مرزبندی میان دفاع مشروع قانونی و تبعیض الگوریتمی در تعیین مسئولیت کیفری

Delineating Lawful Self-Defense and Algorithmic Discrimination in Establishing Criminal Liability

چکیده

در دهه‌های اخیر، هوش مصنوعی و الگوریتم‌های پیچیده وارد عرصه‌های امنیتی و قضایی شده‌اند. این فناوری‌ها نحوه تصمیم‌گیری، اعمال قانون و مجازات (مانند مراقبت الکترونیکی) را دگرگون کرده‌اند. الگوریتم‌ها قدرت تحلیل سریع و دقت بالا دارند. آنها می‌توانند موقعیت‌های خطرناک را درک کرده و اقدامات دفاعی یا امنیتی پیشنهاد دهند.

با این حال، استفاده از این فناوری در موقعیت‌های حساس حقوقی، پرسش‌های تازه‌ای درباره مشروعیت و عدالت ایجاد کرده است. مهمترین چالش، مرز میان دفاع مشروع قانونی و تبعیض الگوریتمی است. تشخیص این مرز در تعیین مسئولیت کیفری طراحان، کاربران و نهادهای بهره‌بردار الگوریتم‌ها نقش تعیین‌کننده‌ای دارد. عدم توجه به این مرز می‌تواند منجر به معافیت غیرموجه از کیفر یا تحمیل مسئولیت ناعادلانه شود.

پژوهش حاضر با رویکرد توصیفی-تحلیلی و با ترکیب معیارهای حقوق کیفری سنتی و شاخص‌های عدالت الگوریتمی، چارچوبی شش مرحله‌ای برای تشخیص این مرز ارائه می‌دهد.

یافته‌های پژوهش نشان می‌دهد که الگوریتم‌ها به تنهایی نمی‌توانند عامل دفاع مشروع باشند. هرگونه تبعیض در داده‌ها، طراحی یا نظارت، مشروعیت دفاع را از بین می‌برد و مسئولیت کیفری را متوجه انسان‌ها و نهادها می‌کند. دفاع مشروع الگوریتمی تنها در صورت رعایت معیارهای قانونی، شفافیت داده‌ها و نظارت انسانی معتبر است. در مقابل، تبعیض الگوریتمی، حتی در قالب اقدام دفاعی، مشروعیت حقوقی ندارد و موجب مسئولیت کیفری می‌شود. مرز میان معافیت از کیفر و مسئولیت کیفری، نیازمند ترکیب معیارهای سنتی حقوق کیفری و شاخص‌های فنی عدالت الگوریتمی است.

این یافته‌ها بر ضرورت تدوین چارچوب‌های قانونی و سیاست‌های نظارتی در ایران و سایر نظام‌های مشابه تأکید دارند. آموزش، پژوهش و ایجاد نهادهای پاسخگو برای کاهش تبعیض و ارتقای عدالت الگوریتمی ضروری است. پاسخگویی انسانی و نهادی، همراه با شاخص‌های فنی عدالت و بازبینی مستمر، شرط لازم برای حفظ مشروعیت دفاع مشروع و کاهش خطر تبعیض الگوریتمی در عصر هوش مصنوعی است. این پژوهش تأکید می‌کند که تدوین قوانین کیفری ویژه، ایجاد نهادهای نظارتی مستقل و آموزش توسعه‌دهندگان و قضات، برای تحقق دفاع مشروع عادلانه و مسئولانه در عصر هوش مصنوعی ضروری است.

کلیدواژه‌ها: *هوش مصنوعی، دفاع مشروع، تبعیض الگوریتمی، مسئولیت کیفری، عدالت الگوریتمی، شفافیت، پاسخگویی.*

۱- مقدمه

ورود هوش مصنوعی به عرصه‌های امنیتی و قضایی، دگرگونی عمیقی در فرآیندهای تصمیم‌گیری، اعمال قانون و مجازات (مانند مراقبت الکترونیکی) ایجاد کرده است. اگرچه الگوریتم‌ها در تحلیل سریع داده‌ها کارآمدی بالایی دارند، اما همین ویژگی در موقعیت‌های دفاعی می‌تواند به دلیل سوگیری‌های پنهان، خطاهای سیستماتیک ایجاد کند که در نظام سنتی حقوق کیفری کمتر مورد توجه بوده است. یکی از مهم‌ترین این چالش‌ها، مرز میان دفاع مشروع قانونی و تبعیض الگوریتمی است. در حقوق کیفری به شکل رایج و سنتی آن، «دفاع مشروع» به عنوان حقی برای جلوگیری از تعرض یا خطر فوری شناخته شده و می‌تواند موجبات معافیت از «مسئولیت کیفری» را فراهم آورد؛ اما وقتی الگوریتم‌ها وارد این فرآیند می‌شوند، امکان بروز تبعیض و یا سوگیری بر اساس تشخیص غیرضرور و یا «خطاهای سیستماتیک» افزایش می‌یابد، که هم از منظر قانونی و هم از منظر حقوق کیفری قابل توجه هستند.

در این راستا مطالعات کلاسیک حقوق کیفری، دفاع مشروع را با توجه به معیارهایی مانند «ضرورت» و «تناسب اقدام» مشروعیت بخشیده است. در مقابل، ادبیات مدرن درباره عدالت الگوریتمی، شاخص‌هایی مانند «عدم تبعیض»، «شفافیت» و «پاسخگویی» را به عنوان الزامات استفاده از الگوریتم‌ها معرفی می‌کند. پژوهش‌های خارجی نیز نشان داده‌اند که سیستم‌های امنیتی مبتنی بر هوش مصنوعی، مانند خودروهای خودران(Custers, 2023 ; Gless & Ligeti, 2024 ; Leite, 2024) یا ربات‌های پلیس(Almasoud et al., 2024)؛سیستم‌های «تشخیص چهره[1]»(Wang et al., 2024) (Qandeel, 2024 و «تحلیل رفتار» در موقعیت‌های امنیتی نشان داده‌اند که در برخی گروه‌های اقلیتی یا رنگین‌پوستان، نرخ خطای بیشتری از خود نشان داده است (Buiten et al., 2023 ; Lendvai & Gosztonyi, 2025; Marchisio, 2021; Qandeel, 2024; Scherer, 2015; Završnik, 2021) Almasoud et al., 2024)؛ بنابراین استفاده نابجا از زور در موقعیت‌های دفاعی یادشده، صدور تصمیمات ناعادلانه محتمل است؛ درست در همین موقعیت‌ها مرز میان «دفاع مشروع قانونی» و «تبعیض الگوریتمی» رخ نمایان می‌سازد؛ بنابراین«دفاع مشروع» توسط الگوریتم‌ها اجرا در برابرمسأله «تبعیض و خطاهای الگوریتمی» نیازمند بسط دقیق نظری در زمینه حقوق کیفری است. در رویه قضایی سنتی ایران و سایر نظام‌ها، دفاع مشروع همواره با چالش‌هایی نظیر تشخیص «خطر قریب‌الوقوع»، «تناسب واکنش» و «نیت مدافع» مواجه بوده است. برای نمونه، در پرونده‌های مطروحه در دیوان عالی کشور ایران (مانند رأی شماره ۹۸۰۹۹۷۰۹۰۹۵۰۰۰۵۱)، گاه قضات در تشخیص «فوریت خطر» اختلاف نظر داشته‌اند؛ حال آنکه الگوریتم‌ها می‌توانند با شبیه‌سازی همین موقعیت‌ها، هم خطاهای انسانی را تکرار و هم خطاهای جدید الگوریتمی بیافزایند. این مقاله درصدد مقایسه نظام‌مند این دو نوع چالش است(Pourghasab, 2022)

