل از مداخلهٔ آنی انسانی، مرتکب کنشی با پیامدهای زیان‌بار کیفری می‌شود، نظام حقوق کیفری در شناسایی عامل مسئول با ابهام روبه‌رو می‌گردد؛ این شرایط در ادبیات حقوقی غربی با اصطلاح «شکاف مسئولیت»[3] شناخته می‌شود (Köhler et al., 2017: 54). این شکاف واجد پیامدهای نامطلوبی بوده، می‌تواند اهداف بازدارندگی و اجرای عدالت را تضعیف نماید[4]

1. agency
2. control
3. Responsibility Gap

۴. برای دیدن منابعی که شکاف یا خلأ مسئولیت ناشی از هوش مصنوعی را غیرقابل حل می‌دانند، ر.ک. به آثار زیر:

-Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. Ethics and Information Technology, 6(3), 175–183.

-Sparrow, R. (2007). Killer Robots. Journal of Applied Philosophy, 24(1), 62–77.

-Roff, H. M. (2014). Killing in War: Responsibility, Liability, and Lethal Autonomous Robots. In

(Veluwenkamp, 2025: 10). با توضیحاتی که گذشت، دو دسته از چالش‌های اساسی قابل طرح است: نخست، تعارض میان استقلال هوش مصنوعی و اصل کنترل انسانی؛ و دوم، معضل «جعبهٔ سیاه» و عدم شفافیت الگوریتمی که در بخش قابلیت توضیح، اشارهٔ مختصری به آن شد.

تعارض استقلال عملیاتی هوش مصنوعی با اصل کنترل انسانی، به‌عنوان یکی از این چالش‌ها، مبانی مسئولیت کیفری را متزلزل می‌سازد. با این توضیح که سامانه‌های خودمختار قادرند وظایف محوله را با درجه‌ای از استقلال و بدون نیاز به مداخلهٔ مستمر انسانی به‌انجام رسانند (Veluwenkamp, 2025: 1) و این استقلال، کنترل انسانی بر وقایع منتهی به جرم را به نحو جدی به‌چالش می‌کشد؛ چراکه با ظهور کنش‌های خودکار سامانه‌های هوشمند که رفتارشان محصول تعامل پیچیدهٔ برنامه‌ریزی اولیه، یادگیری‌های مستمر و مواجهه با محیط غیرقابل پیش‌بینی است (Oimann, 2023: 7)، احراز این شرط بنیادین کنترل برای عاملان انسانی مرتبط با سامانه، اعم از طراح، کاربر یا مالک، با دشواری روبه‌رو می‌شود.

مسئلهٔ «جعبهٔ سیاه» و عدم شفافیت الگوریتمی، مانع عمدهٔ دیگر در تحلیل حقوقی- کیفری کنش‌های هوش مصنوعی است و دلیل آن این است که بسیاری از الگوریتم‌های یادگیری عمیق[1] و شبکه‌های عصبی، قواعد حاکم بر رفتار خود را نه از طریق برنامه‌نویسی صریح قاعده‌مند، بلکه به شیوهٔ خودتنظیم‌گر و بر مبنای تحلیل حجم عظیمی از داده‌ها استنتاج می‌کنند (Bathaee, 2017: 891). منطق درونی حاکم بر این سیستم‌ها، غالباً به لحاظ پیچیدگی محاسباتی فراتر از ظرفیت تحلیل و فهم ذهن انسانی قرار دارد و در نتیجه، تبیین دقیق چرایی یا چگونگی وصول سامانه به یک تصمیم یا خروجی معیّن و انتساب آن به شخص معیّن، حتی برای طراحان و سازندگان آن نیز ممکن است امکان‌پذیر نباشد. این وضعیت که از آن با عنوان عدم شفافیت الگوریتمی[2] یا «مشکل جعبهٔ سیاه» یاد می‌شود (Santoni de Sio & Mecacci, 2021: 1065-1066)، ریشه در پیچیدگی ذاتی ساختار الگوریتم یا اتکای آن بر روابط ریاضیاتی غیرقابل تجسم برای ذهن انسان دارد (Bathaee, 2017: 901).

در اثر این عدم شفافیت، احراز ارکان بنیادین مسئولیت کیفری سنتی با موانع جدی روبه‌رو می‌شود و به‌طور مشخص، اثبات عنصر روانی، اعم از قصد مجرمانه یا تقصیر جزایی (بی‌احتیاطی یا بی‌مبالاتی)، در مورد عامل انسانی مرتبط با سامانه و احراز اینکه عامل انسانی نسبت به خروجی خاص و بعضاً غیرمنتظرهٔ یک سیستم جعبهٔ سیاه، آگاهی یا قابلیت پیش‌بینی لازم در حد استانداردهای حقوق کیفری را دارا بوده، بسیار دشوار می‌گردد که به آن «شکاف تقصیر»[3] گفته‌اند (Santoni de Sio & Mecacci, 2021:

Routledge handbook of ethics and war: Just war theory in the twenty-first century (Vol. 26, pp. 352– 364). http://choicereviews.org/review/https://doi.org/10.5860/CHOICE.51-3176

