Comparative Studies in Jurisprudence, Law, and Politics

From the Rational Actor to the Rational Algorithm: Rethinking the Concept of Criminal Capacity for Autonomous Artificial Intelligence in the U.S. Criminal Justice System

1. Hamideh Elahi: Department of Law, Sar.C., Islamic Azad University, Sari, Iran
2. Asghar Abbasi*: Department of Law, Cha.C., Islamic Azad University, Chalus, Iran. Email: drabbasi191@gmail.com (Corresponding Author)
3. Seyyed Abdollah MirGhiasi: Department of Political Science, CT.C., Islamic Azad University, Tehran, Iran

ABSTRACT

The traditional foundation of criminal law rests on the model of the "rational actor"; that is, a human being endowed with awareness, volition, and the capacity for moral reasoning and choice, whose attributes constitute the basis for the recognition of legal personality and criminal responsibility. With the emergence of advanced and autonomous artificial intelligence agents capable of independent, goal-directed decision-making, this classical model has encountered a fundamental challenge. Through a critical examination of the transition from the "rational actor" to the "rational algorithm," this article argues that the traditional concept of legal capacity requires substantial reconsideration if it is to encompass non-human autonomous agents. On this basis, the article demonstrates that anthropocentric criteria of criminal responsibility, including the actus reus and mens rea elements of an offense, are no longer fully adequate to address emerging technological realities. The study subsequently proposes a functionalist framework for redefining legal capacity, according to which capacity is decoupled from biological consciousness and exclusively human characteristics and is instead assessed on the basis of the degree of autonomy, functional rationality, and the effects and consequences of an agent's conduct. By analyzing the characteristics of advanced AI agents, including adaptive learning, operational independence, and the ability to undertake complex decisions and tasks, the article examines the possibility of establishing a new form of "electronic legal personhood" with graduated levels of responsibility. The article concludes that this conceptual transformation is not merely a theoretical hypothesis but an unavoidable necessity for addressing responsibility gaps, ensuring accountability, preserving social order, and guiding legal policymaking in a world increasingly shaped by algorithmic decision-making.

Keywords: *legal capacity, autonomous agents, functional rationality, artificial intelligence and criminal law*

How to cite: Elahi, H., Abbasi, A., & MirGhiasi, S. A. (2026). From the Rational Actor to the Rational Algorithm: Rethinking the Concept of Criminal Capacity for Autonomous Artificial Intelligence in the U.S. Criminal Justice System. *Comparative Studies in Jurisprudence, Law, and Politics, 8*(1), 1-24.

Submit Date: 22 December 2025
Revise Date: 28 April 2026
Accept Date: 05 May 2026
Publish Date: 18 May 2026

پژوهش‌های تطبیقی فقه، حقوق و سیاست

از کنشگر عقلانی تا الگوریتم عقلانی: بازاندیشی در مفهوم اهلیت کیفری برای هوش مصنوعی خودمختار در نظام کیفری آمریکا

۱. حمیده الهی: گروه حقوق، واحد ساری، دانشگاه آزاد اسلامی، ساری، ایران
۲. اصغر عباسی*: گروه حقوق، واحد چالوس، دانشگاه آزاد اسلامی،چالوس، ایران. پست الکترونیک: drabbasi191@gmail.com (نویسنده مسئول)
۳. سیدعبدالله میرغیاثی: گروه علوم سیاسی، واحد تهران مرکزی، دانشگاه آزاد اسلامی، تهران، ایران

چکیده

بنیاد سنتی حقوق کیفری بر الگوی «کنشگر عقلانی» استوار است؛ یعنی انسانی آگاه، دارای اراده و توانایی استدلال اخلاقی و انتخاب، که مبنای شناسایی شخصیت حقوقی و مسئولیت کیفری را تشکیل می‌دهد. با ظهور عامل‌های هوش مصنوعی پیشرفته و خودمختار که قادر به تصمیم‌گیری مستقل و هدف‌محور هستند، این الگوی کلاسیک با چالشی بنیادین روبه‌رو شده است. این مقاله با بررسی انتقادی گذار از «کنشگر عقلانی» به «الگوریتم عقلانی»، استدلال می‌کند که مفهوم سنتی اهلیت حقوقی نیازمند بازاندیشی اساسی است تا بتواند عامل‌های خودمختار غیرانسانی را نیز در بر گیرد. بر این اساس، مقاله نشان می‌دهد که معیارهای انسان‌محور مسئولیت کیفری، از جمله عنصر مادی و عنصر معنوی جرم، دیگر پاسخگوی واقعیت‌های نوظهور فناوری نیستند. در ادامه، پژوهش چارچوبی کارکردگرایانه برای بازتعریف اهلیت حقوقی پیشنهاد می‌کند که در آن، اهلیت از آگاهی زیستی و ویژگی‌های صرفاً انسانی منفک شده و بر پایه میزان خودمختاری، عقلانیت کارکردی و آثار و پیامدهای رفتار عامل ارزیابی می‌شود. با تحلیل ویژگی‌های عامل‌های پیشرفته هوش مصنوعی، از جمله یادگیری تطبیقی، استقلال عملیاتی و توانایی انجام تصمیم‌ها و وظایف پیچیده، امکان ایجاد گونه‌ای جدید از «شخصیت حقوقی الکترونیکی» با سطوح تدریجی مسئولیت بررسی می‌شود. مقاله نتیجه می‌گیرد که این تحول مفهومی صرفاً یک فرضیه نظری نیست، بلکه ضرورتی اجتناب‌ناپذیر برای رفع خلأهای مسئولیت، تضمین پاسخگویی، حفظ نظم اجتماعی و هدایت سیاست‌گذاری حقوقی در جهانی است که به‌طور فزاینده تحت تأثیر تصمیم‌گیری‌های الگوریتمی قرار دارد.

واژگان کلیدی: *اهلیت حقوقی، عامل‌های خودمختار، عقلانیت کارکردی، هوش مصنوعی و حقوق کیفری*

نحوه استناددهی: الهی، حمیده.، عباسی، اصغر.، و میرغیاثی، سیدعبدالله. (۱۴۰۵). از کنشگر عقلانی تا الگوریتم عقلانی: بازاندیشی در مفهوم اهلیت کیفری برای هوش مصنوعی خودمختار در نظام کیفری آمریکا. *پژوهش‌های تطبیقی فقه، حقوق و سیاست*، ۸(۱)، ۲۴-۱.

تاریخ ارسال: ۱ دی ۱۴۰۴
تاریخ بازنگری: ۸ اردیبهشت ۱۴۰۵
تاریخ پذیرش: ۱۵ اردیبهشت ۱۴۰۵
تاریخ چاپ: ۲۸ اردیبهشت ۱۴۰۵

مقدمه

حقوق کیفری به‌عنوان مهم‌ترین ابزار جامعه برای انتساب مسئولیت، سرزنش مرتکب و پیشگیری از جرم، همواره بر مبنایی انسان‌محور استوار بوده است. اصول بنیادین این شاخه از حقوق، یعنی عنصر مادی جرم و عنصر معنوی جرم، بر این پیش‌فرض استوارند که مرتکب، انسانی برخوردار از آگاهی، اراده، قصد و توانایی داوری اخلاقی است. این تصویر از «کنشگر عقلانی» که ریشه در اندیشه‌های عصر روشنگری دارد، فردی را ترسیم می‌کند که با سنجش هزینه‌ها و منافع، دست به انتخاب می‌زند و به همین دلیل می‌توان او را از منظر حقوقی مسئول اعمال خویش دانست. در نتیجه، مفهوم اهلیت حقوقی که مبنای برخورداری از حقوق، تکالیف و مسئولیت‌های قانونی است، بر همین برداشت از شخصیت و عاملیت انسانی بنا شده است.

بااین‌حال، ظهور و گسترش عامل‌های خودمختار هوش مصنوعی، این بنیان نظری را با چالشی اساسی مواجه ساخته است. سامانه‌های پیشرفته هوش مصنوعی، از خودروهای خودران و سامانه‌های تسلیحاتی خودمختار گرفته تا پلتفرم‌های معاملات الگوریتمی و مدل‌های مولد، دیگر صرفاً ابزارهایی منفعل برای اجرای دستورات انسان نیستند؛ بلکه قادرند محیط پیرامون خود را ادراک کنند، داده‌ها را از طریق مدل‌های پیچیده پردازش نمایند، بر اساس اهداف از پیش تعیین‌شده یا آموخته‌شده تصمیم بگیرند و بدون مداخله مستقیم و لحظه‌ای انسان بر محیط پیرامون اثر بگذارند. این سامانه‌ها از نوعی «عقلانیت کارکردی» برخوردارند؛ بدین معنا که می‌توانند برای دستیابی به اهداف مشخص، تصمیم‌هایی اتخاذ کنند که در بسیاری از موارد از حیث سرعت، دقت و کارآمدی با تصمیم‌گیری انسان برابری کرده یا حتی از آن فراتر رود. با وجود این، چنین سامانه‌هایی فاقد آگاهی ذهنی، تجربه زیسته و ادراک اخلاقی هستند؛ ویژگی‌هایی که تاکنون زیربنای انتساب مسئولیت کیفری در نظام‌های حقوقی محسوب می‌شده‌اند.

همین گسست میان توانایی‌های عملکردی هوش مصنوعی و معیارهای سنتی مسئولیت کیفری، نظام حقوقی را با بحرانی جدی روبه‌رو ساخته است. هرگاه یک عامل خودمختار هوش مصنوعی موجب ورود خسارت یا ارتکاب رفتاری زیان‌بار شود، پاسخ به این پرسش که مسئولیت متوجه چه شخصی است، به‌سادگی امکان‌پذیر نیست. آیا برنامه‌نویس مسئول است؟ آیا بهره‌بردار یا کاربر باید پاسخ‌گو باشد؟ آیا مسئولیت متوجه شرکت مالک یا توسعه‌دهنده سامانه است یا می‌توان خود سامانه هوش مصنوعی را، به‌دلیل استقلال نسبی در تصمیم‌گیری، موضوع مسئولیت حقوقی یا کیفری قرار داد؟ تلاش برای تطبیق این سامانه‌ها با چارچوب سنتی عنصر معنوی جرم، افزون بر دشواری‌های نظری، در عمل نیز با خلأهای جدی در انتساب مسئولیت مواجه است و در بسیاری از موارد به ایجاد شکاف‌های پاسخ‌گویی می‌انجامد.

بر این اساس، مقاله حاضر از ضرورت گذار آگاهانه از الگوی «کنشگر عقلانی» به الگوی «الگوریتم عقلانی» دفاع می‌کند. فرض اساسی پژوهش آن است که حفظ کارآمدی و اعتبار حقوق کیفری در عصر هوش مصنوعی، مستلزم بازاندیشی بنیادین در مفهوم اهلیت حقوقی است؛ به‌گونه‌ای که این مفهوم دیگر صرفاً بر ویژگی‌های زیستی و روان‌شناختی انسان مبتنی نباشد، بلکه ظرفیت‌های عملکردی عامل‌های خودمختار را نیز در بر گیرد. پرسش اصلی پژوهش آن است که آیا می‌توان نظریه‌ای منسجم، عادلانه و عملی درباره اهلیت حقوقی و مسئولیت عامل‌های خودمختار هوش مصنوعی ارائه کرد که مبنای آن نه شباهت وجودشناختی این سامانه‌ها با انسان، بلکه ویژگی‌هایی همچون استقلال عملکردی، عقلانیت کارکردی، قابلیت یادگیری، پیش‌بینی‌پذیری رفتاری و میزان تأثیرگذاری آن‌ها بر محیط باشد؟

در راستای پاسخ به این پرسش، ابتدا مبانی نظری شخصیت و اهلیت حقوقی در نظام‌های حقوقی بررسی خواهد شد، سپس ویژگی‌های عملکردی هوش مصنوعی که می‌توانند مبنای شکل‌گیری نوعی «اهلیت هوش مصنوعی» قرار گیرند، مورد تحلیل قرار می‌گیرد و در نهایت،

دیدگاه‌های موجود درباره ایجاد نظامی تدریجی از «شخصیت حقوقی الکترونیکی» و سطوح مختلف مسئولیت برای عامل‌های خودمختار ارزیابی خواهد شد. هدف این پژوهش اعطای بی‌قیدوشرط حقوق به ماشین‌ها یا پذیرش فوری شخصیت حقوقی آن‌ها نیست؛ بلکه آغاز گفت‌وگویی علمی درباره ضرورت تحول مفاهیم بنیادین حقوق در مواجهه با نسل جدیدی از عامل‌های هوشمند است که به‌طور فزاینده در تصمیم‌گیری‌ها و روابط اجتماعی ایفای نقش می‌کنند.

اهلیت در حقوق کیفری

گفته شده است که «ملموس‌ترین» ویژگی شخصیت حقوقی در حقوق کیفری، قابلیت تحمل مجازات است و این‌که «تنها زمانی که شرایط مجرم شناخته شدن یک شخص در جامعه به‌طور روشن مشخص باشد، می‌توان تعیین کرد که تحت چه شرایطی ربات‌ها نیز در آینده مجرم شناخته خواهند شد». ازاین‌رو، این فصل به‌طور ویژه بر مجموعه‌ای از قابلیت‌ها و توانایی‌هایی تمرکز دارد که برای تشخیص این امر ضروری هستند که آیا یک شخص یا موجود، مخاطب مناسبی برای مجازات کیفری محسوب می‌شود یا خیر؛ به‌ویژه قابلیت‌هایی که در احراز قابلیت سرزنش او نقش دارند و از آن‌ها با عنوان «اهلیت کیفری» یاد می‌شود.