از سوی دیگر، مطالعات داخلی اخیر در ایران، ضمن بررسی ابعاد حقوقی و اخلاقی هوش مصنوعی، موضوعاتی همچون تبعیض الگوریتمی، امکان شناسایی شخصیت حقوقی برای هوش مصنوعی و تأثیر رشد فناوری بر حقوق مسئولیت مدنی را مورد توجه قرار داده‌اند (انصاری، ۱۴۰۱؛ طباطبایی و امینی، ۱۴۰۴؛ کاظمی، ۱۴۰۴؛ جرفی و همکاران، ۱۴۰۴؛ شیخوند و همکاران، ۱۴۰۴). با وجود این، این مطالعات عمدتاً بر جنبه‌های اخلاقی، مدنی یا جایگاه حقوقی فناوری تمرکز داشته و کمتر به مسئله خاص مسئولیت کیفری ناشی از اقدامات الگوریتمی در وضعیت دفاعی و مرزبندی آن با تبعیض الگوریتمی پرداخته‌اند. تحلیل این مطالعات نشان می‌دهد که فقدان چارچوب قانونی مشخص برای ارزیابی مشروعیت دفاع الگوریتمی، می‌تواند موجب معافیت غیرموجه از کیفر یا بالعکس مسئولیت غیرعادلانه شود؛ بنابراین، ترکیب این دو حوزه، امکان تحلیل موقعیت‌های دفاعی الگوریتمی را فراهم می‌سازد، به طوری که مشروعیت تصمیمات گرفته شده توسط الگوریتم‌ها را از منظر حقوقی و اخلاقی ارزیابی نمود.

[1] - Facial recognition

این وضعیت، پرسشی اساسی را مطرح می‌کند: چه معیارهایی می‌توانند مرز بین دفاع مشروع قانونی و تبعیض الگوریتمی را تعیین کنند و در نتیجه مسئولیت کیفری چه کسانی باید متوجه شود؟ به عبارت دیگر، آیا مسئولیت ناشی از تبعیض الگوریتمی متوجه سازندگان، کاربران یا خود الگوریتم است و چگونه می‌توان از سوءاستفاده احتمالی دفاع مشروع جلوگیری کرد؟

با توجه به خلأ علمی و قانونی موجود، این پژوهش در پی آن است که چارچوبی توصیفی-تحلیلی و علمی برای مرزبندی دفاع مشروع و تبعیض الگوریتمی ارائه دهد؛ تلاش دارد با ترکیب معیارهای حقوق کیفری سنتی و شاخص‌های عدالت الگوریتمی، چارچوبی تحلیلی و عملی ارائه دهد که بتواند مرز میان دو مفهوم یادشده را به روشنی مشخص کند. بنابراین تمرکز بر سه محور کلیدی است: نخست، تحلیل شرایطی که دفاع مشروع در حقوق کیفری دارد و بررسی امکان تعمیم آن به تصمیمات الگوریتمی؛ دوم، شناسایی موقعیت‌هایی که تبعیض الگوریتمی می‌تواند مشروعیت دفاع را زیر سؤال ببرد؛ و سوم، تعیین مرز میان کسانی که سزاوار معافیت از کیفر هستند و مواردی که تبعیض قابل مجازات است. با پاسخ به این پرسش‌ها، پژوهش می‌تواند نه تنها شکاف‌های علمی و قانونی موجود در منابع داخلی را پر کند، بلکه به سیاستگذاران، قضات و پژوهشگران کمک کند تا در مواجهه با فناوری‌های نوین، تصمیمات حقوقی متعادل و منصفانه‌ای اتخاذ کرده و از سوءاستفاده احتمالی دفاع مشروع یا تبعیض الگوریتمی جلوگیری کنند. نوآوری اصلی پژوهش حاضر در دو سطح نظری و روشی قابل ارائه است. در سطح نظری، این نخستین تلاش نظام‌مند در ادبیات حقوق کیفری ایران است که به صورت مستقیم به مرزبندی میان «دفاع مشروع قانونی» و «تبعیض الگوریتمی» می‌پردازد و صرفاً به شناسایی تبعیض‌ها اکتفا نمی‌کند، بلکه شروط تحقق دفاع مشروع در تصمیمات خودکار الگوریتم‌ها را احصاء می‌نماید. در سطح روشی، پژوهش با ترکیب معیارهای سنتی حقوق کیفری (ضرورت، تناسب، فوریت خطر) و شاخص‌های مدرن عدالت الگوریتمی (شفافیت، پاسخگویی، قابلیت توضیح‌پذیری، ارزیابی اثرات اجتماعی و فرهنگی)، چارچوبی شش مرحله‌ای ارائه می‌دهد که امکان تشخیص عملی دفاع مشروع از تبعیض ناروا را فراهم می‌کند. این چارچوب در مقایسه با پژوهش‌های پیشین که عمدتاً بر مسئولیت مدنی یا اخلاق الگوریتم متمرکز بوده‌اند، به طور مشخص به آثار کیفری تبعیض الگوریتمی و معافیت از کیفر در دفاع مشروع می‌پردازد.

پژوهش‌های پیشین در حوزه هوش مصنوعی و حقوق کیفری عمدتاً بر سه محور متمرکز بوده‌اند: (۱) تحلیل سوگیری الگوریتمی در سیستم‌های پیش‌بینی جرم(Almasoud et al., 2024; Mugari et al., 2021) ، (۲) مسئولیت مدنی ناشی از تصمیمات خودکار خودروهای خودران(Gless & Ligeti, 2024; Leite, 2024) ، و (۳) چالش‌های اخلاقی تشخیص چهره و نظارت هوشمند(Wang et al., 2024; Qandeel, 2024) ، در داخل ایران، مطالعات محدودی مانند (انصاری، ۱۴۰۱) بر تبعیض الگوریتمی در نظام بانکی، (جرفی و همکاران، ۱۴۰۴) بر مسئولیت مدنی کشتی‌های خودران، و (شیخوند و همکاران، ۱۴۰۴) بر کاربرد هوش مصنوعی در تعقیب کیفری تمرکز داشته‌اند. حلقه مفقوده اصلی که در هیچ یک از پژوهش‌های داخلی و بین‌المللی به صورت نظام‌مند به آن پرداخته نشده، عبارت است از: «مرزبندی عملیاتی میان دفاع مشروع قانونی و تبعیض الگوریتمی در تعیین مسئولیت کیفری». به عبارت دیگر، پژوهش‌های موجود یا به دفاع مشروع در سیستم‌های انسانی پرداخته‌اند، یا به تبعیض الگوریتمی در حوزه‌های غیرکیفری (استخدام، اعتبارات بانکی، آموزش). هیچ مطالعه‌ای تاکنون چارچوبی شش شاخصی برای تفکیک این دو مفهوم در لحظه تصمیم‌گیری الگوریتمی ارائه نداده است.