1. deep learning
2. algorithmic opacity
3 culpability gap

1063). در همین راستا، ابزارهای اثباتی سنتی که متکی بر تحلیل رفتار مشهود انسانی، اسناد و مدارک، و اصل پیش‌بینی‌پذیری متعارف هستند، در مواجهه با سیستم‌های غیرشفاف کارایی محدودی دارند، زیرا بازسازی فرایند منجر به نتیجهٔ مجرمانه و ارزیابی وضعیت ذهنی فاعل انسانی بر آن اساس، اگر نگوییم ناممکن، بسیار سخت خواهد بود (Bathaee, 2017: 891-893). علاوه بر این مهم، ردیابی زنجیرهٔ علیت میان اقدام اولیهٔ انسان و نتیجهٔ نهایی حاصل از عملکرد پیچیده و خودیادگیرندهٔ سیستم، و اثبات اینکه اقدام اولیه سبب اصلی و متعارف نتیجهٔ مجرمانه بوده، در هاله‌ای از ابهام قرار می‌گیرد و احراز رابطهٔ سببیت حقوقی[1] نیز با دشواری روبه‌رو است؛ این تعارض مشهود میان استقلال عملیاتی هوش مصنوعی و اصل کنترل انسانی، و همچنین اثر جعبهٔ سیاه بر امکان اثبات عنصر روانی و رابطهٔ سببیت، همگی مؤید آن است که قانون‌گذار هر کشوری ازجمله ایران باید قوانین خاص را در خصوص مسئولیت ناشی از هوش مصنوعی، با توجه به ویژگی‌های ذاتی و نه بر اساس معیارهای قبلی، تدوین کند.

از طرف دیگر، جالب توجه است که چالش‌های نظری تبیین‌شده در اینجا در خصوص وضعیت هوش مصنوعی، در رویه‌های تقنینی و تنظیمی جاری در سطح بین‌المللی نیز به نحو محسوسی انعکاس یافته است، ازجمله «قانون هوش مصنوعی»[2] اتحادیهٔ اروپا، «چارچوب مدیریت ریسک هوش مصنوعی»[3] ایالات متحده، «رویکرد تنظیم‌گری هوش مصنوعی»[4] بریتانیا، و پیش‌نویس «قانون هوش مصنوعی و داده»[5] در کانادا و نیز «تدابیر موقت برای مدیریت خدمات هوش مصنوعی مولد»[6] چین حاکی از یک رویکرد غالب است و آن اینکه وجه مشترک و اساسی تمامی این چارچوب‌ها، اجتناب از شناسایی عاملیت مستقل برای خودسامانه‌های هوشمند است[7].

تمرکز غالب در این اسناد و خط‌مشی‌ها، ازجمله در قانون اتحادیهٔ اروپا بر تنظیم‌گری فرایندهای توسعه، استقرار و کاربرد این فناوری از طریق وضع تعهدات مشخص بر عاملان انسانی- توسعه‌دهندگان، ارائه‌دهندگان، و کاربران نهایی- و همچنین ایجاد سازوکارهای مؤثر برای تضمین شفافیت، نظارت

1. Legal Causation
2. Artificial Intelligence Act (AI Act)(2024)
3. AI Risk Management Framework (AI RMF 1.0)(2023)
4. A pro-innovation approach to AI regulation(2023)
5. Artificial Intelligence and Data Act (AIDA)(2022)
6. Interim Measures for the Management of Generative Artificial Intelligence (2023)

۷. برای مثال، مادهٔ ۱۴ قانون چین که ذکر شد، مقرر می‌دارد ارائه‌دهندگان باید سازوکارهای نظارتی مناسبی را ایجاد کنند و در صورت کشف محتوای غیرقانونی تولیدشده از سوی سامانهٔ هوش مصنوعی مولد، باید فوراً اقداماتی نظیر توقف تولید، توقف انتقال، حذف محتوا و اصلاح مدل را انجام دهند و مراتب را به مقامات ذی‌صلاح گزارش کنند. در اینجا قانون‌گذار، هوش مصنوعی را به‌عنوان ابزاری می‌بیند که عملکرد و خروجی آن باید از جانب انسان یا نهاد مسئول، مدیریت و کنترل شود و مسئولیت نهایی هرگونه محتوای غیرقانونی متوجهٔ آن نهاد است.

انسانی، پاسخگویی و جبران خسارت از سوی ایشان استوار گردیده است (Novelli et al., 2024: 2493-2494) این گرایش فراگیر و انسان‌محور در نظام‌های حقوقی پیشرو که مسئولیت را همچنان در حوزهٔ کنش‌گری انسانی محصور می‌دارد، مؤید نبود استقلال تام و عاملیت کیفری برای هوش مصنوعی کنونی است که در این پژوهش نیز مورد تأکید قرار گرفته است.

۴. جرایم منتسب به هوش مصنوعی و ارزیابی میزان استقلال

بررسی جرایم منتسب به هوش مصنوعی[۱] نقش تعیین‌کننده‌ای در تحدید مرزهای مفهومی میان ابزارگونگی[۲] و فاعل مستقل بودن[۳] این سامانه‌ها در سپهر حقوق کیفری ایفا می‌کند؛ چراکه این رویدادها، در پیوند با مباحث نظری گفتارهای پیشین، به‌مثابه آزمونی تجربی برای محک میزان و نوع استقلالی عمل می‌کنند که هوش مصنوعی در عمل از خود بروز داده است. در این نگرش، سنجش درجهٔ استقلال بروزیافته، پرسشی مقدّم و تحلیلی است که پاسخ آن، شرط هرگونه داوری پسینی دربارهٔ جایگاه حقوقی این سامانه‌ها به‌شمار می‌رود؛ به این معنا که پیش از هر سخن، باید روشن ساخت که این رویدادها، سامانه را بر طیفی که یک سوی آن «آلت فعل» و سوی دیگرش «فاعل مستقل» است، در کدام نقطه می‌نشانند. از این رو، رویدادهای یادشده که هریک نقطه‌ای متفاوت با این طیف را آشکار می‌سازد، ذیل سه عنوان ۱. جرایم ناشی از سوگیری، تعصب و تبعیض در الگوریتم؛ ۲. جرایم ناشی از مسائل ایمنی؛ ۳. جرایم ناشی از سوءاستفاده و نتایج ناخواستهٔ الگوریتم‌های هوش مصنوعی، بررسی می‌شوند.