برای روشن شدن بحث، لازم است ابتدا به مبانی مفهوم اهلیت بازگردیم. اهلیت معمولاً یکی از آثار شخصیت حقوقی به شمار می‌آید. در واقع، شخصیت حقوقی به‌عنوان مجموعه‌ای از ویژگی‌ها و آثار حقوقی توصیف شده است که برخی از آن‌ها جنبه انفعالی و برخی دیگر جنبه فعال دارند. به اعتقاد کورکی، آثار فعال شخصیت حقوقی را می‌توان به دو دسته تقسیم کرد: نخست، تکالیف حقوقی که اجرای آن‌ها با ضمانت اجرای قانونی همراه است و دوم، صلاحیت‌ها و اختیارات حقوقی. مسئولیت کیفری از مصادیق دسته نخست محسوب می‌شود. در مقابل، گروهی دیگر مسئولیت کیفری را یکی از آثار فعال شخصیت حقوقی می‌دانند، اما آن را در قالب «مسئولیت مبتنی بر اهلیت» تحلیل می‌کنند. مسئولیت مبتنی بر اهلیت به این پرسش می‌پردازد که آیا شخص می‌تواند به سبب رفتار خود، مشمول مسئولیت کیفری یا مدنی و ضمانت اجراهای قانونی قرار گیرد یا خیر. بر اساس این دیدگاه، مسئولیت کیفری مستلزم برخورداری از توانایی انجام رفتاری است که قانون، مسئولیت را بر آن مترتب می‌کند.

انتساب، به معنای نسبت دادن یک عمل یا نتیجه به منشأ یا عامل آن است. در حقوق کیفری، انتساب به سازوکاری اطلاق می‌شود که به موجب آن، یک قاعده کیفری شخص معینی را مخاطب آثار و پیامدهای خود قرار می‌دهد. این سازوکار دو کارکرد اساسی دارد: نخست، شخص را به حکم مقرر در قاعده کیفری مرتبط می‌سازد و دوم، وی را به یک واقعه یا وضعیت مشخص منتسب می‌کند. در این خصوص، نویسندگانی مانند گودینیو معتقدند که «قواعد مربوط به مسئولیت کیفری، خود فرایندی برای انتساب مسئولیت ایجاد نمی‌کنند، بلکه بر اصول دادرسی مبتنی هستند که ساختار مسئولیت را شکل می‌دهند و این اصول، پیش‌فرض آن قواعد حقوقی به شمار می‌روند.» در نتیجه، آنچه معمولاً از آن به‌عنوان بخش عمومی حقوق کیفری یاد می‌شود، شامل شرایط عمومی‌ای است که باید فراهم باشند تا یک رفتار به مرتکب آن منتسب شود، مسئولیت وی احراز گردد و در نهایت به سبب آن مجازات شود (Bohlander, 2009).

نخستین شرط از شرایط یادشده آن است که تنها رفتار انسان می‌تواند منشأ مسئولیت کیفری باشد. در اینجا این پرسش مطرح می‌شود که «انسان بودن» در حقوق کیفری دقیقاً به چه معناست و آیا یک سامانه هوش مصنوعی نیز می‌تواند در قلمرو این مفهوم قرار گیرد یا خیر. در این مرحله، بدون ورود به مباحث مربوط به عنصر مادی و عنصر معنوی جرم، تمرکز پژوهش بر تبیین مفهوم «انسان» و به تبع آن، مفهوم «اهلیت کیفری» است؛ یعنی توانایی شخص برای آنکه بتوان او را موضوع مجازات کیفری قرار داد. در همین زمینه گفته شده است که

«شخصیت حقوقی ارتباطی نزدیک با قابلیت سرزنش دارد، زیرا تنها شخصی که بتواند میان درست و نادرست تمایز قائل شود و قدرت انتخاب داشته باشد، در صورت انتخاب رفتار نادرست، شایسته سرزنش است.»

البته باید توجه داشت که نظام‌های حقوقی‌ای که مسئولیت کیفری اشخاص حقوقی را پذیرفته‌اند، تا اندازه‌ای از این شرط فاصله گرفته‌اند. ازاین‌رو، اهلیت کیفری بیانگر فلسفه واقعی حقوق کیفری است؛ یعنی «واکنشی که تنها نسبت به کسانی اعمال می‌شود که می‌توانستند قانون را رعایت کنند، اما آگاهانه از انجام آن خودداری کرده‌اند.» در حقیقت، تعیین شرایط لازم برای آنکه شخص بتواند مخاطب قواعد کیفری قرار گیرد، خود نوعی احترام به آزادی اراده او در انتخاب رفتار درست یا نادرست است.

اهلیت کیفری بر مجموعه‌ای از فرض‌های قانونی استوار است. بر این اساس، اصل بر آن است که هر شخصی که به سن معینی رسیده باشد، از سلامت روان برخوردار است و بنابراین توان تحمل مسئولیت و مجازات کیفری را دارد. این فرض ارتباطی مستقیم با قابلیت سرزنش شخصی دارد که مخاطب قاعده کیفری و ضمانت اجرای مقرر در آن قرار می‌گیرد.

با نگاهی به نظام‌های حقوقی ملی، از جمله نظام‌های حقوق کیفری آلمان و ایتالیا، می‌توان دریافت که در هر دو نظام، قواعد کلی بر شخصی بودن مسئولیت کیفری تأکید دارند. بی‌تردید، اصل تقصیر یکی از بنیادی‌ترین اصول حقوق کیفری آلمان است و اصل شخصی بودن مجازات نیز جایگاهی مشابه در حقوق کیفری ایتالیا دارد. هر دو کشور این اصول بنیادین را در سطح قانون اساسی نیز به رسمیت شناخته‌اند.

نظام حقوق کیفری آلمان که ساختار سه‌مرحله‌ای جرم را پذیرفته است، میان عنصر روانی جرم و قابلیت سرزنش تفاوت قائل می‌شود؛ به بیان دیگر، میان عنصر روانی در مفهوم توصیفی و عنصر روانی در مفهوم هنجاری تمایز برقرار می‌کند. در قانون کیفری آلمان، اهلیت کیفری به‌صراحت در ماده ۲۰ مورد اشاره قرار گرفته است. این ماده که به دفاع مبتنی بر جنون اختصاص دارد، فقدان اهلیت کیفری را از عوامل زایل‌کننده سومین رکن جرم، یعنی قابلیت سرزنش، می‌داند. ماده ۲۰ قانون کیفری آلمان مقرر می‌دارد:

«هر شخصی که هنگام ارتکاب جرم، به علت ابتلا به اختلال روانی مرضی، اختلال شدید در سطح هوشیاری، ناتوانی ذهنی یا هرگونه نابهنجاری شدید روانی دیگر، قادر به درک نامشروع بودن رفتار خود یا عمل کردن بر اساس چنین درکی نباشد، فاقد تقصیر محسوب می‌شود.» (Bonicalzi & Haggard, 2019)

مقررات مربوط به دفاع مبتنی بر جنون در حقوق کیفری آلمان، تنها تا حدودی با معیار تشخیص جنون پذیرفته‌شده در حقوق کیفری ایالات متحده آمریکا شباهت دارد؛ زیرا در حقوق آلمان، احراز جنون بر پایه بررسی دو توانایی اساسی صورت می‌گیرد: نخست، توانایی درک و تشخیص ماهیت و نامشروع بودن رفتار و دوم، توانایی کنترل و هدایت رفتار بر اساس آن درک.

همانند نظام حقوقی آلمان، در حقوق کیفری ایالات متحده آمریکا نیز نهاد مستقلی تحت عنوان «اهلیت کیفری» وجود ندارد. در این نظام، اصل بر آن است که اشخاص بالغ از سلامت روان برخوردارند و بنابراین دارای اهلیت کیفری هستند و معیارهای تشخیص این اهلیت باید از مقررات مربوط به دفاع مبتنی بر جنون استنباط شود. در واقع، همان‌گونه که بیان شده است، دفاعیات کیفری «در تحلیل پیش‌شرط‌های روان‌شناختی مسئولیت در حقوق کیفری، از اهمیت ویژه‌ای برخوردارند.»

رایج‌ترین معیار دفاع مبتنی بر جنون در ایالات متحده آمریکا، مقرر می‌دارد:

«اصل بر این است که هر شخصی عاقل محسوب می‌شود و برای پذیرش دفاع مبتنی بر جنون باید به‌طور روشن اثبات شود که متهم در زمان ارتکاب رفتار، به علت بیماری روانی، چنان دچار اختلال در قوه تعقل بوده است که یا ماهیت و کیفیت رفتار خود را درک نمی‌کرد، یا اگر آن را درک می‌کرد، نمی‌دانست که رفتار او نادرست است.»

نسخه ارائه‌شده در «قانون نمونه کیفری» که تنها در شمار محدودی از ایالت‌ها پذیرفته شده است، از معیار یادشده فاصله گرفته و به مقررات ماده ۲۰ قانون کیفری آلمان نزدیک‌تر است؛ زیرا علاوه بر جنبه شناختی، جنبه ارادی را نیز در بر می‌گیرد. مطابق بند ۱ از ماده ۴/۰۱ این قانون: «شخصی که در زمان ارتکاب رفتار، در نتیجه ابتلا به بیماری یا اختلال روانی، به میزان قابل توجهی توانایی درک مجرمانه یا نادرست بودن رفتار خود یا توانایی انطباق رفتار خویش با الزامات قانون را از دست داده باشد، مسئولیت کیفری نخواهد داشت.»

بااین‌حال، پذیرش دفاع مبتنی بر جنون در حقوق ایالات متحده آمریکا لزوماً به صدور حکم برائت منتهی نمی‌شود. برخی از ایالت‌ها نهادی را با عنوان «مجرم ولی مبتلا به بیماری روانی» پیش‌بینی کرده‌اند که هدف آن تأمین هم‌زمان اهداف اصلاح و بازپروری مجرم و نیز اجرای عدالت کیفری است (Abel et al., 2016).

یکی از مباحث مهم و مورد اختلاف که ارتباط مستقیمی با موضوع حاضر دارد، تفسیر مفهوم «نادرستی» در دفاع مبتنی بر جنون است. این پرسش مطرح می‌شود که آیا مقصود از نادرستی، مغایرت با قانون است یا مغایرت با ارزش‌های اخلاقی. شایان توجه است که بر اساس آخرین رأی دیوان عالی ایالات متحده آمریکا در پرونده «کالر علیه کانزاس»، ایالت‌ها از نظر قانون اساسی مجاز هستند مقرراتی درباره دفاع مبتنی بر جنون وضع کنند که در آن، توانایی متهم برای درک نادرستی اخلاقی رفتار خود در زمان ارتکاب جرم، شرط پذیرش این دفاع نباشد.

در نظام حقوقی ایتالیا، مفهوم اهلیت کیفری به‌طور صریح در ماده ۸۵ قانون مجازات پیش‌بینی شده است. این ماده مقرر می‌دارد:

«هیچ‌کس را نمی‌توان به دلیل رفتاری که قانون آن را جرم شناخته است مجازات کرد، مگر آنکه در زمان ارتکاب آن رفتار، دارای اهلیت کیفری باشد. شخصی دارای اهلیت کیفری است که از توانایی ادراک و اراده برخوردار باشد.»

بر این اساس، اهلیت کیفری در حقوق ایتالیا بر دو رکن اساسی استوار است: رکن ادراکی و رکن ارادی. توانایی ادراک به معنای توانایی فهم ماهیت و آثار رفتار خویش است، در حالی که توانایی اراده به معنای قدرت شخص برای انجام رفتار بر اساس اراده آزاد و کنترل انگیزه‌ها و تمایلات خود است. تحقق هر دو شرط برای احراز اهلیت کیفری ضروری است. همچنین، ماده ۸۸ قانون مجازات ایتالیا آثار فقدان اهلیت کیفری را بیان کرده و مقرر می‌دارد که هرگاه شخص در زمان ارتکاب رفتار، بر اثر بیماری روانی، به‌گونه‌ای باشد که توانایی ادراک یا اراده را از دست داده باشد، مسئولیت کیفری نخواهد داشت.

پس از تبیین مفهوم اهلیت کیفری، باید آن را با نظریه‌های اصلی توجیه‌کننده مجازات مرتبط ساخت. اشخاصی که دارای اهلیت کیفری شناخته می‌شوند، افرادی هستند که از نظر نظریه‌های کیفری، باید از ارتکاب جرایم آینده بازداشته شوند. این امر مبتنی بر این فرض است که آنان توانایی درک دستورهای قانون کیفری و فهم پیامدهای نقض آن را دارند. تنها در چنین صورتی است که مجازات می‌تواند در اصلاح رفتار آنان یا بازداشتنشان از ارتکاب مجدد جرم مؤثر باشد.

از سوی دیگر، مجازات این اشخاص باید نقش بازدارندگی عمومی نیز ایفا کند؛ بدین معنا که سایر اعضای جامعه با مشاهده مجازات مرتکب، خود را در موقعیت او قرار دهند و از ارتکاب رفتار مشابه خودداری کنند. افزون بر این، بر اساس نظریه سزادهی، مجازات زمانی توجیه‌پذیر

است که شخص با اراده آزاد مرتکب جرم شده باشد؛ زیرا تنها در این صورت است که مجازات می‌تواند پاسخ متناسبی به بی‌عدالتی ناشی از ارتکاب جرم تلقی شود.