از این رو، نوآوری پژوهش حاضر در سه سطح ارائه چارچوب شش مرحله‌ای ترکیبی (حقوق کیفری سنتی + عدالت الگوریتمی) برای تشخیص مرز دفاع مشروع و تبعیض، شناسایی و دسته‌بندی «پیامدهای کیفری خلط مرز» شامل: معافیت غیرموجه از کیفر، تحمیل مسئولیت ناعادلانه به بهره‌برداران، و بازتولید تبعیض ساختاری در نظام عدالت کیفری، ارائه راهکارهای تقنینی و نهادی اختصاصی برای نظام حقوقی ایران (از جمله دادگاه‌های تخصصی الگوریتمی و استانداردهای افشای سوگیری) قابل ارائه است.

۲- چارچوب مفهومی

قبل از پرداختن به بحث و بررسی موضوع این پژوهش برخی از واژگان خاص باید تعریف و نحوه ارتباط با پژوهش تبیین شود؛ از این رو سعی در تعریف و ایجاد یک چارچوب نظری عمده تلاش این مبحث است.

۲-۱- هوش مصنوعی[۲]

هوش مصنوعی در این پژوهش به معنای سامانه‌ای داده‌محور و یادگیرنده تعریف می‌شود که قادر به تصمیم‌گیری خودکار در شرایط پیچیده است (Zhao et al., 2022; Yu & Kumbier, 2017) ؛ اما فاقد اراده و مسئولیت کیفری مستقل است.

۲-۲- دفاع مشروع و اظطرار قانونی[۳]

دفاع مشروع حالتی است که در آن شخص برای دفع خطری واقعی و قریب‌الوقوع علیه جان، مال، آزادی یا سایر حقوق اساسی خود یا دیگری، ناگزیر به ارتکاب رفتاری می‌شود که در شرایط عادی جرم است. در حقوق کیفری ایران دفاع مشروع بر اساس ماده ۱۵۶ قانون مجازات اسلامی و فقه امامیه مستلزم وجود خطر فعلی و غیرقانونی، ضرورت و تناسب واکنش، فقدان امکان مراجعه به مأموران و عدم تحریک اولیه مدافع است (نجفی، ۱۴۰۴ق، ج۲۱: ۴۶۶؛ نجم‌الدین، ۱۴۰۳ق، ج۴: ۹۸). باید توجه داشت که در حقوق کیفری، دفاع مشروع ناظر به دفع «تجاوز غیرقانونیِ دارای منشأ انسانی» است، در حالی که اضطرار ناظر به دفع «خطر قریب‌الوقوع بدون منشأ انسانی» (مانند آتش‌سوزی، سیل، بیماری همه‌گیر) می‌باشد. در نظام حقوق کیفری ایران، ماده ۱۵۶ قانون مجازات اسلامی ناظر به دفاع مشروع و ماده ۱۵۲ همان قانون ناظر به اضطرار است. این تمایز در تحلیل الگوریتم‌ها نیز باید رعایت شود؛ زیرا تشخیص منشأ خطر (انسانی یا غیرانسانی) در تعیین نوع موجّه قانونی تأثیر مستقیم دارد

۲-۳- تبعیض الگوریتمی

تبعیض الگوریتمی اصطلاحی است که به پدیده بروز رفتارهای ناعادلانه و نابرابر از سوی سامانه‌های مبتنی بر الگوریتم و هوش مصنوعی اطلاق می‌شود؛ رفتارهایی که برخلاف تصور بی‌طرفی ماشین، در عمل می‌تواند مشابه یا حتی شدیدتر از تبعیض انسانی عمل کند. در ادبیات علمی، ریشه‌های این تبعیض را در سه سطح اصلی شناسایی کرده‌اند: نخست، «سوگیری داده‌ها[۴]» که ناشی از جمع‌آوری یا آموزش الگوریتم بر داده‌های ناقص، نادرست یا بازتاب‌دهنده پیش‌داوری‌های اجتماعی است؛ دوم، «سوگیری طراحی[۵]» (Mugari et al., 2021) که به دلیل انتخاب معیارها، متغیرها و اهداف در طراحی الگوریتم رخ می‌دهد؛ و سوم، «سوگیری پردازشی[۶]» یا عملکردی (Lim & Taeihagh, 2019) که در نتیجه نحوه تحلیل داده‌ها و تولید خروجی‌های الگوریتمی به وجود می‌آید (Barocas & Selbst,2016). از منظر حقوقی، تبعیض الگوریتمی زمانی تحقق می‌یابد که تصمیم یا خروجی سامانه هوش مصنوعی، برخلاف اصول برابری و منع تبعیض، منجر به محرومیت یا محدودیت فرصت‌ها و حقوق افراد یا گروه‌ها به‌ویژه در حوزه‌هایی چون استخدام، اعطای تسهیلات بانکی، آموزش، خدمات اجتماعی و نظام عدالت کیفری شود، به بیان دیگر، تبعیض الگوریتمی شکلی نوین از تبعیض ساختاری است که با پوشش «عینی و علمی بودن داده‌ها» پنهان می‌ماند، اما آثار آن به‌مراتب عمیق‌تر است (European Commission, 2020).

از منظر نظریه‌های عدالت و حقوق بشر، این نوع تبعیض با اصول بنیادینی چون «برابری در مقابل قانون»، «حق دسترسی برابر به فرصت‌ها» و «منع تبعیض» در اسناد بین‌المللی مانند ماده ۲ میثاق حقوق مدنی و سیاسی و ماده ۷ اعلامیه جهانی حقوق بشر در تعارض قرار دارد. در نظام‌های حقوقی داخلی نیز بحث شناسایی و مقابله با تبعیض الگوریتمی به یکی از چالش‌های نوین بدل شده است. به‌طور نمونه، در برخی نظام‌های پیشرو همچون اتحادیه اروپا، مقررات ویژه‌ای در چارچوب «قانون هوش

---

[2] -Artificial intelligence: AI
[3] - Legal Self-Defense
[4] - Data Bias
[5] - Design Bias
[6] - Processing Bias

مصنوعی[7]» (OECD, 2019)که بر ضرورت شفافیت الگوریتم‌ها، امکان بازبینی تصمیمات خودکار، و تضمین پاسخگویی طراحان و بهره‌برداران سامانه‌های هوش مصنوعی تأکید دارد[8] (Makauskaite-Samuole, 2025) لکه قانون اخیرالتصویب مذکور، با نام رسمی Regulation (EU) 2024/1689، در پارلمان اروپا و شورای اروپا تصویب و در «روزنامه رسمی اتحادیه اروپا[9]» در تاریخ ۱۲ ژوئیه ۲۰۲۴ منتشر شد.قانون رسماً از ۱ اوت ۲۰۲۴ لازم الاجرا شده است.