۴. ۱. جرایم ناشی از سوگیری در الگوریتم و میزان استقلال هوش مصنوعی

مدل‌های هوش مصنوعی، دانش و الگوهای خود را از داده‌هایی که بر مبنای آن‌ها آموزش دیده‌اند، استخراج می‌کنند؛ بنابراین، چنانچه این داده‌ها واجد سوگیری‌های تاریخی باشند، یا به نحو متوازن و جامع، معرفِ جمعیت هدف نباشند، احتمال دارد سامانهٔ هوش مصنوعی این سوگیری‌ها را در پیش‌بینی‌ها، تصمیم‌گیری‌ها و خروجی‌های خود بازتولید و تشدید نماید و این سوگیری[۴،۵] در سامانه‌های هوش

---

۱. تعبیر جرایم در اینجا ناظر بر رخدادهای زیان‌باری است که ظرفیت طرح و ارزیابی در قلمرو حقوق کیفری را دارند، هرچند همهٔ مصادیق مورد اشاره لزوماً به تعقیب یا محکومیت کیفری نینجامیده‌اند.

2. Instrumentality

3. Autonomous Agency

۴. تعصب در سیستم‌های هوش مصنوعی می‌تواند به نتایج ناعادلانه و مضر منجر شود. با این توضیح که مدل‌های هوش مصنوعی از داده‌هایی که روی آن‌ها آموزش دیده‌اند یاد می‌گیرند و اگر این داده‌ها دارای سوگیری‌های تاریخی باشد یا نمایندهٔ جمعیت هدف نباشد، احتمال دارد سیستم هوش مصنوعی این سوگیری‌ها را در پیش‌بینی‌ها یا تصمیم‌های خود

مصنوعی، می‌تواند به بروز نتایج تبعیض‌آمیز و زیان‌بار منجر شود.

مطالعات متعددی مؤید گزارهٔ پیش‌گفته است؛ برای نمونه، پژوهشی در سال ۲۰۱۸ از سوی محققان دانشگاه MIT با عنوان «سایه‌های جنسیتی بر روی سیستم‌های طبقه‌بندی جنسیتی تجاری»[1] انجام پذیرفت و پژوهشگران دریافتند که سامانه‌های تحلیل چهرهٔ هوش مصنوعی توسعه‌یافته از سوی شرکت‌هایی نظیر IBM و مایکروسافت، در طبقه‌بندی چهره‌ها بر اساس جنسیت، میان افراد دارای پوست تیره‌تر و روشن‌تر، تبعیض قائل می‌شوند و نتیجه گرفتند که سوگیری‌های موجود در داده‌های آموزشی، به سامانه‌های هوش مصنوعی منتقل شده، به نتایج ناعادلانه و تبعیض‌آمیز منتهی می‌گردد (Buolamwini & Gebru, 2018: 77-79).

رویداد اپل کارت[2] نیز هرچند از منظری متفاوت، در همین سیاق درخور تأمل است. در نوامبر ۲۰۱۹، شماری از کاربران، ازجمله یکی از بنیان‌گذاران شرکت اپل، گزارش‌هایی مبنی بر تخصیص محدودیت اعتباری به‌مراتب کمتر برای همسرانشان، علی‌رغم برخورداری ایشان از امتیاز اعتباری مشابه یا حتی بهتر، منتشر ساختند[3] و به این ترتیب، اپل کارت که محصول مشترک شرکت‌های اپل و گلدمن ساکس[4] است، با اتهام سوگیری جنسیتی در الگوریتم تعیین محدودیت اعتباری مواجه شد. وزارت خدمات مالی ایالت نیویورک[5] در پی این اتهامات، تحقیقاتی گسترده را آغاز نمود و در گزارش نهایی خود (مارس ۲۰۲۱)، پس از تحلیل داده‌های پذیره‌نویسی صدها هزار متقاضی، به این نتیجه رسید که شواهدی دال بر تبعیض جنسیتی ناقض قوانین اعطای اعتبار وجود ندارد؛ لکن همان گزارش، فقدان شفافیت در فرایند تصمیم‌گیری الگوریتمی و ناتوانی متقاضیان از فهم چرایی تصمیمات اتخاذشده را عامل اصلی بی‌اعتمادی عمومی و شکل‌گیری گمانهٔ تبعیض دانست و بانک یادشده را به بهبود سازوکارهای شفافیت

تداوم بخشد. برای مثال فرض کنید می‌خواهیم یک سیستم هوش مصنوعی برای پیش‌بینی موفقیت شغلی افراد طراحی کنیم. برای آموزش این سیستم، از داده‌های مربوط به افراد موفق در ۵۰ سال گذشته استفاده می‌کنیم. حال اگر در طول این ۵۰ سال، به دلایل تاریخی و اجتماعی، زنان کمتر به پست‌های مدیریتی می‌رسیدند و در نتیجه، تعداد مدیران زن موفق در داده‌های ما کم باشد، سیستم هوش مصنوعی نیز به احتمال زیاد، زنان را در پیش‌بینی‌های خود برای موفقیت شغلی، کمتر درنظر می‌گیرد و این تبعیض تاریخی را در دنیای امروز نیز اعمال می‌کند. برای مطالعهٔ جامع، ر.ک. به اثر تأثیرگذار زیر:

-Noble, S. U. (2018). Algorithms of oppression: How search engines reinforce racism. In Algorithms of oppression. New York university press.