در نهایت، با توجه به مطالب پیش‌گفته، می‌توان نتیجه گرفت که برای آنکه شخصی مخاطب قواعد حقوق کیفری قرار گیرد، باید از دو قابلیت اساسی برخوردار باشد: نخست، توانایی درک ماهیت و آثار رفتار خود و دوم، توانایی کنترل رفتار و هدایت اعمال خویش بر اساس این درک. در ادامه، هر یک از این دو مؤلفه به‌ترتیب مورد بررسی قرار خواهند گرفت (Palazzo & Papa, 2013).

۲. ماشین‌های مجرم یا ماشین‌های اخلاق‌مدار؟

اکنون باید نخستین شرط اهلیت کیفری، یعنی توانایی درک ماهیت رفتار، مورد بررسی قرار گیرد. پیش از هر چیز، لازم است روشن شود که مقصود از «ماهیت رفتار» چیست. این مفهوم در نظام‌های حقوقی معاصر معانی متفاوتی دارد. در حقوق آلمان، اهلیت کیفری به‌طور مستقیم با توانایی شخص در درک نامشروع بودن رفتار خود ارتباط دارد. در ایالات متحده آمریکا، معیار مشهور تشخیص جنون هم به ماهیت و کیفیت رفتار و هم به نادرست بودن آن اشاره می‌کند، در حالی که قانون نمونه کیفری، آگاهانه از ارائه تعریف قطعی در این خصوص خودداری کرده است. در حقوق ایتالیا نیز تأکید اصلی بر توانایی شخص در درک پیامدهای اجتماعی رفتار خویش قرار دارد.

اگر اهلیت کیفری صرفاً بر پایه توانایی درک «ماهیت فیزیکی رفتار» تعریف شود؛ برای مثال، اینکه شخص بداند نزدیک کردن شعله آتش به یک ساختمان موجب سوختن آن می‌شود یا فرو بردن سر فردی در آب به مرگ او می‌انجامد، در این صورت عامل‌های هوش مصنوعی بدون دشواری می‌توانند شرایط لازم برای برخورداری از اهلیت کیفری را احراز کنند. به بیان دیگر، سامانه‌های هوش مصنوعی قادرند شناخت دقیقی از واقعیت‌های مرتبط با رفتار به دست آورند. در حقیقت، توانایی این سامانه‌ها در پیش‌بینی احتمال وقوع پیامدهای فیزیکی ناشی از یک رفتار و ایجاد مدل‌های تحلیلی، به دلیل دسترسی به حجم عظیمی از داده‌ها و قدرت استنباط بسیار بالا، از توانایی انسان فراتر است. تردیدی نیست که یک سامانه هوش مصنوعی می‌تواند در مدت زمانی بسیار کوتاه، دانشی درباره قواعد حاکم بر جهان فیزیکی کسب کند که دستیابی انسان به آن حتی در طول یک عمر نیز ممکن نباشد.

اما اگر مقصود از ماهیت رفتار، درک نادرستی آن باشد، در این صورت باید روشن شود که منظور از نادرستی، مغایرت با قانون است یا مغایرت با اخلاق. همان‌گونه که در بررسی حقوق ایالات متحده آمریکا اشاره شد، این مسئله حتی درباره انسان نیز همچنان محل اختلاف است. بنابراین، این پرسش مطرح می‌شود که آیا یک عامل هوش مصنوعی برای آنکه مخاطب قواعد کیفری قرار گیرد، باید بداند که مرتکب یک رفتار غیراخلاقی شده است یا آنکه صرف آگاهی از غیرقانونی بودن رفتار کافی است؟ برای پاسخ به این پرسش، ابتدا باید بررسی کرد که آیا اصولاً عامل‌های هوش مصنوعی توانایی یادگیری قواعد اخلاقی یا هنجارهای حقوقی را دارند یا خیر.

در خصوص هنجارهای حقوقی، می‌توان تصور کرد که الزام به رعایت قانون در ساختار یک سامانه هوش مصنوعی تعبیه شود. برخی از پژوهشگران میان معماری سامانه‌های هوش مصنوعی و قواعد و استانداردهای حقوقی نوعی شباهت قائل شده‌اند. بر این اساس، امکان طراحی مجموعه‌ای از قوانین ویژه ماشین‌ها مطرح شده است؛ مجموعه‌ای که قواعد حقوقی را به زبانی قابل فهم و اجرا برای سامانه‌های هوش مصنوعی تبدیل می‌کند. از نظر نظری، حتی می‌توان انتظار داشت که سامانه‌های هوش مصنوعی در رعایت قانون از انسان موفق‌تر باشند؛ زیرا قادرند حجم بسیار گسترده‌ای از مقررات، حتی تمامی قوانین کیفری یک کشور، را در حافظه خود نگهداری کنند و هرگز آن‌ها را فراموش

نکنند. افزون بر این، همان‌گونه که در ادامه خواهد آمد، می‌توان آثار بازدارنده مجازات را نیز در طراحی این سامانه‌ها پیش‌بینی و به‌گونه‌ای برنامه‌ریزی کرد که در محاسبات هزینه و فایده آن‌ها مورد توجه قرار گیرد.

با وجود این، سامانه‌های کنونی هوش مصنوعی هنوز قادر نیستند به‌صورت مستقل هنجارهای حقوقی را به‌عنوان «قاعده‌ای الزام‌آور» درک کنند؛ هرچند می‌توان گفت که برای مثال، خودروهای خودران هم‌اکنون توانایی شناسایی مستقیم برخی هنجارهای نمادین، مانند علائم راهنمایی و رانندگی، را دارند. بااین‌حال، تا امروز تمامی سامانه‌های هوش مصنوعی برای درک الزام‌آور بودن قواعد حقوقی، نیازمند مداخله و برنامه‌ریزی انسان هستند؛ به بیان دیگر، این انسان است که باید به آن‌ها بیاموزد رعایت قانون الزامی است (Rengier, 2019).

افزون بر این، به نظر می‌رسد که اصول اخلاقی نیز تا حدی قابلیت تبدیل شدن به دستورهای قابل اجرا در سامانه‌های هوش مصنوعی را دارند. در واقع، تبدیل ارزش‌های اخلاقی به قواعد محاسباتی موضوع تازه‌ای نیست. مشهورترین نمونه آن، سه قانون رباتیک است که عبارت‌اند از:

قانون نخست: ربات نباید به انسان آسیب برساند یا با خودداری از اقدام، موجب آسیب دیدن انسان شود.

قانون دوم: ربات باید از دستورهای انسان اطاعت کند، مگر آنکه اجرای آن دستورها با قانون نخست تعارض داشته باشد.

قانون سوم: ربات باید از موجودیت خود محافظت کند، مشروط بر آنکه این امر با قانون نخست یا دوم در تعارض نباشد.

بااین‌حال، پرسش اساسی این است که آیا واقعاً می‌توان قواعد اخلاقی را به سامانه‌های هوش مصنوعی آموزش داد؟ آزمایش ذهنی مطرح‌شده پیش‌تر، نمونه‌ای از مشکلاتی است که در این زمینه ممکن است پدید آید. این مسئله در واقع نمونه‌ای از «مسئله همسوسازی» است؛ یعنی چالش هماهنگ کردن ارزش‌های هوش مصنوعی با ارزش‌های انسانی. این مشکل را می‌توان چنین توضیح داد:

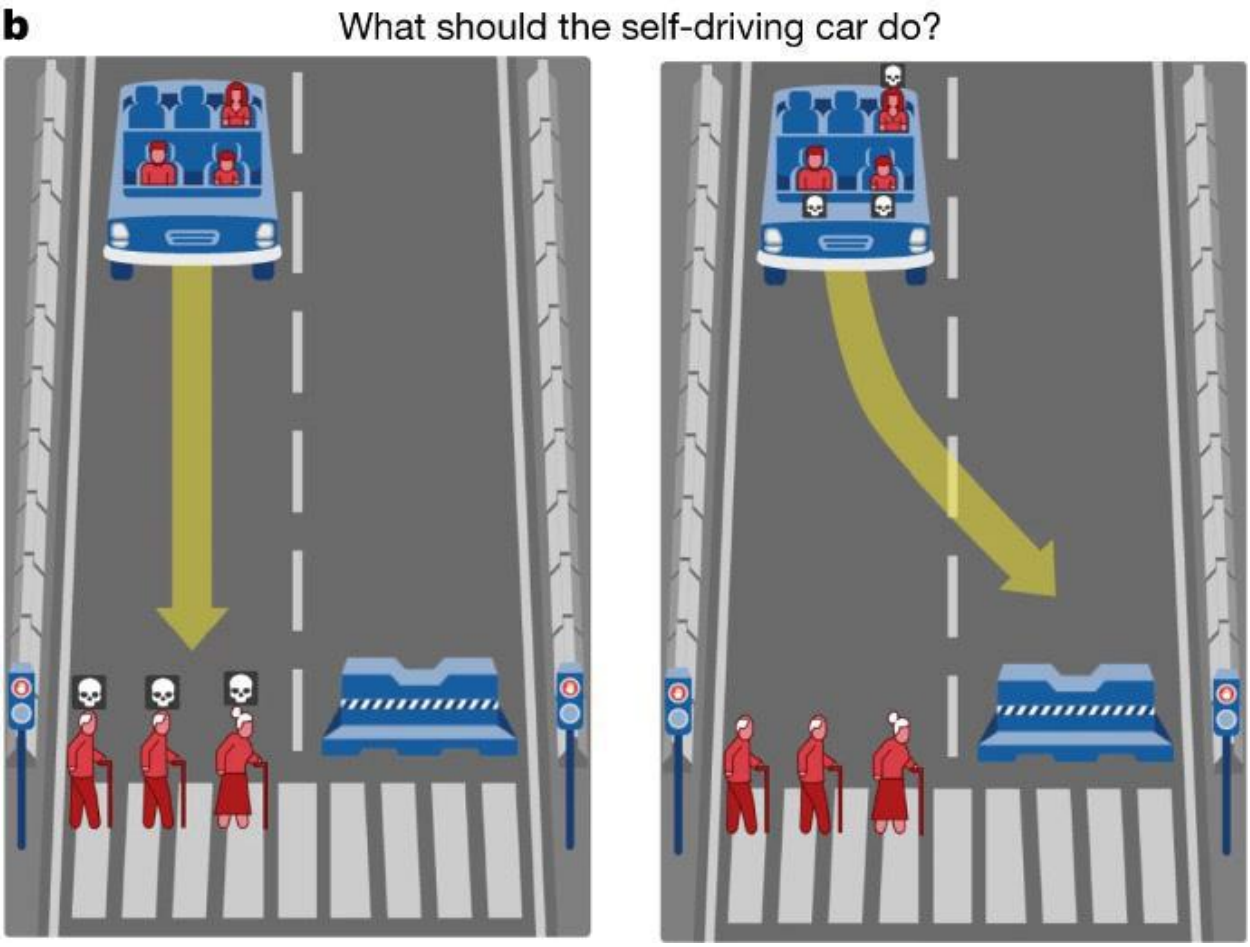
هر قاعده یا هدف، قابلیت تفسیر به شیوه‌های بی‌شمار و گاه نامحدود را دارد. در حوزه هوش مصنوعی، این امر ممکن است موجب شود که یک سامانه بسیار پیشرفته، برداشت‌ها و تفسیرهای جدیدی از قواعدی که به آن داده شده است ارائه کند؛ تفسیرهایی که از نظر منطقی نادرست نیستند و حتی می‌توان آن‌ها را برداشت‌هایی معتبر از همان قواعد دانست، اما از آنجا که با مقاصد و ارزش‌های انسانی سازگار نیستند، مورد پذیرش انسان قرار نمی‌گیرند.

پژوهش درباره طراحی سامانه‌های هوش مصنوعی اخلاق‌مدار هم‌اکنون آغاز شده و در سال‌های اخیر رشد چشمگیری داشته است. بخش مهمی از این پژوهش‌ها به توسعه «عامل‌های اخلاقی مصنوعی» اختصاص یافته است؛ یعنی سامانه‌هایی که اصول اخلاقی و توانایی تصمیم‌گیری اخلاقی در آن‌ها پیاده‌سازی می‌شود تا رفتارشان در برابر انسان‌ها و سایر سامانه‌ها از نظر اخلاقی قابل قبول باشد. هدف از طراحی چنین سامانه‌هایی آن است که بتوانند در موقعیت‌های اخلاقی دشوار، غیرمنتظره یا مبهم، در تصمیم‌گیری به انسان یاری رسانند یا حتی در برخی موارد جایگزین قضاوت انسانی شوند.

ماشین اخلاقی مؤسسه فناوری ماساچوست

یکی از مشهورترین نمونه‌های پژوهش در زمینه اخلاق هوش مصنوعی، آزمایش «ماشین اخلاقی» است که توسط مؤسسه فناوری ماساچوست طراحی شده است. این آزمایش در قالب یک بازی اینترنتی اجرا می‌شود که در آن کاربران با یک موقعیت مبتنی بر دوراهی اخلاقی مواجه می‌شوند. در این سناریو، شرکت‌کنندگان با یک حادثه اجتناب‌ناپذیر روبه‌رو هستند و باید میان دو گزینه تصمیم بگیرند: ادامه مسیر یا تغییر مسیر. هر یک از این دو تصمیم، به مرگ تعداد معینی از افراد منجر می‌شود (شکل ۱) (Lima, 2018).