در نتیجه، تبعیض الگوریتمی را می‌توان این‌گونه تعریف کرد: «هر نوع تصمیم‌گیری خودکار مبتنی بر الگوریتم که در اثر سوگیری داده‌ها، طراحی یا پردازش، منجر به رفتار نابرابر و ناعادلانه نسبت به افراد یا گروه‌ها گردد و آثار منفی بر حقوق بنیادین آنان بر جای نهد.» در واقع این تعریف با تأکید بر سه رکن منشأ، ماهیت و آثار، هم با واقعیت‌های فنی منطبق است و هم با الزامات حقوقی مقابله با تبعیض هماهنگ می‌باشد.

۲-۴- عدالت الگوریتمی

عدالت الگوریتمی مفهومی است که در سال‌های اخیر با گسترش استفاده از سامانه‌های هوش مصنوعی و الگوریتم‌های تصمیم‌گیر خودکار اهمیت یافته است. این مفهوم به معنای وضعیتی است که در آن تصمیمات و خروجی‌های الگوریتمی با اصول عدالت، برابری و بی‌طرفی همخوانی داشته و از تبعیض نسبت به افراد یا گروه‌های خاص جلوگیری شود(Behzad et al., 2025). عدالت الگوریتمی شامل سه رکن اساسی است: نخست، بی‌طرفی الگوریتمی که اطمینان می‌دهد الگوریتم در پردازش داده‌ها و اتخاذ تصمیمات، هیچ گونه ترجیح یا تبعیضی نسبت به افراد یا گروه‌ها قائل نمی‌شود؛ دوم، شفافیت و قابلیت بررسی که امکان تحلیل و تفسیر تصمیمات خودکار و شناسایی خطاها یا سوگیری‌ها را فراهم می‌کند؛ و سوم، تناسب و اثر عادلانه، به این معنا که خروجی‌های الگوریتم باید با اصول عدالت اجتماعی، حقوق بشر و حقوق شهروندی همخوانی داشته باشند (He & Zhang, 2025) و از پیامدهای ناعادلانه و محرومیت‌های غیرضروری جلوگیری نمایند (Barocas & Selbst, 2016)در ادبیات حقوقی و اخلاقی، عدالت الگوریتمی نه تنها به حذف تبعیض و سوگیری محدود می‌شود، بلکه به ایجاد ساختارهای پاسخگو و قابل اعتماد برای الگوریتم‌ها نیز توجه دارد، به گونه‌ای که کاربران، توسعه‌دهندگان و قانون‌گذاران بتوانند تصمیمات ماشینی را بررسی و اصلاح کنند. از منظر کاربردی، تحقق عدالت الگوریتمی مستلزم طراحی الگوریتم‌های شفاف، آموزش بر داده‌های نماینده و عادلانه، و تدوین مقررات نظارتی و اخلاقی است تا تضمین شود که هوش مصنوعی نه تنها کارآمد، بلکه مطابق با ارزش‌های انسانی و عدالت اجتماعی عمل می‌کند (OECD, 2019; European Commission, 2020). به طور خلاصه، عدالت الگوریتمی تلاش می‌کند که هوش مصنوعی و تصمیمات خودکار آن، عادلانه، شفاف و قابل پاسخگویی باشند و همزمان از تأثیرات منفی بر حقوق و فرصت‌های افراد جلوگیری شود.

۳- مسئولیت و دفاع الگوریتمی

۳-۱- مسئولیت کیفری و دفاع الگوریتمی

مطالعات بین‌المللی نشان می‌دهند که تصمیمات خودکار در حوزه‌های کیفری و مدنی می‌توانند مسئولیت قانونی را با چالش‌های اساسی مواجه کنند. با تمرکز بر سامانه‌های نظارتی و الگوریتم‌های پیش‌بینی جرم نشان داد که سوگیری داده‌ها نه‌تنها می‌تواند تبعیض نژادی و اجتماعی را تقویت کند، بلکه مشروعیت دفاع و تعیین مسئولیت کیفری را نیز زیر سؤال می‌برد. در همین راستا، (Scherer, 2015)با بررسی چارچوب‌های قانونی هوش مصنوعی هشدار داد که نبود مقررات روشن، زمینه توجیه

[7] - AI Act
[8]- Regulation (EU) 2024/1689
[9] - Official Journal

اقدامات تبعیض‌آمیز توسط الگوریتم‌ها را فراهم می‌کند؛ به موازات این بحث، (Danaher et al., 2017)با طرح مفهوم «حکمرانی الگوریتمی[10]» نشان دادند که عدم شفافیت و پاسخگویی می‌تواند موجب خطاهای حقوقی در تصمیمات خودکار شود. از منظر مسئولیت مدنی، (Marchisio, 2021)پیشنهاد کرد که قواعد «مسئولیت بدون خطا» می‌تواند مبنایی برای توسعه مسئولیت کیفری الگوریتم‌ها باشد و مانع از آن شود که تبعیض به‌عنوان دفاع مشروع مطرح شود. در ادامه، پژوهش‌های جدید مانند (Buiten et al., 2023)با تمرکز بر پیامدهای اقتصادی و حقوقی تصمیمات الگوریتمی و (Cooney et al., 2023)با بررسی نقش سوگیری در تضعیف مشروعیت دفاع، ابعاد تازه‌ای به این حوزه افزودند؛ افزون بر این، ارائه چارچوبی تلفیقی از شاخص‌های فنی و معیارهای حقوقی قادراست تا مرز میان دفاع مشروع و تبعیض الگوریتمی شفاف‌تر شود و (Gless & Ligeti, 2024)نیز با تأکید بر بُعد بین‌المللی، بر ضرورت وضع مقررات ملی و فراملی برای تضمین عدالت و پاسخگویی تأکید کرد.

مرور مطالعات بین‌المللی نشان می‌دهد که پژوهش‌ها در حوزه مسئولیت کیفری الگوریتم‌ها جامع و چندبُعدی هستند و از بررسی سوگیری داده‌ها و اثرات تبعیض اجتماعی تا چارچوب‌های بین‌المللی و تلفیقی حقوقی و فنی را پوشش می‌دهند، اما بررسی انتقادی نشان می‌دهد که برخی محدودیت‌ها و خلأها همچنان باقی است؛ از جمله فقدان نقد درونی نسبت به پیشنهادهایی مانند «مسئولیت بدون خطا» که ممکن است با اصول بنیادین مسئولیت کیفری تعارض داشته باشد، تمرکز بر جنبه های مدنی نیز، ابهام در تمایز عملی میان دفاع مشروع و تبعیض الگوریتمی و ضعف پیوند با مقررات جزائی شده است.