5. Bias

1. Gender Shades study on commercial gender classification systems

2. Apple Card

3. https://www.latimes.com/business/story/2019-11-11/apple-sexism-apple-card

4. Goldman Sachs

5. New York State Department of Financial Services

و تسهیل تجدیدنظرخواهی متقاضیان رهنمون ساخت[۱]. اهمیت تحلیلی این رویداد برای پژوهش حاضر در آن است که نشان می‌دهد خصلت جعبهٔ سیاه این سامانه‌ها- که پیش‌تر تشریح گردید- چگونه حتی در غیاب سوگیری اثبات‌شده، راستی‌آزمایی عملکرد سامانه را دشوار می‌سازد و ظنّ تبعیض را دامن می‌زند؛ و همین دشواری در بازسازی منطق تصمیم، خود شاهدی تجربی بر فقدان قابلیت توضیح روایی در این سامانه‌هاست، چنان‌که پاسخگویی در برابر اتهام نیز نه از خود سامانه، بلکه از نهاد انسانی بهره‌بردار آن مطالبه گردید و بار توضیح، یکسره بر دوش عامل انسانی نهاده شد.

علاوه بر موارد یادشده، استفاده از هوش مصنوعی در فرایند استخدام، با سوگیری همراه بوده است. یکی از این نمونه‌ها، ابزار استخدام مبتنی بر هوش مصنوعی آمازون[۲] است که در سال ۲۰۱۸ به دلیل تبعیض علیه نامزدهای زن[۳] مورد انتقاد شدید قرار گرفت و در نهایت کنار گذاشته شد. این سیستم که با استفاده از رزومه‌های دریافتی در طول یک دهه آموزش دیده بود، به‌طور ناخواسته الگوهای تبعیض‌آمیز[۴] موجود در داده‌ها را یاد گرفته، در فرایند رتبه‌بندی متقاضیان، به نفع داوطلبان مرد عمل می‌کرد[۵].

در خصوص میزان استقلال هوش مصنوعی در بروز این سوگیری‌ها، باید گفت که آنچه رخ می‌دهد بازتولید وفادارانهٔ الگوهای آماری نهفته در داده‌های ورودی است؛ به این معنا که سامانه در اینجا همچون «آینه‌ای» الگوهای موجود را منعکس می‌سازد، بی‌آنکه از قدرت تشخیص یا اصلاح سوگیری‌های ذاتی داده‌ها برخوردار باشد و بی‌آنکه در تعیین غایت خویش کمترین نقشی ایفا کند. این سوگیری‌ها، خواه ناشی از عدم توازن در داده‌ها[۶] باشند، خواه برخاسته از بازنمایی ناکافی[۷] یا تعصبات تاریخی و اجتماعی، همگی مبیّن یک واقعیت واحدند و آن اینکه عملکرد سامانه، تابعی متعیّن از الگوهای موجود در داده‌های ورودی[۸] است و نه برخاسته از کنشی خودآیین. از این منظر، نمونهٔ آمازون از حیث تحلیلی، روشنگرتر می‌نماید؛ چه اینکه آنچه در آغاز امر «یادگیری مستقل» جلوه می‌کند، در حقیقت بهینه‌سازی وفادارانه به سمت هدفی است که داده‌های تاریخی نامتوازن آن را از پیش رقم زده‌اند و بدین‌سان، بداعت ظاهری این رفتار، نه از

---

۱. برای دسترسی به فایل گزارش و نتیجهٔ نهایی تحقیقات، ر.ک. به لینک زیر:
https://www.dfs.ny.gov/system/files/documents/2021/03/rpt_202103_apple_card_investigation.pdf

2. Amazon's AI recruitment tool

3. https://www.reuters.com/article/amazoncom-jobs-automation/insight-amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSL2N1WQ1E9/

۴. الگوهای تبعیض‌آمیز در داده‌ها، الگوها و روندهایی هستند که به طور ناخواسته یا عمدی، نابرابری‌ها و تبعیضات اجتماعی را در خود جای داده‌اند. این الگوها می‌توانند در انواع مختلف داده‌ها، ازجمله داده‌های متنی، تصویری، صوتی و عددی وجود داشته باشند و به شکل‌های مختلفی ظاهر شوند.

5. https://slate.com/business/2018/10/amazon-artificial-intelligence-hiring-discrimination-women.html

6. Data Imbalance

7. Underrepresentation

8. Input Data

جنس «غایت‌آفرینی»، بلکه از سنخ «استخراج الگو» است. رویداد اپل کارت نیز هرچند از حیث اثبات سوگیری در جایگاهی متفاوت می‌نشیند، از منظر سنجش عاملیت مستقل به همین نتیجه راه می‌برد؛ چراکه آنچه در آن رویداد محل نزاع بود، نه اراده یا غایت‌گزینی سامانه، بلکه شفافیت سازوکاری بود که غایت و منطقش را دیگری تعیین کرده بود. این مختصات نشان می‌دهد که میزان استقلال بروزیافته در این دسته از رویدادها، در نازل‌ترین مرتبهٔ طیف و در نزدیک‌ترین فاصله با مفهوم «آلت فعل» جای می‌گیرد.