طراحان این آزمایش با تحلیل انتخاب‌های کاربران میان «دو گزینه نامطلوب» و بررسی اینکه کدام نتیجه از دیدگاه آنان ترجیح دارد، و همچنین با استفاده از اطلاعات جمعیت‌شناختی و موقعیت جغرافیایی شرکت‌کنندگان، موفق شدند سه الگوی اصلی را شناسایی کنند که به اعتقاد آنان، بازتابی از اصول نسبتاً جهان‌شمول اخلاق قابل استفاده برای ماشین‌ها است. این سه الگو عبارت‌اند از: ترجیح نجات جان انسان‌ها، ترجیح نجات تعداد بیشتری از افراد، و ترجیح نجات افراد جوان‌تر. برای نمونه، نتایج این پژوهش نشان می‌دهد که میان شهروندان آلمان و ایالات متحده آمریکا در نحوه انتخاب میان این گزینه‌های اخلاقی تفاوت‌هایی وجود دارد که در شکل ۸ گزارش شده است.



شکل ۱. نمونه‌ای از پرسش مطرح‌شده در آزمایش «ماشین اخلاقی» مؤسسه فناوری ماساچوست. منبع: (Awad, 2022).

نقد و ارزیابی

پژوهش‌هایی مانند پروژه «ماشین اخلاقی» مؤسسه فناوری ماساچوست و سایر تحقیقات مشابه، با چالش‌های متعددی روبه‌رو هستند. نخست آنکه برای آموزش یک قاعده اخلاقی به یک سامانه هوش مصنوعی، بر اساس رویکردی که قواعد از پیش تعیین‌شده را به سامانه تحمیل می‌کند، لازم است آن قاعده به زبانی رسمی و قابل فهم برای سامانه تبدیل شود. تبدیل تصمیم‌های اخلاقی به مجموعه‌ای از قواعد محاسباتی، فرایندی بسیار پیچیده و دشوار است؛ زیرا قواعد اخلاقی ذاتاً مبهم هستند و به همین دلیل، تبدیل آن‌ها به دستورهای دقیق و قابل اجرا برای طراحی سامانه‌ها و الگوریتم‌ها آسان نیست.

همان‌گونه که بیان شده است:

«این خطر وجود دارد که اگر هوش مصنوعی با دقت و بر مبنای ارزش‌های انسانی طراحی نشود، پیامدهای فاجعه‌باری برای بشریت به همراه داشته باشد. برای مثال، اگر در طراحی سامانه‌های هوشمند، ارزش‌های انسانی مورد توجه قرار نگیرد، این سامانه‌ها ممکن است تصمیم‌هایی اتخاذ کنند که به انسان آسیب برساند. همچنین، هرچه توانایی‌های هوش مصنوعی با سرعت بیشتری افزایش یابد، اهمیت و حساسیت مسئله همسوسازی ارزش‌های هوش مصنوعی با ارزش‌های انسانی نیز بیشتر خواهد شد.» (MacCormick, 2008)

اگر برای لحظه‌ای بر موضوع نهادینه کردن احترام به قواعد حقوقی در سامانه‌های هوش مصنوعی تمرکز کنیم، مشاهده می‌شود که مشکلات مشابهی درباره مفاهیم حقوقی مورد استفاده در قوانین کیفری نیز وجود دارد. برای نمونه، مفهوم «زیان» دارای قلمرو بسیار گسترده‌ای است؛ به‌گونه‌ای که ممکن است شامل آسیب جسمانی واقعی، خطر ورود آسیب احتمالی، خسارت مالی، آسیب روانی یا حتی زیان حقوقی ناشی از قرار دادن مالک یک سامانه هوش مصنوعی در معرض مسئولیت قانونی باشد. بنابراین، هنگامی که حتی میان انسان‌ها نیز درباره حدود و قلمرو چنین مفاهیمی توافق کامل وجود ندارد، این پرسش مطرح می‌شود که چگونه می‌توان چنین مفاهیمی را به‌گونه‌ای دقیق در یک سامانه هوش مصنوعی برنامه‌ریزی کرد.

هرچند قوانین کیفری باید از بالاترین سطح صراحت و قطعیت برخوردار باشند، اما این قوانین نیز همواره درجه‌ای از ابهام ذاتی را حفظ می‌کنند. در بسیاری از نظام‌های حقوقی، قضات وظیفه دارند با تفسیر مفاهیم قانونی، خلأهای معنایی موجود در قوانین را برطرف کنند و این مفاهیم را متناسب با تحولات اجتماعی تفسیر نمایند. اگر قرار باشد سامانه‌های هوش مصنوعی نیز بر همین مبنا عمل کنند، ساختار آن‌ها باید به‌طور مستمر به‌روزرسانی شود تا بتواند خود را با آخرین تفسیرهای قضایی و تحولات حقوقی هماهنگ سازد.

نکته مهم آن است که مشکلات مربوط به همسوسازی ارزش‌ها و تفسیر قواعد، تنها به سامانه‌های هوش مصنوعی عمومی محدود نمی‌شود، بلکه در سامانه‌های تخصصی نیز مشاهده می‌شود. برای مثال، سامانه هوش مصنوعی «لیبراتوس» که در سال ۲۰۱۷ توسط پژوهشگران دانشگاه کارنگی ملون برای انجام بازی پوکر طراحی شد، توانست شیوه‌هایی مانند بلوف زدن و بازی تهاجمی را بیاموزد و در نهایت قهرمانان جهان را در یکی از پیچیده‌ترین انواع بازی پوکر شکست دهد. اهمیت این دستاورد در آن است که نشان می‌دهد سامانه‌های هوش مصنوعی می‌توانند در شرایطی که اطلاعات ناقص و نامطمئن است نیز تصمیم‌های راهبردی موفقی اتخاذ کنند. همین الگوریتم قابلیت آن را دارد که در سایر موقعیت‌هایی که انسان ناچار است بر پایه اطلاعات ناقص تصمیم‌گیری کند نیز مورد استفاده قرار گیرد. بدیهی است هرچه چنین سامانه‌هایی در حوزه‌های واقعی زندگی به کار گرفته شوند، میزان تأثیرگذاری آن‌ها بر جهان خارج و در نتیجه احتمال بروز خسارت نیز افزایش خواهد یافت.

بازگشت به بحث هوش مصنوعی اخلاق‌مدار نشان می‌دهد که پژوهشگران همچنان با چالش بازآفرینی انعطاف‌پذیری ذهن اخلاقی انسان مواجه هستند. در این زمینه گفته شده است:

«انسان می‌تواند درباره موقعیت‌هایی که هرگز پیش از آن با آن‌ها مواجه نشده است، داوری اخلاقی انجام دهد. او می‌تواند تشخیص دهد که در برخی شرایط باید از قواعد از پیش تعیین‌شده عدول کند و حتی در موقعیت‌های جدید، قواعد اخلاقی تازه‌ای را شکل دهد. بازآفرینی چنین سطحی از انعطاف‌پذیری، یکی از اساسی‌ترین چالش‌های توسعه سامانه‌های هوش مصنوعی است که بتوانند همانند انسان درباره مسائل اخلاقی قضاوت و تصمیم‌گیری کنند.» (Simmler & Markwalder, 2019)

افزون بر این، تاکنون هیچ معیار قابل اعتمادی برای سنجش میزان «اخلاقی بودن» یک سامانه هوش مصنوعی ارائه نشده است. برخی پژوهشگران حتی معتقدند طراحی چنین معیاری اساساً امکان‌پذیر نیست؛ زیرا هنوز توافقی درباره بهترین یا کامل‌ترین نظام اخلاقی وجود ندارد. افزون بر آن، در هر فرایند تبدیل ارزش‌های اخلاقی به قواعد قابل اجرا، همواره یک انسان حضور دارد که ارزش‌ها، باورها، پیش‌فرض‌ها و حتی کلیشه‌های ذهنی خود را در این فرایند دخالت می‌دهد.

از منظر سیاست‌گذاری و تنظیم‌گری نیز برخی صاحب‌نظران معتقدند نباید پیشرفت فناوری را تا زمان حل کامل مباحث فلسفی اخلاق متوقف کرد؛ زیرا نمی‌توان انتظار داشت که تدوین قواعد حاکم بر یکی از مهم‌ترین حوزه‌های پیشرفت فناوری، تا زمانی که فیلسوفان تمامی مسائل اخلاقی فرضی را حل کنند، به تعویق افتد. افزون بر این، می‌توان استدلال کرد که مشروط کردن اهلیت کیفری به توانایی مرتکب در تشخیص غیراخلاقی بودن رفتار خود، ضرورتی ندارد. همین استدلال نیز مبنای رأی دیوان عالی ایالات متحده آمریکا در پرونده «کالر علیه کانزاس» قرار گرفته است.

همچنین، حتی اگر فرض کنیم برای همه مسائل اخلاقی پاسخ صحیحی وجود داشته باشد، باز هم اگر یک سامانه هوش مصنوعی در موقعیتی جدید و پیش‌بینی‌نشده، تصمیمی اتخاذ کند که از دیدگاه طراحان آن «از نظر اخلاقی نادرست» باشد، دادگاه‌ها یا قانون‌گذاران می‌توانند همین امر را مبنایی برای ایجاد مسئولیت حقوقی تلقی کنند (Kirchner, 2024).

در نهایت، این پرسش که در چنین وضعیتی مسئولیت باید متوجه کدام انسان باشد، اعم از برنامه‌نویس، طراح، تولیدکننده یا کاربر، موضوعی است که در بخش‌های بعدی بررسی خواهد شد. آنچه در شرایط کنونی اهمیت دارد این است که تفکیک مسئولیت میان «خالق» و «مخلوق» هنوز با دشواری‌های فراوانی همراه است؛ و شاید زمانی که چنین تفکیکی از نظر فنی امکان‌پذیر شود، مسئولیت کیفری دیگر مهم‌ترین دغدغه نظام‌های حقوقی در برابر هوش مصنوعی نخواهد بود.

کنترل رفتار

دو مؤلفه اهلیت کیفری ارتباطی جدایی‌ناپذیر با یکدیگر دارند. برای آنکه شخصی بتواند مخاطب قواعد حقوق کیفری قرار گیرد، باید علاوه بر توانایی درک رفتار خود و تشخیص مغایرت آن با قانون، توانایی کنترل رفتار خویش و خودداری از ارتکاب آن را نیز داشته باشد. به بیان دیگر، متهم باید بتواند با اراده خود قانون را نقض کند یا از نقض آن خودداری ورزد. از این منظر، اهلیت کیفری مستلزم آن است که شخص از آزادی لازم برای ارتکاب یا عدم ارتکاب جرم برخوردار باشد.

کنترل رفتار را می‌توان توانایی شخص در هدایت و تنظیم حرکات و اعمال خود، انجام رفتار در راستای اهداف مورد نظر یا خودداری از اقدام در مواقع لازم تعریف کرد. تحقق این مؤلفه، مبتنی بر وجود مؤلفه شناختی است؛ زیرا تنها کسی که توانایی درک ماهیت رفتار خود را دارد، می‌تواند رفتار خویش را بر اساس آن شناخت کنترل کند (Fernandes, 2022).

به اعتقاد برخی نویسندگان، مهم‌ترین ایرادی که به پذیرش اراده آزاد برای عامل‌های هوش مصنوعی وارد می‌شود، این است که آن‌ها صرفاً ماشین‌هایی برنامه‌ریزی‌شده هستند. بااین‌حال، نباید صرفاً به این دلیل که انسان برنامه‌ریزی نشده و سامانه‌های هوش مصنوعی بر پایه برنامه‌ریزی عمل می‌کنند، میان این دو تفاوتی مطلق قائل شد. همان‌گونه که پیش‌تر نیز اشاره شد، سامانه‌های هوش مصنوعی در آینده ممکن است به همان اندازه مستقل به نظر برسند که انسان‌ها مستقل تلقی می‌شوند؛ در حالی که تصمیم‌های به‌ظاهر آزاد انسان نیز همواره تحت تأثیر مجموعه‌ای از عوامل پنهان و خارج از اراده او شکل می‌گیرد.

نکته مسلم آن است که اراده آزاد از منظر زیست‌شناختی یا فیزیکی قابل اثبات نیست و وجود آن را نمی‌توان به‌طور عینی نشان داد. در واقع، اراده آزاد مفهومی است که جامعه آن را به اشخاص نسبت می‌دهد و از این حیث، نوعی فرض یا ساختار حقوقی محسوب می‌شود. بحث درباره اراده آزاد از موضوعات مهم فلسفه اخلاق است. گروهی از اندیشمندان که به تردید در وجود اراده آزاد باور دارند، معتقدند رفتارها و ویژگی‌های انسان در نهایت محصول عواملی هستند که خارج از کنترل او قرار دارند و به همین دلیل، انسان هرگز مسئولیت اخلاقی واقعی نسبت به اعمال خود ندارد. در مقابل، گروهی دیگر بر این باورند که تأثیرپذیری انسان از عوامل بیرونی، منافاتی با این ندارد که او در نهایت عامل اصلی رفتارهای خویش محسوب شود (Keiler & Roef, 2019).

افزون بر این، برخی نویسندگان معتقدند که مسئولیت انسان دقیقاً از آن رو معنا پیدا می‌کند که رفتار او تحت تأثیر عوامل گوناگون قرار دارد. در غیر این صورت، راهبردهای پیشگیرانه‌ای که نظام‌های حقوقی معاصر بر آن‌ها استوارند، توجیه خود را از دست خواهند داد. این راهبردها از یک سو بر تهدید به اعمال مجازات در صورت نقض قانون و از سوی دیگر بر نقش آموزشی، هنجارساز و انگیزشی اجرای مجازات در جامعه استوار هستند. اگر تصمیم‌های انسان تحت تأثیر تهدیدها، تشویق‌ها و سایر عوامل محیطی قرار نمی‌گرفت، اصولاً پیشگیری کیفری نمی‌توانست کارآمد باشد.