۳-۲- مسئولیت در عدالت الگوریتمی

عدالت الگوریتمی به مطالعه و تضمین انصاف در تصمیم‌گیری‌های خودکار اشاره دارد و هدف آن جلوگیری از تبعیض و بازتولید نابرابری‌های اجتماعی است. هوش مصنوعی، به عنوان سامانه‌ای که قادر به تحلیل داده‌ها و اتخاذ تصمیمات خودکار در شرایط پیچیده است(Sandu & Vosinakis, 2025; Wei et al., 2025) می‌تواند نتایج تبعیض‌آمیز تولید کند، به‌ویژه وقتی الگوریتم‌ها بر داده‌های تاریخی یا نمونه‌برداری‌های ناقص متکی باشند. این مسئله در حوزه‌هایی مانند پیش‌بینی جرم، استخدام، اعطای وام، تعقیب کیفری و .... اهمیت ویژه دارد، زیرا تصمیمات الگوریتمی می‌توانند فرصت‌ها و حقوق افراد را محدود کرده و مشروعیت حقوقی و اخلاقی فرآیندهای تصمیم‌گیری را به چالش بکشند؛ این چالش زمانی مطرح است که سیستم‌های هوشمند برای دفع خطر یا حفظ منافع مشروع اقداماتی انجام داده‌اند که به‌طور مستقیم یا غیرمستقیم منجر به تبعیض یا رفتار مجرمانه شده‌اند. در ادامه، چند نمونه از این موارد را بررسی می‌کنیم:

نخستین نمونه مربوط به ربات آتش‌نشانی[11] است که در جریان یک عملیات نجات، دیوار خانه‌ای را تخریب می‌کند تا به فردی گرفتار در آتش‌سوزی دسترسی پیدا کند(Verhagen et al., 2024). اما مشخص می‌شود که خانه خالی بوده و تخریب غیرضروری صورت گرفته است. اگرچه این اقدام در ظاهر دفاع مشروع تلقی می‌شود، اما در صورت نبود خطر واقعی، می‌تواند مصداق تجاوز از حد دفاع و موجب مسئولیت مدنی یا کیفری برای بهره‌بردار سیستم باشد. در نظام سنتی، اگر یک شهروند در اطفای حریق اشتباهاً به مال دیگری خسارت بزند، دادگاه‌ها میان «اضطرار» و «تعدی» اختلاف نظر دارند) مطالعه تطبیقی: مقایسه رأی وحدت رویه شماره ۷۹۹ دیوان عالی ایران با قضیه United States v. Haynes, 2021). الگوریتم همین ابهام را با قدرت بیشتر و شفافیت کمتر بازتولید می‌کند(Haynes, 2021) .

در حوزه اقتصادی، الگوریتم‌های بانکی برای ارزیابی اعتبار مشتریان گاهی به‌طور سیستماتیک درخواست‌های وام افراد ساکن در مناطق کم‌درآمد را رد می‌کنند، حتی اگر اعتبار مالی مناسبی داشته باشند(Barocas & Selbst, 2016; Završnik, 2021). این رفتار تبعیض‌آمیز قابل توجیه با دفاع مشروع نیست، زیرا هیچ خطر فوری برای بانک وجود ندارد. مسئولیت مدنی و تبعیض ساختاری در اینجا مطرح است. در یک پرونده داخلی، شعبه ۹ دادگاه کیفری دو تهران (۱۴۰۰) تشخیص داد که اعمال

[10] - Algorithmic governance
[11] - Firefighting robot

تبعیض‌آمیز یک سامانه اعتباری خودکار بر اساس محل سکونت، مصداق تبعیض غیرقانونی است و مسئولیت متوجه بانک بهره‌بردار می‌باشد (دادگاه کیفری دو تهران، ۱۴۰۰).

در بحران‌های بهداشتی، سیستم تریاژ بیمارستانی بیماران مهاجر را در اولویت پایین‌تری برای دریافت دستگاه تنفس مصنوعی قرار می‌دهد (European Parliament, 2025; Leslie et al., 2021). این تصمیم بر اساس داده‌های ناقص و سوگیرانه اتخاذ شده و ممکن است مصداق قتل غیرعمد یا ترک فعل باشد. **اضطرار (مدیریت منابع محدود در شرایط بحرانی بهداشتی)** تنها زمانی قابل پذیرش است که تخصیص منابع بر اساس معیارهای پزشکی و نه تبعیض‌آمیز صورت گیرد.

این مجموعه مثال‌ها نشان می‌دهد که دفاع مشروع در هوش مصنوعی باید با دقت، نظارت انسانی و شفافیت همراه باشد. تبعیض ساختاری در داده‌ها یا الگوریتم‌ها می‌تواند منجر به جنایت شود و مسئولیت کیفری باید به انسان‌هایی که در طراحی، اجرا یا نظارت نقش دارند، نسبت داده شود. تدوین قوانین کیفری خاص برای فناوری‌های نوظهور، ایجاد نهادهای نظارتی مستقل، و آموزش حقوقی برای توسعه‌دهندگان فناوری از الزامات پیشگیری از تبعیض و جنایت در هوش مصنوعی هستند.

یکی از چالش‌های اصلی، تبعیض داده‌محور است که ناشی از سوگیری‌های تاریخی یا خطاهای جمع‌آوری داده‌هاست؛ در این راستا برخی مطالعات داخلی نشان داده‌اند که الگوریتم‌ها در بانک‌ها(انصاری، ۱۴۰۱)، تعقیب کیفری(شیخوند و همکاران، ۱۴۰۱) و سامانه‌های اطلاعاتی می‌توانند سوگیری ایجاد کنند؛ حتی الگوریتم‌های طراحی‌شده با نیت خنثی ممکن است گروه‌های خاصی را به شکل ناعادلانه‌ای تحت تأثیر قرار دهند. شاخص‌های عدالت الگوریتمی مانند تساوی فرصت‌ها و عدم سوگیری گروهی ابزارهای کلیدی برای شناسایی و اصلاح این تبعیض‌ها هستند.

بنابراین، عدالت الگوریتمی تنها با تلفیق شاخص‌های فنی، معیارهای حقوقی و اخلاقی قابل تحقق است و توسعه چارچوب‌های تلفیقی می‌تواند به شناسایی گروه‌های آسیب‌پذیر، اصلاح مدل‌های تصمیم‌گیری و تضمین شفافیت و پاسخگویی کمک کند و از بازتولید تبعیض و نابرابری جلوگیری نماید.