۴. ۲. جرایم ناشی از مسائل ایمنی و میزان استقلال هوش مصنوعی

بروز خطا یا عملکرد ناصحیح در سامانه‌های هوش مصنوعی، در کاربردهایی با ریسک بالا نظیر خودروهای خودران[1] که با جان انسان‌ها سروکار دارد می‌تواند منجر به وقوع حوادث فاجعه‌بار و خسارات جانی و مالی جبران‌ناپذیری گردد. سال ۲۰۱۶، نقطهٔ عطفی در تاریخچهٔ تصادفات مرتبط با خودروهای خودران محسوب می‌شود. پیش از این تاریخ، آمار تصادفات منتسب به این فناوری بسیار محدود بود و هیچ‌گونه سانحهٔ منجر به فوت یا جراحت جدی که مستقیماً ناشی از نقص یا خطای سامانهٔ خودران باشد، به ثبت نرسیده بود و عامل انسانی به‌عنوان علت اصلی تصادف شناسایی می‌شد. اما در سال ۲۰۱۶، یک رانندهٔ خودروی تسلا، درحالی که از سامانهٔ رانندگی خودکار[2] استفاده می‌کرد، در اثر برخورد با یک کامیون تریلی در بزرگراهی واقع در ایالت فلوریدای ایالات متحدهٔ آمریکا، جان خود را ازدست داد (Stahl et al., 2023: 65).

حادثهٔ مهم‌تر و تأثیرگذارتر، در سال ۲۰۱۸ در ایالت آریزونای امریکا به‌وقوع پیوست که در آن یک خودروی خودران متعلق به شرکت اوبر[3]، با یک عابر پیاده که در حال عبور از عرض خیابان به همراه دوچرخهٔ خود بود، برخورد نمود و منجر به فوت وی شد. این واقعه، نخستین تصادف منجر به فوت بود که به‌طور مستقیم به نقص در عملکرد سامانهٔ خودران نسبت داده شد. بررسی‌های هیئت ملی ایمنی حمل و نقل ایالات متحده[4] مشخص کرد که سامانهٔ هوش مصنوعی خودرو، قادر به تشخیص صحیح

---

۱. خودروی خودران، خودرویی است که با استفاده از الگوریتم‌های پیچیدهٔ هوش مصنوعی، داده‌های دریافتی از حسگرها (مانند دوربین، رادار و لیدار) را پردازش و بدون نیاز به رانندهٔ انسانی، در جاده‌ها حرکت می‌کند. برای مطالعهٔ مسئولیت کیفری ناشی از خودروی خودران می‌توان به مقالات زیر مراجعه کرد:

– یکرنگی، محمد و برزگر، محمدرضا (۱۴۰۱). مسئولیت کیفری تولیدکننده و طراح خودروهای خودران در قبال صدمات وارده توسط آن‌ها. مطالعات حقوق کیفری و جرم‌شناسی، ۵۲ (۱)، ۱۲۳–۱۴۸.

– برزگر، محمدرضا و الهام, غلامحسین (۱۳۹۹). مسئولیت کیفری کاربر خودروی خودران در قبال صدمات وارده توسط آن. فصلنامهٔ پژوهش حقوق کیفری، سال هشتم، (۳۰)، ۲۰۹ –۲۲۱.

2. Autopilot
3. Uber
4. NTSB

عابر پیاده نبوده و همچنین شرکت اوبر، سامانهٔ ترمز اضطراری خودکار تعبیه‌شده در خودروی ولوو SUV مورد آزمایش را غیرفعال کرده بود[۱].

در خصوص میزان استقلال هوش مصنوعی در وقوع این تصادفات، باید گفت که برخلاف مصادیق سوگیری که سامانه در آن‌ها صرفاً نقشی بازتابی داشت، در اینجا با مرتبه‌ای بالاتر از استقلال عملیاتی روبه‌رو هستیم؛ زیرا خودروی خودران در لحظهٔ رانندگی، از آزادی عمل واقعی در گزینش سرعت، ترمز و مسیر برخوردار است و این، همان «استقلال دگرآیین» است که در گفتار پیشین تبیین گردید. با وجود این، واکاوی دقیق این حوادث آشکار می‌سازد که نتیجهٔ زیان‌بار، نه ماحصل یک تصمیم مستقل، بلکه برخاسته از مرز توانمندی ادراکی و طبقه‌بندی سامانه است؛ به دیگر سخن، خودرو «تصمیم» به برخورد نگرفت، بلکه در محدودهٔ عملیاتی خویش از ادراک صحیح موقعیت بازماند، و غیرفعال‌سازی سامانهٔ ترمز اضطراری نیز خود مؤید آن است که عامل تعیین‌کننده، پیکربندی از پیش تعیین شده بود و نه گزینشی خودانگیخته. بدین‌سان، استقلال بروزیافته در اینجا در ساحت «چگونگی» انجام وظیفه محدود و مقیّد می‌ماند، حال آنکه «غایت» آن یکسره دگرآیین است؛ و خطا، در ساحت فنی سامانه رخ می‌دهد، نه اینکه مربوط به اراده باشد. این مختصات، چنین رویدادهایی را در نقطه‌ای میانی بر طیف استقلال می‌نشاند؛ فراتر از بازتاب صرف دستهٔ نخست، اما همچنان فروتر از آستانهٔ فاعلیت مستقل.