همان‌گونه که در ادامه این فصل با تفصیل بیشتری بیان خواهد شد، همین استدلال می‌تواند مبنای نسبتاً محکمی برای پذیرش نوعی نظام واکنش کیفری نسبت به سامانه‌های هوش مصنوعی نیز فراهم آورد.

از سوی دیگر، این موضوع نیز محل اختلاف است که آیا توانایی اراده و کنترل رفتار، اساساً مستلزم برخورداری از توانایی انتخاب اخلاقی است یا خیر. برخی معتقدند حتی اگر بپذیریم انسان نیز همانند یک ماشین تا حد زیادی تحت تأثیر عوامل از پیش تعیین‌شده قرار دارد، این امر تأثیری بر مسئولیت کیفری او نخواهد داشت؛ زیرا اگر قابلیت سرزنش را مفهومی هنجاری بدانیم، وجود یا عدم وجود اراده اخلاقی اهمیت تعیین‌کننده‌ای نخواهد داشت. بر اساس این دیدگاه، حتی اشخاصی که فاقد درک اخلاقی هستند نیز ممکن است مسئول شناخته شوند؛ زیرا توانایی اراده چیزی جز برخورداری از وضعیت روانی متعارفی نیست که شخص را قادر می‌سازد در برابر انگیزه‌های ارتکاب جرم مقاومت کند.

از سوی دیگر، یکی از ویژگی‌های اساسی هر عامل، برخورداری از تجربه ذهنی انجام رفتار است؛ یعنی اینکه شخص احساس کند خود منشأ و انجام‌دهنده رفتار خویش است. به بیان دیگر، تجربه‌هایی همچون انتخاب کردن، تصمیم گرفتن یا آغاز یک حرکت ارادی، همواره با نوعی تجربه ذهنی ویژه همراه هستند؛ تجربه‌ای که در رفتارهای انعکاسی وجود ندارد و در رفتارهای عادت‌گونه نیز بسیار ضعیف‌تر است. از دیدگاه برخی دانشمندان علوم اعصاب، تنها عاملی می‌تواند مخاطب واقعی قواعد حقوقی قرار گیرد که نوعی احساس انتساب رفتار به خود داشته باشد؛ یعنی بتواند خود را علت رفتارهایش بداند و رابطه میان حرکت ارادی و پیامدهای آن را درک کند. تنها در این صورت است که شخص می‌تواند ارتباط میان رفتار و نتایج آن را بیاموزد و بر اساس آن، رفتارهای آینده خود را تکرار یا اصلاح کند (Bohlander, 2009).

آیا سامانه‌های هوش مصنوعی نیز می‌توانند چنین احساسی نسبت به رفتارهای خود داشته باشند؟ دست‌کم با توجه به فناوری‌های کنونی، پاسخ این پرسش منفی است. بااین‌حال، شاید اصل پرسش به‌درستی مطرح نشده باشد؛ زیرا در اغلب نظام‌های حقوقی، برخورداری از چنین تجربه ذهنی، شرط لازم برای داشتن اهلیت کیفری محسوب نمی‌شود. آنچه اهمیت دارد، بعد بیرونی رفتار و نحوه ارزیابی آن از منظر نظام

حقوقی است، نه احساس درونی عامل نسبت به رفتار خود. از همین رو، مبانی سنتی مسئولیت کیفری سال‌هاست که نخست با یافته‌های علوم اعصاب و اکنون با ظهور هوش مصنوعی، با چالش‌های جدی روبه‌رو شده‌اند.

درباره بزهکاران فاقد سلامت روان و اطفال بزهکار؛ الگویی برای بررسی مسئولیت سامانه‌های هوش مصنوعی

حقوق کیفری حتی در مورد اشخاصی که فاقد اهلیت کیفری هستند نیز کاملاً بی‌تفاوت نیست و همچنان به نوعی نسبت به رفتار آنان واکنش نشان می‌دهد. همان‌گونه که دیامانتیس و همکارانش بیان کرده‌اند، این مسئله که «الگوریتم‌ها می‌توانند موجب ورود زیان شوند، اما خود مسئول شناخته نمی‌شوند، پدیده‌ای منحصر به الگوریتم‌ها نیست.»

در قرون وسطی و حتی تا اوایل سده هفدهم میلادی ــ که از آن به‌عنوان دوره‌ای نسبتاً روشن‌اندیش یاد شده است ــ و در برخی موارد تا حدود سال ۱۹۰۰، حیوانات نیز به‌عنوان مرتکب جرم تحت تعقیب کیفری قرار می‌گرفتند، برای آن‌ها محاکمه برگزار می‌شد و در نهایت مجازات می‌شدند. امروزه نیز در تعداد محدودی از نظام‌های حقوقی، سگ‌هایی که خطرناک تشخیص داده می‌شوند و به دیگران آسیب رسانده‌اند، پس از برگزاری جلسه رسیدگی و احراز خطرناک بودن آن‌ها توسط مرجع صلاحیت‌دار، ممکن است معدوم شوند.

از سوی دیگر، کودکان و اشخاص فاقد سلامت روان، هرچند ممکن است از نظر حقوقی فاقد اهلیت کیفری برای ارتکاب جرم شناخته شوند، اما در صورتی که برای خود یا جامعه خطرناک تشخیص داده شوند، همچنان ممکن است از سوی دولت تحت نگهداری یا محدودیت قرار گیرند. برای نمونه، اشخاص مبتلا به اختلالات شدید روانی ممکن است تا زمان بازیابی سلامت روان یا رفع خطر، در مراکز درمانی روان‌پزشکی نگهداری شوند. همچنین، کودکان بزهکار ممکن است به مراکز اصلاح و تربیت فرستاده شوند، در دادگاه‌های ویژه اطفال به‌عنوان بزهکار تحت رسیدگی قرار گیرند یا در برخی موارد استثنایی و بر اساس مقررات انتقال صلاحیت، رسیدگی به جرم آنان از دادگاه اطفال به دادگاه کیفری بزرگسالان واگذار شود (Bonicalzi & Haggard, 2019).

نکته اساسی آن است که حقوق همواره می‌کوشد نوعی کنترل بر رفتار این اشخاص اعمال کند و در وهله نخست، از وقوع آسیب‌های احتمالی جلوگیری نماید. با وجود این، نه کودکان و نه حیوانات از نظر حقوقی مسئول اعمال خود محسوب نمی‌شوند و حتی اگر امکان طرح دعوا علیه آن‌ها وجود داشت، توانایی تحمل آثار مسئولیت را نداشتند. به بیان دیگر، این اشخاص به دلیل فقدان اهلیت کیفری، شخصیت حقوقی خود را از منظر حقوق کیفری از دست نمی‌دهند.

این وضعیت می‌تواند در بررسی جایگاه سامانه‌های هوش مصنوعی نیز آموزنده باشد. بدین معنا که شاید برای اعمال برخی اقدامات محدودکننده نسبت به سامانه‌های هوش مصنوعی، ضرورتی نداشته باشد که ابتدا برای آن‌ها شخصیت حقوقی مستقلی شناسایی شود. برای مثال، ممکن است بدون اعطای شخصیت حقوقی، بتوان اقداماتی مشابه سلب آزادی را از طریق جمع‌آوری یک محصول از بازار، غیرفعال کردن سامانه یا لغو مجوز بهره‌برداری از آن اعمال کرد.

در این بخش، کودکان بزهکار و اشخاص فاقد سلامت روان به‌عنوان الگوهایی برای بررسی امکان مسئولیت مستقیم کیفری سامانه‌های هوش مصنوعی مورد توجه قرار می‌گیرند. در مقابل، از بررسی وضعیت حیوانات صرف‌نظر می‌شود؛ زیرا به استثنای مواردی بسیار محدود، مجازات حیوانات دیگر با ارزش‌ها و رویکردهای حقوقی جوامع معاصر سازگار نیست و از این رو، نمی‌تواند الگویی مناسب و متناسب برای تحلیل مسئولیت کیفری سامانه‌های هوش مصنوعی باشد (Kurki, 2019).

بزهکاران فاقد سلامت روان در سامانه‌های هوش مصنوعی

نخستین پرسشی که باید مورد توجه قرار گیرد این است که آیا سامانه‌های هوش مصنوعی باید در نظام عدالت کیفری همانند بزهکاران فاقد سلامت روان تلقی شوند؟ آیا یک سامانه هوش مصنوعی می‌تواند با موفقیت به دفاع مبتنی بر فقدان سلامت روان استناد کند؟

اگر این دیدگاه پذیرفته شود که سامانه‌های هوش مصنوعی اساساً نمی‌توانند دارای اهلیت کیفری باشند، پاسخ به این پرسش طبیعتاً منفی خواهد بود. زیرا مفهوم فقدان سلامت روان بر این پیش‌فرض استوار است که شخص مبتلا به این وضعیت، در مرحله‌ای از زندگی خود دارای ویژگی‌ها و توانایی‌هایی بوده است که بعدها بر اثر یک بیماری یا اختلال روانی با چنان شدتی آسیب دیده‌اند که اهلیت کیفری او را از بین برده است. به بیان دیگر، مفهوم فقدان سلامت روان کیفری بر این فرض استوار است که همه اشخاصی که از سن معینی بالاتر هستند، اصولاً سالم و دارای اهلیت کیفری فرض می‌شوند. در حالی که عنوان «سالم» به حکم قانون و از ابتدا به شخص نسبت داده می‌شود، عنوان «فاقد سلامت روان» تنها پس از ارتکاب جرم و در مرحله رسیدگی قضایی به شخص اختصاص می‌یابد (Fletcher, 2001).

برخی نویسندگان معتقد شده‌اند که می‌توان سامانه‌های هوش مصنوعی را با نوعی «روان‌پریش ناقص» مقایسه کرد؛ یعنی موجودی که «فاقد توانایی درک اخلاقی است، اما از توانایی سنجش مصلحت و اقدام بر اساس آن برخوردار است». چنین شخصی دارای «عقلانیت ابزاری» است؛ بدین معنا که هدف‌هایی دارد و می‌تواند رفتار خود را با توجه به این هدف‌ها و پیامدهای احتمالی آن، از جمله احتمال تحمل مجازات کیفری، تنظیم کند. در واقع، او در ارتباط با پیامدهای قانونی رفتار خود، نوعی دوراندیشی و محاسبه‌گری دارد.

بااین‌حال، افراد دارای روان‌پریشی ناقص فاقد «توانایی عاطفی برای درک نادرستی اخلاقی رفتار خود» هستند و بنابراین انگیزه‌ای درونی برای رعایت قواعد ندارند، مگر آنکه رعایت آن قواعد با منافع شخصی آنان هماهنگ باشد. برخی استدلال کرده‌اند که همین سطح از پاسخ‌گویی به دلایل و محاسبه عقلانی، برای تحقق مسئولیت کیفری کافی است؛ بنابراین، افراد دارای روان‌پریشی ناقص نیز می‌توانند مسئولیت کیفری داشته باشند.

همان‌گونه که پیش‌تر بیان شد، می‌توان استدلال کرد که سامانه‌های هوش مصنوعی قادرند محیط پیرامون خود و آثار رفتارهایشان را تا حد زیادی تحلیل کنند. افزون بر این، ممکن است به‌گونه‌ای طراحی شوند که قواعد و ضمانت اجراها را در قالب محاسبات هزینه و فایده مورد توجه قرار دهند. بنابراین، آیا برای مسئول شناختن آن‌ها واقعاً نیاز است که از «انگیزه اخلاقی واقعی» برخوردار باشند؟ احتمالاً پاسخ منفی است.

پیش‌تر نیز مشخص شد که در برخی نظام‌های حقوقی، حتی اشخاص فاقد سلامت کامل روان نیز می‌توانند مسئول شناخته شده و در معرض واکنش‌های کیفری قرار گیرند. برای مثال، برخی نظام‌ها مانند نظام حقوقی ایتالیا، امکان اعمال مجازات کاهش‌یافته نسبت به اشخاصی را پذیرفته‌اند که تنها به‌صورت «نسبی» فاقد سلامت روان تشخیص داده می‌شوند. همچنین، برخی نظام‌های حقوقی به‌صراحت مقرر کرده‌اند که توانایی شخص در درک اخلاقی نادرست بودن جرم ارتکابی، شرط تعیین‌کننده‌ای برای احراز وجود یا فقدان اختلال روانی کیفری نیست.

بااین‌حال، اگر کسی دیدگاه مثبتی نسبت به اهلیت کیفری سامانه‌های هوش مصنوعی اتخاذ کند و معتقد باشد که این سامانه‌ها می‌توانند مخاطب قواعد حقوق کیفری قرار گیرند، پرسش دیگری مطرح خواهد شد: چه چیزی می‌تواند در مورد یک سامانه هوش مصنوعی به‌عنوان «بیماری روانی» تلقی شود؟ آیا یک ویروس یا اختلال فنی می‌تواند علت فقدان سلامت روان محسوب شود؟ (Chopra & White, 2011)

برخی پژوهشگران معتقدند که ربات‌ها ممکن است دچار اختلالات مشابه بیماری‌های روانی شوند، مشروط بر آنکه بتوان وجود نوعی آگاهی را در آن‌ها اثبات کرد. این اختلالات می‌توانند ناشی از واکنش‌های سازگارانه یا ناسازگارانه آن‌ها در برابر عوامل بیرونی یا حتی ناشی از طراحی اولیه انسانی آن‌ها باشند؛ به این معنا که ممکن است بازتابی از محدودیت‌ها یا ویژگی‌های طراحان انسانی خود باشند.