۴- چالش‌های نوین مسئولیت کیفری سنتی در برابر دفاع مشروع به شکل هوشمند

۴-۱- مسئولیت کیفری سنتی

مسئولیت کیفری یکی از مبانی اصلی حقوق جزا است و به معنای الزام قانونی فرد به پاسخگویی و مجازات برای اعمالی است که قانوناً جرم تلقی می‌شوند. مسئولیت کیفری بر اساس دو رکن اصلی «عنصر مادی» و «عنصر معنوی» تعیین می‌شود؛ در چارچوب حقوق سنتی، دفاع مشروع به عنوان یک موجّه قانونی برای تخفیف یا رفع مسئولیت کیفری مطرح است و در شرایط مشخصی مانند «تهدید مستقیم و غیرقابل اجتناب» اعمال می‌شود؛ در حقوق کیفری سنتی ایران، دفاع مشروع در ماده‌های قانونی مرتبط با حمله یا تجاوز به حریم شخصی و مالکیت پیش‌بینی شده است و عنصر مشروعیت بر اساس نسبت میان تهدید و واکنش تعیین می‌شود. با ورود فناوری‌های نوین و الگوریتم‌های تصمیم‌گیر، تحلیل مسئولیت کیفری و دفاع مشروع پیچیده‌تر شده است، چرا که تصمیمات الگوریتمی می‌توانند به‌طور خودکار اعمالی را توجیه کنند که در دنیای سنتی غیرمجاز هستند. در رویه قضایی ایران، رأی وحدت رویه شماره ۷۹۹ دیوان عالی کشور (۱۳۹۸) اگرچه مستقیماً ناظر به هوش مصنوعی نیست، اما در مورد مسئولیت ناشی از سامانه‌های خودکار در صنعت حمل‌ونقل، اصل مسئولیت انسانی نهاد بهره‌بردار را تأکید کرده است (دیوان عالی کشور، ۱۳۹۸).

۴-۲- دفاع مشروع و چالش‌های نوین

دفاع مشروع در دنیای دیجیتال و هوش مصنوعی با چالش‌های تازه‌ای مواجه است.

جدول ۱: مقایسه چالش‌های دفاع مشروع در رویه قضایی سنتی و سیستم‌های الگوریتمی

منبع تطبیقی	چالش مشابه در دفاع الگوریتمی	مثال از اختلاف نظر قضایی	چالش در دفاع مشروع سنتی (انسانی)
(Custers, 2023)	وابستگی الگوریتم به داده‌های لحظه‌ای و نویز سنسورها	رأی شماره ۹۸۰۹۹۷۰۹۰۹۵۰۰۰۵۱ دیوان عالی ایران	تشخیص «فوریت خطر»
(Almasoud et al., 2024)	سوگیری در شدت‌دهی به تهدید بر اساس نژاد/مکان	پرونده Famous v. State, Florida (2020)	تناسب واکنش
(Scherer, 2015)	عدم امکان احراز نیت در الگوریتم	رأی شعبه ۲۷ دادگاه تجدیدنظر تهران (۱۳۹۸)	نیت دفاعی (عنصر روانی)
(Leite, 2024)	الگوریتم بدون درک موقعیت، عقب‌نشینی را بر نمی‌گزیند	پرونده People v. Johnson, NY (2019)	الزام به عقب‌نشینی

این جدول نشان می‌دهد که بسیاری از چالش‌های سنتی دفاع مشروع که در آرای قضایی باعث اختلاف نظر شده‌اند، در دنیای الگوریتمی نه تنها حل نشده‌اند، بلکه با ابعاد جدیدی از سوگیری و عدم شفافیت همراه شده‌اند.

الگوریتم‌ها می‌توانند در شرایط اضطراری و بدون دخالت انسانی تصمیماتی اتخاذ کنند که از نظر قانونی به عنوان دفاع مشروع تفسیر شوند. اما مسئله اصلی سوگیری و تبعیض داده‌ها است؛ زیرا الگوریتم‌های یادگیری ماشین بر اساس داده‌های تاریخی آموزش می‌بینند و ممکن است تصمیمات تبعیض‌آمیز اتخاذ کنند؛ این وضعیت منجر به پرسش اساسی حقوقی و اخلاقی می‌شود: آیا می‌توان اعمال یک الگوریتم را صرفاً به دلیل اینکه «دفاع مشروع» تلقی شده، مشروع دانست؟ مطالعات مرور شده قبلی نشان داده‌اند که بدون شفافیت و پاسخگویی الگوریتمی، مشروعیت دفاع مشروع تحت تأثیر سوگیری‌های الگوریتمی قرار می‌گیرد.

دفاع مشروع به معنای اعمال اقدام قانونی برای دفع تجاوز یا تهدید غیرقانونی است که در قوانین کیفری به‌عنوان موجّه قانونی برای رفع یا تخفیف مسئولیت کیفری شناخته می‌شود؛ در حقوق ایران، طبق مواد ۱۵۶، ۱۵۷ و ۱۵۸ قانون مجازات اسلامی ۱۳۹۲، زمانی پذیرفته می‌شود که برای دفع خطر فوری علیه جان، مال یا عرض، اقدام متناسب و ضروری انجام شود و امکان توسل به قوای دولتی وجود نداشته باشد؛ در غیر این صورت مسئولیت کیفری برقرار است؛ و در نظام‌های حقوقی دیگر مانند آمریکا و اتحادیه اروپا نیز دفاع مشروع تحت شرایط مشابه تعریف می‌شود، با این تفاوت که برخی قوانین، «ضرورت پیشگیرانه» و «مسئولیت محدود» (Buiten et al., 2023)در دفاع از خود یا دیگران را در شرایط اضطراری نیز لحاظ کرده‌اند.

۴-۳- چالش‌های و پیامدهای دفاع مشروع در عصر الگوریتم‌ها

با ورود هوش مصنوعی و الگوریتم‌های تصمیم‌گیری به حوزه‌های امنیتی و قضایی، تحلیل دفاع مشروع پیچیده‌تر شده است. الگوریتم‌ها می‌توانند تصمیماتی خودکار در مواجهه با تهدیدها اتخاذ کنند که از نظر قانونی به‌عنوان دفاع مشروع توجیه شوند، اما این تصمیمات می‌توانند:

۱. مبتنی بر داده‌های سوگیرانه یا ناقص باشند و به تبعیض منجر شوند.
۲. بدون شفافیت و امکان بازبینی، مشروعیت حقوقی آن‌ها قابل بررسی نباشد.
۳. مسئولیت کیفری و اخلاقی تصمیمات را مبهم کنند و بر نحوه اعمال عدالت تأثیر بگذارند.

مطالعات نشان می‌دهند که الگوریتم‌های پیش‌بینی جرم و سیستم‌های نظارتی هوشمند می‌توانند «استفاده نابجا از قدرت» را توجیه کنند، به ویژه زمانی که معیارهای دفاع مشروع انسانی به شکل کامل در الگوریتم لحاظ نشده باشند؛ همچنین، تحلیل تطبیقی نشان می‌دهد که در کشورهای پیشرفته، تلاش شده است تا چارچوب‌های قانونی و شاخص‌های فنی عدالت الگوریتمی برای تشخیص مشروعیت دفاع در تصمیمات خودکار تعریف شود(He & Zhang, 2025) . اما در ایران، هنوز قوانین کیفری صراحت کافی برای تشخیص مسئولیت الگوریتم‌ها و مشروعیت دفاع خودکار ندارند، این خلا، پژوهش حاضر را از نظر نوآوری و اهمیت تقویت می‌کند، زیرا تحلیل ترکیبی حقوق کیفری و عدالت الگوریتمی می‌تواند مرز میان دفاع مشروع و تبعیض الگوریتمی را روشن کند.