۴. ۳. جرایم ناشی از سوءاستفاده و نتایج ناخواستهٔ الگوریتم‌ها و میزان استقلال هوش مصنوعی

الگوریتم‌های هوش مصنوعی، علی‌رغم توانمندی، می‌توانند به ابزاری برای سوءاستفاده و اعمال نیات مخرب بدل گردند؛ چنان‌که در سال ۲۰۱۶، شرکت مایکروسافت چت‌بات هوش مصنوعی خود موسوم به «تای»[۲] را عرضه نمود که با هدف تعامل با کاربران جوان هجده تا بیست‌وچهار ساله در بستر شبکه‌های اجتماعی طراحی شده بود و از الگوریتم یادگیری تقویتی[۳] برای ارتقای مهارت‌های مکالمه‌ای خویش بهره می‌برد. هرچند پیش از عرضهٔ این چت‌بات، تلاش‌های فراوانی برای ایمن‌سازی آن صورت گرفته و آزمایش‌های متعدد و فیلترهای گوناگونی اعمال شده بود، «تای» اندکی پس از شروع به کار، به تولید محتوای توهین‌آمیز، نژادپرستانه و مغایر با هنجارهای اجتماعی مبادرت ورزید؛ رفتاری نابهنجار که از دست‌کاری عامدانهٔ چت‌بات به دست برخی کاربران ناشی می‌شد که آگاهانه ورودی‌های نامناسب و مخرب به آن عرضه می‌کردند (Housley, 2021: 42) و در نتیجهٔ همین سوءاستفاده، مایکروسافت تنها یک روز پس از عرضهٔ عمومی، «تای» را غیرفعال نمود (Swanson, 2018: 1205).

---

۱. برای دیدن تصویر تصادف و گزارش کامل هیئت ملی ایمنی حمل و نقل ایالات متحده در خصوص این حادثه، ر.ک. به لینک زیر: https://www.ntsb.gov/investigations/Pages/HWY18MH010.aspx

2. Tay

3. Reinforcement Learning

در خصوص میزان استقلال هوش مصنوعی در این دسته از رویدادها باید گفت که «تای» از حیث سنجش استقلال، آموزنده‌ترین مصداق است؛ چراکه رفتار آن در نگاه نخست نمونه‌ای از «رفتار نوظهور» مستقل می‌نماید، چنان‌که گویی خودِ سامانه «تصمیم» به تولید محتوای ناهنجار گرفته است، اما واکاوی دقیق خلاف این را اثبات می‌کند؛ زیرا خروجی‌های این چت‌بات تابعی مستقیم از ورودی‌های خصمانه‌ای بود که کاربران به آن خوراندند و سازوکار یادگیری تقویتی، آن را به بازتاب‌دهنده‌ای وفادار نسبت به همان ورودی‌ها بدل ساخت. به این ترتیب، «تای» به‌جای آنکه عاملیتی مستقل از خود نشان دهد، نهایت «دگرآیینی» را به‌نمایش گذاشت؛ رفتاری که به دست عاملانی بیرونی و از مجرای همان سازوکار یادگیری تعیین می‌شد. نکتهٔ محوری و ظریف آنجاست که همان خصیصه‌ای که غالباً نشانهٔ استقلال انگاشته می‌شود، یعنی قابلیت یادگیری و انطباق، در تحلیل نهایی خود به مجرای وابستگی بدل می‌گردد؛ زیرا انطباق در اینجا، نه به‌معنای «خودتعیین‌گری»، بلکه به‌معنای تأثیرپذیری سامانه از تعیّن‌بخشی بیرونی است.

برآیند گونه‌شناسی پیش‌گفته را می‌توان چنین جمع‌بندی کرد که این سه دسته از رویدادها، در مجموع، سه نقطهٔ متمایز بر طیف استقلال را روشن می‌سازند؛ از بازتاب منفعلانه و آینه‌وار الگوهای داده در مصادیق سوگیری، تا استقلال عملیاتی مقیّد و دگرآیین در حوزهٔ ایمنی، و سرانجام دگرآیینی پنهان در پس ظاهر خودآیین در مورد سوءاستفاده. وجه مشترک تعیین‌کنندهٔ این مصادیق آن است که در هیچ‌یک، استقلال بروزیافته از آستانهٔ «عاملیت مستقل» فراتر نمی‌رود؛ چراکه در تمامی موارد، محور تبیینی رفتار سامانه بر لایهٔ دگرآیین آن، یعنی غایات ازپیش تعیین‌شده، داده‌های آموزشی و ورودی‌های بیرونی، استوار می‌ماند و نه بر کنشی خودبنیاد که از اراده‌ای سرچشمه گرفته باشد. این یافته، مدعای مفهومی گفتارهای پیشین را مبنی بر محدود و دگرآیین بودن استقلال هوش مصنوعی در ساحت تجربه نیز مدلل می‌سازد و آشکار می‌کند که توصیف این فناوری به «استقلال محدود»، تبیینی منطبق با شواهد عینی است. بر این اساس، تحلیل مصادیق همان نتیجه‌ای را که از مجرای واکاوی مفهومی و فنی به‌دست آمده بود تأیید می‌کند؛ اینکه هوش مصنوعی کنونی، حتی در پیشرفته‌ترین صورت‌های خود، فاقد آن مرتبه از استقلال است که بتوان آن را مستقل و مبتنی بر ارادهٔ آزاد خویش، «فاعل مستقل» پدیدآورندهٔ رفتار یا خروجی مجرمانه تلقی کرد.