علاوه بر این، سامانه‌های هوش مصنوعی، به‌ویژه سامانه‌هایی که بر یادگیری ماشینی مبتنی هستند، دارای آسیب‌پذیری‌های خاصی هستند. نخستین آسیب‌پذیری، امکان «مسموم‌سازی داده‌ها» است. این وضعیت زمانی رخ می‌دهد که داده‌های نادرست یا فریبنده در مجموعه داده‌های آموزشی سامانه وارد شوند و در نتیجه، سامانه به نتایج اشتباه برسد.

برای مثال، ممکن است یک مدل یادگیری ماشینی توسط مهاجم ــ که حتی می‌تواند یک سامانه هوش مصنوعی دیگر باشد ــ فریب داده شود و به‌اشتباه یاد بگیرد که هر تصویر دارای یک مربع سفید کوچک، تصویر یک سگ است. این امر از طریق وارد کردن ارتباطات نادرست در داده‌های آموزشی رخ می‌دهد و سامانه در جریان آموزش این ارتباطات اشتباه را در ساختار خود جذب می‌کند. پس از تکمیل آموزش، کافی است تصویری دارای همان علامت سفید یا حتی نشانه‌ای کوچک‌تر به سامانه ارائه شود تا رفتار نادرست آن فعال گردد. آیا در چنین وضعیتی، سامانه هوش مصنوعی می‌تواند به دفاع مبتنی بر فقدان سلامت روان استناد کند؟

آسیب‌پذیری دیگر، حملات فریبنده است؛ یعنی شرایطی که در آن مجموعه‌ای از تغییرات بسیار ظریف در ورودی سامانه ایجاد می‌شود تا باعث شود مدل هدف، داده ورودی را به‌اشتباه طبقه‌بندی کند. این نمونه‌های دستکاری‌شده برای انسان قابل تشخیص نیستند، اما می‌توانند عملکرد سامانه را به‌طور جدی تغییر دهند (Seher, 2016).

ترنر معتقد است که می‌توان میان دو وضعیت تفاوت قائل شد: نخست، مواردی که سامانه هوش مصنوعی درباره یک واقعیت دچار اشتباه شده است و دوم، مواردی که سامانه نسبت به یک واقعیت شناخته‌شده، قاعده نادرستی را اعمال کرده است. برای مثال، زمانی که یک ربات کارخانه‌ای تصور می‌کند سر یک کارگر انسانی بخشی از فرایند تولید است و تصمیم می‌گیرد آن را خرد کند و موجب مرگ او می‌شود، این وضعیت مشابه اشتباه در تشخیص واقعیت است.

در چنین شرایطی، می‌توان تصور کرد که اگر اشتباه ربات ناشی از یک فرایند عصبی ـ الکترونیکی معیوب، یعنی نوعی فرایند ذهنی ناشی از یک ویروس یا اختلال فنی، باشد، استدلالی برای پذیرش دفاع مبتنی بر فقدان سلامت روان وجود خواهد داشت.

برای بررسی بهتر این موضوع، می‌توان به نمونه‌ای اشاره کرد که ترنر از پرونده «هولبروک علیه پرودومکس اتومیشن» ارائه کرده است. واندا هولبروک، تکنسین تعمیرات باتجربه در یک کارخانه تولید قطعات خودرو در ایالت میشیگان بود. این شرکت از ربات‌ها برای تولید قطعات اتصال‌دهنده یدک استفاده می‌کرد. کل خط تولید توسط یک رایانه مرکزی کنترل می‌شد که نرم‌افزار کنترل‌گر قابل برنامه‌ریزی بر روی آن اجرا می‌شد.

هولبروک بدون رعایت دستورالعمل‌های ایمنی تعیین‌شده توسط کارفرما، برای انجام تعمیرات وارد محدوده بازوی رباتیک شد. در این هنگام، ربات سر او را میان قطعه‌ای که از قبل در محل نصب شده بود و بخش متحرک دستگاه فشرد و موجب مرگ او شد. سپس ربات دیگری تلاش کرد قطعه جدید را جوش دهد که در نتیجه آن، صورت، بینی و دهان وی به‌شدت سوخت (Awad, 2022).

به اعتقاد ترنر، پرونده هولبروک با مواردی متفاوت است که در آن یک سامانه هوش مصنوعی به شکلی غیرمنتظره از دستورهای انسانی فاصله می‌گیرد؛ مانند اینکه یک دستگاه توستر برای خنک کردن نان‌ها، خانه را به آتش بکشد. همچنین با وضعیتی که در آن یک سامانه هوش مصنوعی توانایی نافرمانی عمدی از دستورهای روشن انسانی را پیدا کند، تفاوت دارد.

چنین مواردی بیشتر به مفهومی نزدیک می‌شوند که می‌توان آن را «ذهن مجرمانه و واجد تقصیر» دانست؛ مفهومی که در حقوق کیفری سنتی با عنصر معنوی جرم شناخته می‌شود.

بزهکاران کودک‌مانندِ هوش مصنوعی

اکنون باید به بررسی پرسش دوم پرداخت؛ یعنی آیا نظام عدالت کیفری باید با سامانه‌های هوش مصنوعی همانند کودکان برخورد کند؟ با بررسی نحوه تنظیم وضعیت کودکان در برخی نظام‌های حقوقی، می‌توان به برخی ویژگی‌های مشترک دست یافت. برای نمونه، ماده ۹۷ قانون مجازات ایتالیا، درباره افراد کمتر از چهارده سال، فرض فقدان مطلق اهلیت کیفری را مقرر کرده است. با وجود این، مطابق مواد ۲۲۲ و ۲۲۴ همان قانون، قاضی همچنان می‌تواند در برابر کودک، برخی اقدامات الزام‌آور و محدودکننده را اتخاذ کند. همچنین، ماده ۱۹ قانون کیفری آلمان مقرر می‌دارد که شخص کمتر از چهارده سال نمی‌تواند مسئول ارتکاب جرم شناخته شود؛ هرچند این امر مانع از مسئولیت معاونان جرم نخواهد بود (ماده ۲۹) (Badea & Artus, 2022).

در ایالات متحده آمریکا، رسیدگی و واکنش نسبت به رفتار مجرمانه کودکان عمدتاً در صلاحیت حقوق خانواده قرار دارد. برخی ایالت‌ها دفاع مبتنی بر کودکی را که ریشه در حقوق عرفی دارد، در قوانین خود وارد کرده‌اند؛ بر اساس این قاعده، کودکان کمتر از هفت سال به‌طور مطلق فاقد اهلیت کیفری فرض می‌شوند. بیشتر ایالت‌ها مسئله کودکان بزهکار را از طریق تعیین سنی حل می‌کنند که صلاحیت دادگاه اطفال را ایجاد می‌کند؛ این سن معمولاً چهارده سال است. تعداد محدودی از این ایالت‌ها نیز فرض فقدان اهلیت را پذیرفته‌اند که امکان رد آن در دادگاه وجود دارد؛ به این معنا که اگر ثابت شود کودک از نادرستی رفتار خود آگاه بوده است، ممکن است مسئول شناخته شود.

بااین‌حال، در بسیاری از موارد، دادگاه‌های اطفال حداقل سنی مشخصی برای رسیدگی تعیین نمی‌کنند. نتیجه این امر آن است که اگر یک ایالت دفاع مبتنی بر کودکی را در نظام حقوق اطفال خود وارد نکرده باشد، ممکن است حتی کودکی کمتر از هفت سال نیز به‌عنوان بزهکار شناخته شود؛ حتی در شرایطی که فاقد مسئولیت کیفری بوده است.

همان‌گونه که هالوی اشاره کرده است، و در بررسی امکان تعمیم این استدلال به سامانه‌های هوش مصنوعی نیز تأمل کرده است، بسیاری از سامانه‌های هوش مصنوعی امروزی قادرند میان امور مجاز و ممنوع تمایز قائل شوند. از این منظر، مقایسه هوش مصنوعی با کودکان، به‌ویژه از دو جهت، مناسب‌تر از مقایسه آن با اشخاص فاقد سلامت روان به نظر می‌رسد: نخست، از آن جهت که در بسیاری از موارد، نظام حقوقی برای مواجهه با کودکان به حوزه‌های دیگر حقوقی که مناسب‌تر تلقی می‌شوند ارجاع می‌دهد؛ و دوم، از حیث نقش و مسئولیت اشخاص انسانی بزرگسال که در ارتکاب رفتار کودک دخالت داشته‌اند.

تشبیه میان هوش مصنوعی و کودکان غالباً برای نشان دادن محدودیت‌های سامانه‌های هوش مصنوعی کنونی مطرح می‌شود. این سامانه‌ها ممکن است در انجام برخی فعالیت‌های تخصصی، مانند بازی گو، شطرنج یا حتی رانندگی در شرایط خاص، عملکرد بسیار خوبی داشته باشند، اما در همان حال، در انجام بسیاری از کارهایی که یک کودک دوساله به‌طور طبیعی قادر به انجام آن‌هاست، ناتوان باشند.

در واقع، دانشمندان علوم رایانه تلاش می‌کنند سامانه‌های هوش مصنوعی‌ای ایجاد کنند که بتوانند توانایی‌های مشابه کودکان، به‌ویژه فرایند یادگیری آن‌ها، را تقلید کنند (Chesterman, 2020).

به‌طور خلاصه، هدف آن‌ها ایجاد سامانه‌هایی است که بتوانند کاری فراتر از آنچه به‌طور مستقیم به آن‌ها آموزش داده شده است انجام دهند. برای توضیح این موضوع می‌توان از این تشبیه استفاده کرد:

«اگر به مدل‌های کنونی هوش مصنوعی نگاه کنیم، گویی برای این سامانه‌ها مادرانی بیش از حد مراقب و کنترل‌گر ایجاد کرده‌ایم. برنامه‌نویسی همواره بر عملکرد سامانه نظارت دارد و به آن می‌گوید: بله، این مورد را درست انجام دادی؛ این تصویر گربه است، آن تصویر سگ است، این یکی هم گربه است و دیگری سگ. یا سامانه‌ای وجود دارد که هنگام بازی، صرفاً بر اساس نتیجه دریافت‌شده هدایت می‌شود؛ زمانی که امتیاز آن افزایش می‌یابد، همان روش را ادامه می‌دهد و زمانی که شکست می‌خورد، باید راه دیگری پیدا کند.

نتیجه طبیعی چنین رویکردی آن است که سامانه‌های هوش مصنوعی در انجام کارهایی که برای آن آموزش دیده‌اند، بسیار موفق می‌شوند، اما در انجام کارهای متفاوت، انعطاف‌پذیری بسیار کمتری دارند. برای مثال، یک سامانه می‌تواند در بازی شطرنج بسیار قدرتمند باشد، اما اگر به یک کودک گفته شود که از این پس قوانین حرکت مهره اسب تغییر کرده است، کودک معمولاً می‌تواند خود را با شرایط جدید تطبیق دهد؛ در حالی که انجام چنین کاری برای سامانه‌های هوش مصنوعی بسیار دشوارتر است.»

افزون بر این، همان‌گونه که دیامانتیس و همکارانش بیان کرده‌اند، تشبیه میان الگوریتم‌ها و کودکان زمانی که مسئله تعیین مسئولیت ناشی از زیان‌های هوش مصنوعی مطرح می‌شود، با محدودیت مواجه می‌گردد (Mokhtarian, 2020):

«برای هر کودک، معمولاً تنها یک یا دو انسان وجود دارند که به‌طور آشکار می‌توانند مسئول شناخته شوند. اما درباره الگوریتم‌ها وضعیت بسیار پیچیده‌تر است. تیم‌های بزرگ و پراکنده‌ای از افراد بر روی مهم‌ترین الگوریتم‌های امروزی کار می‌کنند. بنابراین، شخص انسانی مشخصی که بتوان مسئولیت را متوجه او کرد، وجود ندارد. حتی اگر چنین شخصی وجود داشته باشد، باز هم روشن نیست که مسئول دانستن او چه فایده‌ای خواهد داشت؛ زیرا این الگوریتم‌های پیچیده غالباً فراتر از توانایی تأثیرگذاری معنادار هر فرد انسانی قرار دارند.»

راه‌حلی که این نویسندگان پیشنهاد می‌کنند، پذیرش مسئولیت اشخاص حقوقی و به‌ویژه شرکت‌ها در قبال خسارات ناشی از عملکرد سامانه‌های هوش مصنوعی است.

نتیجه‌گیری

مسائلی که در این مقاله مورد بررسی قرار گرفتند، در چارچوب چالش کلی «شایستگی برای قرار گرفتن در قلمرو حقوق کیفری» قرار می‌گیرند. این چالش کلی را می‌توان به سه جزء تفکیک کرد: دو پیش‌فرض (الف) و (ب) که در نهایت به نتیجه (ج) منتهی می‌شوند ($A \wedge B \rightarrow C$). بر این اساس، چالش کلی شایستگی بیان می‌دارد:

الف) توانایی انجام رفتار قابل سرزنش، یعنی اهلیت کیفری که در حقوق کیفری از طریق نهادهای دفاع مبتنی بر فقدان اهلیت (مانند کودکی و فقدان سلامت روان) بیان شده است، یک شرط عمومی و اساسی برای اعمال حقوق کیفری محسوب می‌شود و «عدم برخورداری از این توانایی موجب خارج شدن موجود مورد نظر از قلمرو مجازات مناسب خواهد شد»؛

و

ب) هوش مصنوعی فاقد توانایی‌های استدلال عملی لازم برای قابل سرزنش بودن است؛

بنابراین:

ج) هوش مصنوعی در قلمرو حقوق کیفری قرار نمی‌گیرد.