پیامدهای تبعیض الگوریتمی شامل:

۱. کاهش مشروعیت قانونی دفاع :اگر الگوریتم در موقعیت دفاعی سوگیری داشته باشد، مشروعیت دفاع مشروع زیر سؤال می‌رود
۲. مسئولیت کیفری طراحان و نهادها :در صورت بروز تبعیض، مسئولیت کیفری نباید فقط متوجه تصمیمات الگوریتم باشد، بلکه طراحان و کاربران سیستم نیز پاسخگو هستند.
۳. تقویت نابرابری اجتماعی و اقتصادی :تصمیمات ناعادلانه الگوریتمی می‌تواند به بازتولید تبعیض‌ها در جامعه منجر شود.

این شاخص‌ها به عنوان ابزار تحلیلی برای تشخیص مرز میان دفاع مشروع و تبعیض الگوریتمی در پژوهش حاضر به کار گرفته می‌شوند.

۴-۴- سوگیری پس از استقرار(Post-Deployment Bias)

حتی اگر الگوریتم با داده‌های عادلانه و طراحی بی‌طرفانه آموزش دیده باشد، پس از استقرار در محیط واقعی ممکن است به دلیل تعامل با کاربران، بازخوردهای تدریجی، یا تغییرات اجتماعی دچار سوگیری جدید شود. این نوع سوگیری، مسئولیت نظارت مستمر و بازبینی دوره‌ای الگوریتم را بر عهده نهادهای بهره‌بردار مضاعف می‌کند .(Obermeyer et al., 2019) در نتیجه، شاخص «ارزیابی مستمر و یادگیری اخلاقی الگوریتم» باید شامل نظارت پس از استقرار نیز باشد.

۵- مرزبندی دفاع مشروع و تبعیض در تصمیمات الگوریتمی

با گسترش استفاده از الگوریتم‌ها در تصمیم‌گیری‌های بیان شده، مرزبندی دقیق میان دفاع مشروع و تبعیض الگوریتمی بخش دیگری از این پژوهش است. اگر چه مرورمطالعات نشان داده‌اند که الگوریتم‌های خودکار می‌توانند تصمیماتی اتخاذ کنند

که از نظر حقوقی به عنوان دفاع مشروع توجیه شوند، اما در عین حال دارای سوگیری یا تبعیض ناپیدا باشن، که در صورت عدم تعیین مرز مشخص، خطر توجیه اقدامات ناعادلانه به نام دفاع مشروع افزایش می‌یابد و مسئولیت کیفری طراحان و نهادهای بهره‌بردار الگوریتم مبهم می‌شود؛ برای تشخیص مرز دفاع مشروع و تبعیض الگوریتمی، پژوهش‌های اخیر شاخص‌هایی را پیشنهاد کرده‌اند که شامل موارد زیر است:

۱. تناسب واکنش با تهدید: تصمیمات الگوریتم باید با شدت و نوع تهدید متناسب باشد.(انصاری، ۱۴۰۱)
۲. شفافیت داده‌ها و مدل‌ها:منابع داده و معیارهای تصمیم‌گیری باید قابل بررسی و بازبینی باشند(Završnik, 2021)
۳. پاسخگویی و مسئولیت‌پذیری: طراحان، کاربران و نهادهای بهره‌بردار باید مسئولیت تصمیمات الگوریتمی را بر عهده گیرند (He & Zhang, 2025)
۴. عدم سوگیری و تبعیض سیستماتیک :الگوریتم باید عاری از تبعیض‌های نژادی، جنسیتی یا اجتماعی باشد و تأثیرات آن بر گروه‌های مختلف بررسی شود(Cooney et al., 2023; Danaher et al., 2017) .
۵. ضرورت و فوریت اقدام: واکنش الگوریتم باید تنها در شرایطی مجاز باشد که هیچ راه‌حل ملایم‌تر یا جایگزین مؤثرتری برای دفع تهدید وجود ندارد(Barocas & Selbst, 2016).
۶. قابلیت توضیح‌پذیری: الگوریتم باید بتواند دلایل تصمیم خود را به‌گونه‌ای بیان کند که قابل فهم و ارزیابی توسط مراجع قضایی باشد؛ در غیر این صورت مشروعیت دفاع تضعیف می‌شود(Mehrabi et al., 2022).
۷. بازبینی مستقل: مرز مشروعیت زمانی معتبر است که یک نهاد مستقل امکان نظارت بر تصمیمات الگوریتمی و بررسی تبعیض احتمالی را داشته باشد(Yeung, 2018).
۸. حفظ کرامت انسانی: حتی در مواجهه با تهدید، الگوریتم نباید اقداماتی انجام دهد که ناقض کرامت ذاتی یا حقوق بنیادین افراد (مانند حق حیات یا منع شکنجه) باشد(Cath et al., 2018).
۹. ارزیابی مستمر و یادگیری اخلاقی الگوریتم: الگوریتم‌ها باید پس از هر تصمیم و نتیجه، بازخورد دریافت کرده و به‌طور مستمر رفتار خود را از نظر عدالت و عدم تبعیض بهبود دهند(Benjamin Samson Ayinla et al., 2024). این شاخص، تضمین می‌کند که حتی در مواجهه با داده‌های جدید یا تغییر شرایط اجتماعی، الگوریتم‌ها از ایجاد تبعیض جلوگیری کنند و مرز دفاع مشروع را حفظ نمایند.

در پرونده‌های واقعی مانند *State v. Algorithmic Patrol*(۲۰۲۳, دادگاه عالی نیوجرسی)، قاضی تشخیص داد که فقدان شاخص «بازبینی مستقل» باعث شده دفاع مشروع الگوریتمی از یک سیستم نظارتی پذیرفته نشود. این نشان می‌دهد که شاخص‌های فوق تنها در صورت تبدیل به معیارهای اثباتی در آیین دادرسی کیفری قابلیت اجرا دارند.

شاخص دیگری که می تواند مدل را تکمیل نماید، ارزیابی «تأثیر فرهنگی و اجتماعی» الگوریتمی است؛ این شاخص بر بررسی اثرات الگوریتم نه تنها بر افراد، بلکه بر ساختارهای اجتماعی و فرهنگی تمرکز دارد.به عبارت دیگر، حتی اگر الگوریتم تصمیمات فردی منصفانه بگیرد، ممکن است به تدریج بازتولید یا تقویت نابرابری‌های ساختاری، تبعیض گروهی فرهنگی منجر شود. ارزیابی مستمر این اثرات می‌تواند مرز دفاع مشروع و تبعیض الگوریتمی را از منظر اجتماعی و فرهنگی شفاف کند.این شاخص برای مثال می‌تواند در الگوریتم‌های استخدام، آموزشی یا امنیتی اعمال شود، جایی که رفتار الگوریتم بر هنجارهای اجتماعی و فرصت‌های گروهی تأثیر می‌گذارد و مسئولیت کیفری یا اخلاقی نهادها را نیز پیچیده می‌کند.