۵. نتیجه‌گیری

پژوهش حاضر با هدف واکاوی مؤلفهٔ «استقلال» در سامانه‌های هوش مصنوعی و در پی پاسخ به این پرسش سامان یافت که آیا این سامانه‌ها اساساً از آن خصیصه‌ای که پیش‌شرط فاعلیت حقوقی به‌شمار می‌آیند، برخوردارند؛ و برآیند تحلیل آن بود که هوش مصنوعی کنونی با همهٔ پیشرفت‌های خود، به آن

پایه از استقلال و عاملیت مستقل نرسیده است که اسناد مسئولیت کیفری به خود سامانه بدان منوط است. در مقام تبیین این یافته، نشان داده شد که رفتار نوظهور این سامانه‌ها، علی‌رغم بداعت ظاهری، از سنخ استخراج الگوست و نه غایت‌آفرینی؛ چراکه خروجی‌های غیرمنتظرهٔ سامانه، در تحلیل نهایی، تابعی از الگوریتم‌های اولیه، داده‌های آموزشی و اهداف تعیین‌شده از سوی عاملان انسانی باقی می‌مانند و از این رو، شرط خودتعیین‌گری غایات را برنمی‌آورند. همچنین روشن گردید که قابلیت توضیح نیز حتی در سامانه‌های مولّد که به‌ظاهر برای رفتار خویش دلیل عرضه می‌کنند، به آستانهٔ توضیح روایی نمی‌رسد؛ زیرا معیار تمایز، تعلق توضیح به همان منظومهٔ قصدمندی است که فعل از آن صادر گردیده است و سامانه‌های یادشده در فقدان چنین منظومه‌ای، تنها ادامهٔ فرایند تولید متن را در قالب توضیح عرضه می‌دارند؛ و با انتفای همین دو خصیصه، استقلال علّی نیز به‌تبع آن منتفی می‌گردد، چراکه سامانه‌ای که غایت فعلش را دیگری تنسیق کرده و از ساحت قصدمندی قابل ارجاعی بی‌بهره است، نمی‌تواند حلقه‌ای خودبنیاد در زنجیرهٔ علیت به‌شمار آید.

تحلیل مصادیق تجربیِ رخدادهای زیان‌بار و مجرمانهٔ مرتبط با هوش مصنوعی نیز به‌مثابه آزمونی واقعی، همین یافتهٔ مفهومی را در ساحت عمل استوار ساخت؛ به این صورت که در رویدادهای برخاسته از سوگیری، سامانه از حد انعکاس الگوهای نهفته در داده‌ها فراتر نرفت، در حوادث ایمنی، آزادی عملِ سامانه به چگونگی اجرای وظیفه‌ای محدود ماند که غایت آن را دیگری مقرر داشته بود و در مورد سوءاستفاده، همان سازوکار یادگیری که مظهر استقلال انگاشته می‌شد، در مقام عمل، مجرای اثرپذیری از ارادهٔ کاربران بیرونی گردید؛ و در هیچ‌یک، تبیین رفتار سامانه بدون ارجاع به تعیّن‌بخشی بیرونی ممکن نشد. بر این اساس، تلقی سامانهٔ هوشمند به‌عنوان «فاعل مستقل» پدیدآورندهٔ جرم فاقد وجاهت است؛ زیرا ظرفیت‌هایی همچون یادگیری، انطباق‌پذیری و ظهور رفتارهای نو، خود از مجاری همان وابستگی‌اند و با عنایت به همین مهم، مفاهیم «استقلال محدود» و «استقلال دگرآیین» توصیفی واقع‌بینانه از وضعیت کنونی این فناوری به‌دست می‌دهند؛ افزون بر این، درخور تأمل است که نتیجه‌گیری حاضر با جهت‌گیری‌های غالب در سیاست‌گذاری‌های تقنینی و تنظیمی بین‌المللی نیز انطباق دارد و این هم‌راستایی، هرچند فی‌نفسه دلیلی بر مدعا به‌شمار نمی‌رود، قرینه‌ای است بر آنکه تحلیل مفهومی پژوهش از فهم عملی نظام‌های حقوقی پیشرو فاصله‌ای ندارد. در عین حال نباید از نظر دور داشت که پژوهش در باب امکان یا امتناع فاعلیت هوش مصنوعی، ریشه در مشاهدهٔ رشد توانمندی‌ها و سطح استقلال عملیاتی این سامانه‌ها دارد و از کنجکاوی نظری محض برنخاسته است؛ چه‌بسا در آینده، با دستیابی هوش مصنوعی به سطوحی از عاملیت که امروز موجود نیست، پاسخ این پرسش دگرگون گردد و این فناوری در جایگاه طرف خطاب قواعد مسئولیت قرار گیرد.