همان‌گونه که پیش‌تر در متن مقاله اشاره شد، این نویسندگان سه راه‌حل احتمالی برای این چالش ارائه می‌کنند: مسئولیت مبتنی بر عمل دیگری، مسئولیت مطلق و ایجاد چارچوبی برای تحلیل مستقیم عنصر معنوی جرم در مورد هوش مصنوعی.

اکنون باید بر آخرین راه‌حل تمرکز کرد؛ به‌ویژه اینکه قابل سرزنش بودن یک سامانه هوش مصنوعی «به خودی خود» چه معنایی خواهد داشت.

به‌طور کلی، حقوق کیفری نسبت به حالت‌های ذهنی پنهان یا انگیزه‌های فرد برای نقض قانون بی‌تفاوت است. نظام‌های عدالت کیفری بیش از هر چیز به افرادی توجه دارند که بی‌اعتنایی ناکافی خود نسبت به منافع و ارزش‌های مورد حمایت قانون را در رفتار بیرونی خود آشکار می‌کنند. در واقع، می‌توان استدلال کرد که حقوق کیفری:

«از ما نمی‌خواهد که الزاماً با انگیزه احترام به دیگران یا حتی احترام به قانون رفتار کنیم؛ بلکه تنها انتظار دارد که بی‌احترامی خود را از طریق رفتارهایی که با اعطای وزن مناسب به منافع و ارزش‌های مورد حمایت قانون ناسازگار است، آشکار نسازیم. بنابراین، مسئولیت کیفری را می‌توان بیشتر ناظر بر آنچه رفتار فرد آشکار می‌کند دانست تا ظرافت‌های مربوط به انگیزه‌ها، افکار و احساسات خصوصی او.»

بر این اساس، در این دیدگاه تنها مفهوم «حقوقی» تقصیر و سرزنش‌پذیری اهمیت دارد؛ یعنی تا زمانی که شخص «از مرز قانونی عبور کند» و توانایی ارتکاب رفتار مجرمانه را داشته باشد، شایستگی تحمل مجازات کیفری را خواهد داشت. آنچه اهمیت دارد این است که شخص «توانایی انجام رفتارهایی را داشته باشد که نشان‌دهنده بی‌اعتنایی ناکافی نسبت به دلایل و ملاحظات مورد شناسایی قانون است».

می‌توان مفهوم اهلیت کیفری اشخاص حقوقی را نیز بر اساس همین برداشت از سرزنش‌پذیری تبیین کرد؛ زیرا اشخاص حقوقی دارای فرایندهای جمع‌آوری اطلاعات، استدلال و تصمیم‌گیری هستند. آن‌ها از طریق اعضای خود، دلایلی را که حقوق کیفری انتظار دارد مورد توجه قرار گیرند، ارزیابی کرده و بر اساس آن‌ها اقدام می‌کنند. از این رو، اشخاص حقوقی می‌توانند رفتاری انجام دهند که بی‌اعتنایی ناکافی آن‌ها نسبت به منافع حقوقی دیگران را آشکار سازد.

این تحلیل چگونه درباره سامانه‌های هوش مصنوعی قابل اعمال است؟ بدون تردید، سامانه‌های هوش مصنوعی نیز دارای فرایندهای جمع‌آوری اطلاعات، استدلال و تصمیم‌گیری هستند. رفتار آن‌ها ــ که نتیجه تعیین «کارآمدترین ابزارها» برای دستیابی به اهدافشان است ــ ممکن است از نظر کارکردی معادل آشکار کردن بی‌اعتنایی ناکافی نسبت به ارزش‌ها و الزامات قواعد کیفری تلقی شود؛ به عبارت دیگر، رفتاری که از منظر حقوق کیفری واجد وصف مجرمانه است.

این استدلال با نظریه هو نیز هماهنگ است؛ زیرا وی معتقد است نخستین شرط برای اعمال مسئولیت کیفری نسبت به ربات‌ها، برخورداری آن‌ها از الگوریتم‌های اخلاقی است.

در ادامه می‌توان فرض کرد که همه نظام‌های حقوق کیفری تا حدودی این پیش‌فرض را پذیرفته‌اند که شهروندان از محتوای قوانین آگاهی دارند و بر همین اساس، آنان را مسئول می‌دانند. در واقع، ناآگاهی از قانون اصولاً مانع مسئولیت نیست (قاعده «جهل به قانون رافع مسئولیت نیست»)، مگر در موارد استثنایی اشتباه نسبت به حکم قانون.

همین فرض درباره آگاهی از قواعد اخلاقی نیز تا حدودی وجود دارد:

«به نظر می‌رسد قانون، افزون بر آگاهی از قواعد مشخص حقوقی، فرض می‌کند که افراد به سطح معینی از درک اخلاقی دست یافته‌اند. از این منظر، مرتکب باید دست‌کم توانایی پیروی عملی از قواعد مقرر قانونی را داشته باشد. این امر بدین معناست که او باید بتواند الزامات قانون را درک کرده و رفتار خود را بر اساس آن تغییر دهد. برای تحقق این امر، مرتکب باید از حداقلی از توانایی عقلانی برخوردار باشد که به او اجازه دهد قواعد کلی را به رفتارهای روزمره تبدیل کند.»

افزون بر این، می‌توان فرض کرد که قواعد کیفری، به دلیل ویژگی ذاتی و کارکرد دستوری خود، نسبت به قواعد اخلاقی آسان‌تر در سامانه‌های هوش مصنوعی قابل پیاده‌سازی هستند. بنابراین، درباره سامانه‌های هوش مصنوعی، فرض آگاهی از قانون می‌تواند تقریباً مطلق باشد. اما چنین فرضی درباره قواعد اخلاقی قابل پذیرش نیست؛ زیرا معیار مورد انتظار از هوش مصنوعی در این زمینه بسیار پایین‌تر از معیار مورد انتظار از انسان‌هاست.

به بیان دیگر، ممکن است در آینده با سامانه‌های هوش مصنوعی‌ای مواجه شویم که از نظر حقوقی بسیار مطیع باشند، اما از نظر اخلاقی فاقد ارزش‌های انسانی تلقی شوند.

بااین‌حال، مسئله اصلی اندکی متفاوت است. مسئله واقعی این نیست که آیا ماشین‌ها می‌توانند مانند شهروندان خوب رفتار کنند؛ یعنی رفتار اخلاقاً درست داشته باشند. مسئله اساسی این است که آیا آن‌ها می‌توانند به اختیار خود رفتار نادرست انجام دهند.

چگونه می‌توان آن بخش از شرارت انسانی را که شاید همه ما در وجود خود داریم اما هرگز آن را آشکار نمی‌کنیم، در یک سامانه مصنوعی ایجاد کرد؟ همان بخشی که موجب می‌شود انسان بتواند یک قاعده را نادیده بگیرد. چگونه می‌توان چنین امری را به یک تصمیم مبتنی بر محاسبه هزینه و فایده تبدیل کرد؟

اگر شخصی یک سامانه هوش مصنوعی را به‌گونه‌ای طراحی کند که ــ در فرضی نظری ــ مرتکب جرم شود، خود او مسئولیت کیفری خواهد داشت. در همان حال، سامانه هوش مصنوعی نیز احتمالاً تنها به این دلیل مرتکب جرم می‌شود که خالق آن امکان چنین رفتاری را برای آن فراهم کرده است.

این واقعیت که سامانه‌های هوش مصنوعی محصولات ساخته دست انسان هستند، به نظر می‌رسد مانعی جدی و حتی غیرقابل عبور ایجاد می‌کند. با وجود این، بدون بررسی نحوه حل چنین مشکلاتی در حوزه مهم‌ترین ساخته انسانی، یعنی مسئولیت کیفری اشخاص حقوقی، نمی‌توان به نتایج قطعی دست یافت.

در نهایت، همان‌گونه که در ابتدای این مقاله اشاره شد، شخصیت حقوقی در حقوق کیفری ارتباطی مستقیم و جدایی‌ناپذیر با نظریه‌های مجازات دارد؛ زیرا تنها در این چارچوب است که مفهوم شخصیت کیفری معنا پیدا می‌کند. بنابراین، باید این پرسش اساسی را مطرح کرد:

چه فایده‌ای خواهد داشت که یک سامانه هوش مصنوعی را به‌عنوان مخاطب قواعد حقوق کیفری شناسایی کنیم؟

در پاسخ باید عنوان کرد که شناسایی سامانه‌های هوش مصنوعی به‌عنوان مخاطب قواعد حقوق کیفری، پیش از آنکه با هدف مجازات‌کردن یک ماشین به معنای سنتی آن مطرح باشد، پاسخی به ضرورت‌های کارکردی و سیاست جنایی در جوامع فناورانه است. مهم‌ترین فایده چنین رویکردی، پر کردن خلأهای مسئولیت در مواردی است که رفتار زیان‌بار ناشی از عملکرد مستقل و پیچیده یک سامانه هوش مصنوعی است و انتساب آن به اشخاص انسانی ــ مانند برنامه‌نویسان، کاربران یا مدیران شرکت‌ها ــ با دشواری‌های جدی مواجه می‌شود. در چنین شرایطی، شناسایی مسئولیت مستقل برای هوش مصنوعی می‌تواند مانع از ایجاد حوزه‌ای بدون پاسخ‌گویی در حقوق کیفری شود.

نخستین فایده این رویکرد، تقویت کارکرد بازدارندگی حقوق کیفری است. یکی از اهداف اساسی مجازات، جلوگیری از ارتکاب رفتارهای زیان‌بار آینده است. اگر سامانه‌های هوش مصنوعی بتوانند بدون هیچ پیامد حقوقی مستقیم، تصمیم‌هایی اتخاذ کنند که منجر به آسیب‌های گسترده شوند، انگیزه کافی برای طراحی سامانه‌های ایمن‌تر و کنترل‌پذیرتر کاهش خواهد یافت. پذیرش نوعی مسئولیت کیفری یا شبه‌کیفری برای این سامانه‌ها می‌تواند سازوکاری برای وادار کردن طراحان و بهره‌برداران به ایجاد معماری‌های ایمن، قابل نظارت و پاسخ‌گو فراهم کند. دومین فایده، حفظ انسجام مفهومی حقوق کیفری است. حقوق کیفری پیش‌تر نیز از چارچوب کاملاً انسان‌محور فاصله گرفته و مسئولیت کیفری اشخاص حقوقی را پذیرفته است؛ در حالی که شرکت‌ها نیز مانند انسان‌ها دارای آگاهی، اراده و تجربه ذهنی نیستند. پذیرش مسئولیت برای اشخاص حقوقی نشان می‌دهد که ملاک مسئولیت کیفری الزاماً وجود ذهن انسانی نیست، بلکه می‌تواند توانایی سازمان‌یافته برای تصمیم‌گیری، پردازش اطلاعات و ایجاد آثار اجتماعی باشد. از این منظر، سامانه‌های هوش مصنوعی پیشرفته نیز ممکن است در آینده واجد ویژگی‌هایی شوند که توجیه‌کننده نوعی مسئولیت مستقل باشد.

سومین فایده، تغییر تمرکز حقوق کیفری از جست‌وجوی صرف «ذهن مجرمانه» به ارزیابی توانایی عملی یک عامل در تولید رفتارهای قابل انتساب است. همان‌گونه که در مورد اشخاص حقوقی، معیار مسئولیت نه وجود احساس گناه یا انگیزه اخلاقی، بلکه توانایی تصمیم‌گیری و آشکار ساختن بی‌اعتنایی نسبت به منافع حمایت‌شده قانونی است، درباره هوش مصنوعی نیز می‌توان بر قابلیت تصمیم‌گیری مستقل، پیش‌بینی پیامدها و تأثیرگذاری واقعی بر جهان خارج تمرکز کرد.

چهارمین فایده، ایجاد مرز روشن‌تر میان مسئولیت انسانی و مسئولیت فناوری است. در وضعیت کنونی، خسارات ناشی از هوش مصنوعی غالباً میان چندین عامل انسانی پراکنده می‌شود؛ به‌گونه‌ای که تعیین شخص مسئول ممکن است دشوار یا حتی غیرممکن باشد. شناسایی جایگاهی مستقل برای سامانه هوش مصنوعی می‌تواند به‌عنوان یک نقطه اتصال حقوقی عمل کرده و امکان اعمال تدابیر کنترلی مانند محدودسازی فعالیت، توقف عملکرد، اصلاح الگوریتم یا حذف سامانه‌های خطرناک را فراهم کند.

با وجود این، پذیرش هوش مصنوعی به‌عنوان مخاطب حقوق کیفری لزوماً به معنای اعطای حقوق کامل انسانی یا پذیرش شخصیت حقوقی مشابه انسان نیست. همان‌گونه که درباره اشخاص حقوقی یا حتی کودکان و اشخاص فاقد سلامت روان مشاهده می‌شود، می‌توان نظامی تدریجی و محدود از مسئولیت را طراحی کرد که هدف آن نه انتقام یا سرزنش اخلاقی، بلکه کنترل خطر، تضمین پاسخ‌گویی و حمایت از منافع عمومی باشد.

بنابراین، مهم‌ترین فایده شناسایی هوش مصنوعی به‌عنوان مخاطب حقوق کیفری، ایجاد چارچوبی حقوقی برای مدیریت عاملانی است که اگرچه فاقد آگاهی و تجربه انسانی هستند، اما می‌توانند به‌صورت مستقل تصمیم بگیرند، بر محیط اجتماعی اثر بگذارند و موجب ورود خسارت شوند. در نهایت، مسئله اصلی این نیست که آیا ماشین‌ها شایسته مجازات اخلاقی هستند، بلکه این است که آیا حقوق کیفری می‌تواند در برابر اشکال جدید عاملیت و قدرت تصمیم‌گیری، بدون ایجاد خلأ مسئولیت، همچنان کارآمد باقی بماند.

تعارض منافع

در انجام مطالعه حاضر، هیچ‌گونه تضاد منافعی وجود ندارد.

مشارکت نویسندگان

در نگارش این مقاله تمامی نویسندگان نقش یکسانی ایفا کردند.

حامی مالی

این پژوهش حامی مالی نداشته است.

EXTENDED SUMMARY

The traditional architecture of criminal law has historically been constructed around the model of the rational human actor: an individual presumed to possess consciousness, volition, practical reasoning, moral understanding, and the capacity to choose between lawful and unlawful courses of conduct. Criminal responsibility consequently presupposes an agent who can perform the external elements of an offense while simultaneously possessing a legally relevant mental state, and whose conduct can therefore be attributed to that agent as blameworthy. This anthropocentric model has shaped the doctrines of criminal capacity, culpability, punishment, and legal personhood in modern legal systems. The rapid development of autonomous artificial intelligence, however, increasingly destabilizes these assumptions. Advanced AI systems are no longer limited to the passive execution of predefined commands but are capable of processing large volumes of information, learning from environmental inputs, adapting their strategies, selecting among alternative courses of action, and pursuing complex goals with limited or even negligible real-time human intervention. Such capacities have generated an emerging category of functional agency that does not fit comfortably within conventional distinctions between human actors and technological instruments. The central problem is therefore not simply whether artificial intelligence resembles a human mind, but whether criminal law can continue to rely exclusively on human-centered criteria of responsibility when technologically autonomous systems increasingly exercise forms of decision-making that generate legally significant consequences. Existing scholarship on autonomous artificial agents and legal personality suggests that the legal system may need to distinguish between biological personhood and functional forms of agency, particularly where autonomous systems are capable of producing risks, making decisions, and affecting legally protected interests in ways not entirely reducible to the immediate conduct of programmers, owners, or users (Chesterman, 2020; Chopra & White, 2011; Kurki, 2019; Lima, 2018).

Against this background, the present study reconsiders the concept of criminal capacity by examining the possible transition from the “rational actor” to the “rational algorithm.” The analysis begins from the premise that criminal capacity is not merely synonymous with legal personality but constitutes a more specific threshold condition for being an appropriate addressee of criminal norms and punitive consequences. In comparative criminal-law theory, capacity is closely associated with the ability to understand the nature and legal significance of one’s conduct and with the ability to regulate behavior according to that understanding. German criminal law, for example, distinguishes the psychological elements of an offense from normative culpability and treats the absence of the capacity to appreciate wrongfulness or act in accordance with such appreciation as a basis for excluding culpability. Italian criminal law similarly links imputability to the capacities of understanding and volition, while criminal responsibility in the United States is indirectly structured by doctrines such as insanity, which investigate whether a defendant was capable of understanding the nature, quality, or wrongfulness of the relevant conduct. Despite important doctrinal differences, these approaches converge on two basic requirements: cognitive competence and behavioral control. Criminal responsibility therefore presupposes not only that conduct can be causally linked to an actor, but also that the actor possesses sufficient capacities to make punishment normatively meaningful. These requirements are deeply connected to broader theories of culpability, free will, deterrence, and retribution, since punishment is generally justified only where an actor can understand legal reasons and modify conduct in response to them (Bohlander, 2009; Fernandes, 2022; Fletcher, 2001; Keiler & Roef, 2019; Palazzo & Papa, 2013; Rengier, 2019).

The first substantive question is whether autonomous AI systems can satisfy the cognitive dimension of criminal capacity. If this requirement is understood narrowly as the capacity to comprehend the physical nature and likely consequences of conduct, sophisticated AI systems may in many contexts satisfy it more effectively than human beings. Through predictive modeling, probabilistic inference, environmental sensing, and access to extensive datasets, an AI system may accurately identify causal relations, foresee physical consequences, and distinguish among alternative outcomes. The greater difficulty emerges when cognition is understood to include awareness of legal wrongfulness or moral impermissibility. Legal norms could, at least theoretically, be embedded within artificial systems in forms that permit machines to identify prohibited conduct, process applicable rules, and incorporate sanctions into computational decision-making. Indeed, a technologically advanced system might store and process legal rules with a degree of consistency and comprehensiveness unavailable to human decision-makers. Nevertheless, contemporary AI does not independently experience a norm as binding in the phenomenological or moral sense in which human beings may understand obligation. The same difficulty becomes more pronounced when the relevant standard is moral rather than legal wrongfulness. Computational ethics and the development of artificial moral agents attempt to translate moral values into machine-readable decision frameworks, yet the ambiguity, contextual sensitivity, and contestability of moral principles create substantial interpretive difficulties. The “alignment problem” demonstrates that even a formally specified objective can be interpreted by an advanced system in unexpected ways that remain computationally coherent while diverging from human purposes and values. Research on artificial moral agency therefore illustrates both the possibility of machine-guided ethical decision-making and the continuing limitations of formalizing complex moral judgments (Abel et al., 2016; Awad, 2022; Badea & Artus, 2022; Kirchner, 2024; Simmler & Markwalder, 2019).

The second dimension concerns behavioral control, which may be even more significant for determining whether AI can become a meaningful subject of criminal-law regulation. Criminal capacity traditionally assumes that an individual can not only understand relevant reasons but can also regulate conduct accordingly, meaning that the actor possesses a practical capacity to comply with or violate the law. At first glance, artificial systems appear unable to satisfy this requirement because their behavior ultimately depends on programming, architecture, training data, optimization procedures, and system design. Yet an absolute distinction between programmed machines and freely choosing humans may be conceptually overstated. Human conduct is itself influenced by biological dispositions, social conditions, environmental incentives, learned habits, and psychological constraints, while free will remains a normative attribution rather than an empirically demonstrable property. Criminal law nevertheless attributes responsibility where an individual possesses a legally sufficient capacity to respond to reasons, resist prohibited impulses, and alter conduct in light of expected consequences. From this perspective, the relevant question for AI is not whether a machine possesses metaphysical freedom identical to human free will, but whether it has sufficient operational autonomy to select among options, respond to legal incentives, adapt conduct, and modify future behavior based on consequences. A functional conception of control could therefore focus on observable capacities rather than subjective consciousness. If an autonomous system can distinguish legally relevant alternatives, evaluate predicted sanctions, revise behavior after negative outcomes, and independently pursue strategies not directly specified at the moment of action by a human operator, then it may possess a limited but legally significant analogue of behavioral self-control. Such an approach is consistent with normative accounts of culpability that emphasize control, responsiveness to reasons, and legally relevant agency rather than inaccessible

internal experiences (Bonicalzi & Haggard, 2019; Fernandes, 2022; Keiler & Roef, 2019; Simmler & Markwalder, 2019).

A further dimension of the analysis concerns whether existing doctrines governing persons with diminished or absent criminal capacity can provide useful analogies for regulating artificial intelligence. Criminal law does not simply ignore actors who lack full capacity. Children, persons affected by serious mental disorders, and other individuals who cannot satisfy ordinary requirements of culpability may nevertheless be subjected to preventive, rehabilitative, protective, or restrictive measures. This demonstrates an important distinction between moral blameworthiness and the broader regulatory functions of criminal justice. A similar distinction may become relevant for AI. An autonomous system need not initially be recognized as a full legal person in order to become the object of legally structured restrictions such as deactivation, removal from the market, suspension of operational authorization, mandatory modification, or technological confinement. Comparisons with mental incapacity also illuminate possible cases in which an AI system malfunctions because of corrupted data, adversarial manipulation, technical defects, or software infection. Data poisoning and adversarial attacks may alter machine perception or decision-making in ways functionally comparable, although not ontologically identical, to impairments that affect human cognition. The analogy with children may be even more productive because AI systems are often trained through iterative feedback and may display high competence within narrow domains while remaining incapable of generalizing basic contextual knowledge. Yet the analogy also has important limits: unlike children, complex AI systems are commonly produced, trained, deployed, and modified by diffuse networks of programmers, companies, users, and data providers, making the attribution of derivative human responsibility substantially more difficult. These responsibility gaps strengthen arguments for corporate liability, technological regulation, and potentially graduated forms of direct accountability for autonomous systems (Bonicalzi & Haggard, 2019; Chesterman, 2020; Chopra & White, 2011; Kurki, 2019; Mokhtarian, 2020; Seher, 2016).

The study concludes that the transition from the rational actor to the rational algorithm should not be understood as a demand for the immediate recognition of artificial intelligence as a human-equivalent legal person, nor as an attempt to attribute moral consciousness to machines that do not possess subjective experience. Rather, it represents a functional reconsideration of the criteria by which criminal law identifies actors capable of producing legally attributable conduct. As autonomous systems become increasingly capable of independent learning, complex decision-making, environmental adaptation, prediction of consequences, and goal-directed action, a legal framework that treats every AI system exclusively as a passive instrument may become incapable of addressing responsibility gaps created by technological autonomy. Criminal capacity should therefore be reconsidered as a graduated and functional concept grounded in measurable capacities such as operational autonomy, practical reasoning, responsiveness to legal norms, behavioral control, predictability, learning ability, and the capacity to produce significant effects in the external world. Such a model would allow different levels of responsibility to correspond to different levels of technological autonomy, without requiring artificial systems to possess human consciousness, emotions, or moral experience. The principal value of recognizing AI as a possible addressee of criminal-law norms lies not in imposing traditional punishment on machines, but in preserving accountability, strengthening preventive regulation, clarifying the division of responsibility between human and technological actors, and enabling the legal system to impose meaningful restrictions on autonomous systems that generate serious risks. The fundamental question for future criminal law is therefore not whether machines can become morally guilty in the human sense, but whether the

legal system can remain coherent and effective if increasingly autonomous decision-making agents continue to operate outside any direct framework of criminal accountability.

References

Abel, D., MacGlashan, J., & Littman, M. L. (2016). *Reinforcement Learning as a Framework for Ethical Decision Making* AAAI Workshop: AI, Ethics, and Society,

Awad, E. (2022). Computational Ethics. *Trends in Cognitive Sciences*, *26*(5).

Badea, C., & Artus, G. (2022). Morality, Machines, and the Interpretation Problem: A Value-Based, Wittgensteinian Approach to Building Moral Agents. In *Artificial Intelligence XXXIX. SGAI-AI 2022* (Vol. 13652). Springer.

Bohlander, M. (2009). *Principles of German Criminal Law*. Hart Publishing.

Bonicalzi, S., & Haggard, P. (2019). Responsibility Between Neuroscience and Criminal Law: The Control Component of Criminal Liability. *Rivista Internazionale di Filosofia e Psicologia*, *10*(2).

Chesterman, S. (2020). AI and the Limits of Legal Personality. *International and Comparative Law Quarterly*, *69*(4).

Chopra, S., & White, L. (2011). *A Legal Theory for Autonomous Artificial Agents*. University of Michigan Press.

Fernandes, G. (2022). Law and Science: The Autonomy and Limits of Culpability as a Cornerstone to the Ascription of Liability (or the Subject of Criminal Law: Three Maxims, a Problem and a Glimpse into the Future). *International Journal for the Semiotics of Law*.

Fletcher, G. (2001). Criminal Theory in the Twentieth Century. *Theoretical Inquiries in Law*, *2*(1).

Keiler, J., & Roef, D. (2019). *Comparative Concepts of Criminal Law* (3rd ed.). Intersentia.

Kirchner, J. H. (2024). *Understanding AI Alignment Research: A Systematic Analysis*. https://arxiv.org/abs/2206.02841

Kurki, V. A. J. (2019). *A Theory of Legal Personhood*. Oxford University Press.

Lima, D. (2018). Could AI Agents Be Held Criminally Liable? Artificial Intelligence and the Challenges for Criminal Law. *South Carolina Law Review*, *69*(3).

MacCormick, N. (2008). *Institutions of Law: An Essay in Legal Theory (Law, State, and Practical Reason)*. Oxford University Press.

Mokhtarian, E. (2020). The Bot Legal Code: Developing a Legally Compliant Artificial Intelligence. *Vanderbilt Journal of Entertainment and Technology Law*, *21*.

Palazzo, F. C., & Papa, M. (2013). *Lezioni di Diritto Penale Comparato* (3rd ed.). Giappichelli.

Rengier, R. (2019). *Strafrecht Allgemeiner Teil*. C.H. Beck.

Seher, G. (2016). Intelligente Agenten als 'Personen' im Strafrecht? In *Intelligente Agenten und das Recht* (Vol. 9).

Simmler, M., & Markwalder, N. (2019). Guilty Robots? Rethinking the Nature of Culpability and Legal Personhood in an Age of Artificial Intelligence. *Criminal Law Forum*(30).