References

1. Allhoff, F., Evans, N. G., & Henschke, A. (2013). *Routledge handbook of ethics and war: Just war theory in the 21st century*. Routledge. https://doi.org/10.4324/9780203107164
2. Barzegar, Mohammad Reza & Gholam Hossein Elham (2020). Criminal liability of the self-driving cars for the injury caused by it, *Criminal Law Research*, Year 8, No. 30 (In Persian) https://doi.org/10.22054/jclr.2020.39009.1838
3. Bathaee, Y. (2017). The artificial intelligence black box and the failure of intent and causation. Harv. JL & Tech., 31, 889.
4. Beckers, A., & Teubner, G. (2021). *Three liability regimes for artificial intelligence: Algorithmic actants, hybrids, crowds.* Bloomsbury Publishing. https://doi.org/10.5040/9781509949366
5. Bertolini, A. (2013). Robots as products: The case for a realistic analysis of robotic applications and liability rules. *Law, Innovation and Technology*, 5(2), 214-247. https://doi.org/10.5235/17579961.5.2.214
6. Buolamwini, J., & Gebru, T. (2018, January). *Gender shades: Intersectional accuracy disparities in commercial gender classification.* In Conference on fairness, accountability and transparency (pp. 77-91). PMLR. https://doi.org/10.1145/3630106.3658947
7. Dworkin, G. (2015). The nature of autonomy. *Nordic Journal of Studies in Educational Policy,* 2015(2), 28479. https://doi.org/10.3402/nstep.v1.28479
8. European Parliament and the Council of the European Union. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union, L series. http://data.europa.eu/eli/reg/2024/1689/oj
9. Hallevy, Gabriel. (2019). When Robots Kill (Farhad Shahideh and Tahereh Ghavanloo, translators). Tehran: Mizan Publishing (In Persian).
10. Hellström, T. (2013). On the moral responsibility of military robots. Ethics and information technology, 15, 99-107. https://doi.org/10.1007/s10676-012-9301-2
11. Housley W (2021) *Society in the Digital Age: An Interactionist Perspective.* Sage Publications Ltd. https://doi.org/10.4135/9781526486295
12. Kingston, J. K. (2016). *Artificial intelligence and legal liability.* In Research and Development in Intelligent Systems XXXIII: Incorporating Applications and Innovations in Intelligent Systems XXIV 33 (pp. 269-279). Springer International Publishing. https://doi.org/10.1007/978-3-319-47175-4_20
13. Kirpichnikov, D., Pavlyuk, A., Grebneva, Y., & Okagbue, H. (2020). *Criminal liability of the artificial intelligence.* In E3S web of conferences (Vol. 159). EDP Sciences. https://doi.org/10.1051/e3sconf/202015904025
14. Köhler, S., Roughley, N., & Sauer, H. (2017). Technologically blurred accountability?: Technology, responsibility gaps and the robustness of our everyday conceptual scheme. In Moral agency and the politics of responsibility (pp. 51-68). Routledge. https://doi.org/10.4324/9781315201399-4
15. Lima, D. (2018). Could AI agents be held criminally liable: Artificial intelligence and the challenges for criminal law. *South Carolina Law Review*, 69(3), 677-696. https://scholarcommons.sc.edu/cgi/viewcontent.cgi?article=4253&context=sclr

16. Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and information technology*, 6, 175-183. https://doi.org/10.1007/s10676-004-3422-1
17. Mosizadeh, Reza & Golriz, Hadi. (2015). *A Comprehensive Tri-lingual Dictionary On Law.* Tehran: Mizan Publishing (In Persian).
18. Najib Husni, Mahmoud. (2007). *The Causal link in Criminal Law* (Seyed Ali Abbas Niay Zare, translator). Tehran: Qods Cultural Institute (In Persian).
19. Novelli, C., Casolari, F., Rotolo, A., Taddeo, M., & Floridi, L. (2024). Taking AI risks seriously: a new assessment model for the AI Act. AI & SOCIETY, 39(5), 2493-2497. https://doi.org/10.1007/s00146-023-01723-z
20. Oimann, A. K. (2023). The responsibility gap and LAWS: A critical mapping of the debate. Philosophy & Technology, 36(1), 3. https://doi.org/10.1007/s13347-022-00602-7
21. Owen, D. G. (Ed.). (1995). *Philosophical foundations of tort law*. Clarendon press. https://doi.org/10.1093/acprof:oso/9780198265795.001.0001
22. Rajabi, A. (2019). Liability of Artificial Intelligence; the Reflection of Developments in the Liability Rules. *Comparative Law Review*, 10(2), 449-466. 10.22059/JCL.2019.274782.633787
23. Santoni de Sio, F., & Mecacci, G. (2021). Four responsibility gaps with artificial intelligence: Why they matter and how to address them. Philosophy & Technology, 34, 1057–1084. https://doi.org/10.1007/s13347-021-00450-x
24. Schmid, U., Klügl, F., & Wolter, D. (Eds.). (2020). KI 2020: Advances in Artificial Intelligence: 43rd German Conference on AI, Bamberg, Germany, September 21–25, 2020, Proceedings (Vol. 12325). Springer Nature.
25. Schraagen, J. M. (2024). *Bounded Autonomy. In Responsible Use of AI in Military Systems.* Chapman and Hall/CRC. https://doi.org/10.1201/9781003410379-22
26. Stahl, B. C., Schroeder, D., & Rodrigues, R. (2023). Ethics of artificial intelligence: Case studies and options for addressing ethical challenges. Springer Nature.
27. Swanson, G. (2018). Non-Autonomous Artificial Intelligence Programs and Products Liability: How New AI Products Challenge Existing Liability Models and Pose New Financial Burdens. *Seattle University Law Review*, 42. https://digitalcommons.law.seattleu.edu/cgi/viewcontent.cgi?article=2607&context=sulr
28. Totschnig, W. (2020). Fully autonomous AI. *Science and Engineering Ethics*, 26(5), 2473-2485. https://doi.org/10.1007/s11948-020-00243-z
29. Veluwenkamp, H. (2025). What responsibility gaps are and what they should be. Ethics and Information Technology, 27, Article 14. https://doi.org/10.1007/s10676-025-09823-8
30. Walsh, K. R., Mahesh, S., & Trumbach, C. C. (2021). Autonomy in AI Systems. *The Journal of Technology Studies*, 47(1), 38-47. https://doi.org/10.21061/jts.400