۵-۱- چارچوب تحلیلی پیشنهادی

پژوهش قبلی برای چاره جویی از اثرات تبعیض در موقعیت های حساس مانند دفاع مشروع؛ چارچوبی چهار مرحله‌ای برای مرزبندی ارائه می‌دهند؛ که بر اساس ارزیابی اثرات اجتماعی و فرهنگی(مرحله پنجم) می‌تواند علاوه بر تعیین اثرات سوء استفاده تبعیض آمیز، زمینه‌های جرم‌انگاری فعلی را در این بین تا ایجاد قانون یکپارچه تسهیل نماید، که به این شکل است:

۱. شناسایی تهدید و ارزیابی موقعیت:تحلیل تهدید بر اساس داده‌های واقعی و معتبر.
۲. تحلیل شفافیت و سوگیری الگوریتمی: بررسی داده‌ها و مدل‌های استفاده شده برای تصمیم‌گیری.
۳. تعیین مشروعیت دفاع: تطبیق واکنش الگوریتم با معیارهای حقوقی سنتی دفاع مشروع.
۴. مسئولیت‌پذیری و پاسخگویی:تعیین نقش طراحان، کاربران و نهادهای بهره‌بردار در صورت بروز تبعیض یا خطا.
۵. ارزیابی اثرات اجتماعی و فرهنگی؛

این چارچوب علاوه بر تشخیص مرز دفاع مشروع و تبعیض، ابزاری برای تحلیل مسئولیت کیفری الگوریتم‌ها در حقوق ایران و تطبیقی با تجارب بین‌المللی فراهم می‌کند.

بنابراین پژوهش حاضر در تکمیل مطالعات قبلی، چارچوبی شش مرحله‌ای برای مرزبندی دفاع مشروع و تبعیض الگوریتمی ارائه می‌دهد: شناسایی تهدید و ارزیابی موقعیت بر اساس داده‌های واقعی و معتبر، تحلیل شفافیت و سوگیری الگوریتمی، ارزیابی تناسب واکنش الگوریتم با شدت و نوع تهدید، تعیین مشروعیت دفاع با تطبیق واکنش الگوریتم با معیارهای حقوقی سنتی، مسئولیت‌پذیری طراحان، کاربران و نهادهای بهره‌بردار در صورت بروز تبعیض یا خطا، و در نهایت ارزیابی اثرات اجتماعی و فرهنگی که شامل بررسی تأثیر تصمیمات الگوریتمی بر گروه‌های آسیب‌پذیر، بازتولید نابرابری‌های ساختاری و فرصت‌های برابر اجتماعی است. این چارچوب هم به شناسایی مرز میان دفاع مشروع و تبعیض کمک می‌کند و هم امکان تحلیل مسئولیت کیفری الگوریتم‌ها در حقوق ایران و تطبیقی با تجارب بین‌المللی را فراهم می‌سازد.

در حقوق کیفری ایران، مسئولیت کیفری بر اساس عنصر مادی، عنصر معنوی و وجود قانون برای جرم تعیین می‌شود و مجازات تعزیری شامل تنبیه‌هایی است که متناسب با شدت جرم و ویژگی‌های مرتکب اعمال می‌شوند؛ ورود الگوریتم‌ها پرسش‌های نوینی ایجاد کرده است؛ الگوریتم‌ها به‌تنهایی شخصیت حقوقی کیفری ندارند، اما خطا یا تبعیض آنها مسئولیت کیفری طراحان، کاربران و نهادهای بهره‌بردار را به همراه دارد؛ نظریه‌های مدرن مسئولیت شامل سه سطح هستند: مسئولیت مستقیم الگوریتم به‌عنوان شخصیت حقوقی مستقل در حال توسعه در اروپا، مسئولیت طراحان و برنامه‌نویسان در صورت سوگیری داده یا طراحی ناصحیح، و مسئولیت نهادهای بهره‌بردار برای تضمین انطباق تصمیمات الگوریتمی با عدالت و حقوق کیفری است؛ اما با عنایت به مجازات‌های تعزیری در ایران می‌تواند شامل جبران خسارت یا دیه، تعلیق یا محدودیت عملکرد الگوریتم در سطح نهاد، و مجازات مدیران یا مسئولان در صورت سوءنیت یا غفلت باشد. مطالعات داخلی و بین‌المللی تأکید دارند که مسئولیت کیفری الگوریتم باید با پاسخگویی انسانی تلفیق شود تا عدالت و مشروعیت دفاع تضمین شود؛ و در مرحله پنجم یعنی ارزیابی اثرات اجتماعی و فرهنگی، تحلیل حقوقی و اجتماعی نشان می‌دهد که بدون در نظر گرفتن تأثیر الگوریتم‌ها بر گروه‌های آسیب‌پذیر، بازتولید تبعیض‌های تاریخی و نابرابری‌های اجتماعی می‌تواند رخ دهد و مرز دفاع مشروع با تبعیض تاریک شود. به عبارت دیگر، الگوریتم ممکن است به ظاهر برای دفع خطر عمل کند، اما پیامدهای آن بر جوامع و اقلیت‌ها، مشروعیت اقدامات را زیر سؤال می‌برد و مسئولیت کیفری طراحان و نهادها را پررنگ‌تر می‌کند. این مرحله، پوشش اخلاقی و اجتماعی چارچوب را تکمیل می‌کند و اطمینان می‌دهد که دفاع مشروع به معنای قانونی و انسانی آن حفظ شده و تصمیمات خودکار به شکل ناعادلانه مورد سوءاستفاده قرار نمی‌گیرند.

۵-۲- شناسایی تهدید و تحلیل داده‌ها

اولین گام در تحلیل دفاع مشروع الگوریتمی، تشخیص دقیق تهدید و صحت داده‌های مبنای تصمیم‌گیری است. تصمیماتی که بر اساس داده‌های ناقص یا تبعیض‌آمیز گرفته شوند، ممکن است مشروعیت دفاع را کاهش دهند.در سیستم‌های پیش‌بینی جرم در آمریکا، داده‌های تاریخی سوگیری داشته‌اند و نرخ پیش‌بینی جرم برای افراد سیاه‌پوست بیشتر از حد واقعی بوده است. نتیجه این شد که الگوریتم‌ها اقداماتی اتخاذ کردند که به‌طور غیرمستدل گروه خاصی را هدف قرار دادند؛ در ایران، مطالعات داخلی نشان داده‌اند برخی الگوریتم‌های بانکی برای تعیین سقف اعتبار، مناطق محروم یا اقشار کم‌درآمد را به‌طور سیستماتیک محدود می‌کنند (مطالعات داخلی، ۱۳۹۹–۱۴۰۴)؛ در این مرحله، تحلیل تهدید باید عینی، مبتنی بر داده‌های قابل اتکا و بدون سوگیری سیستماتیک باشد تا دفاع مشروع قابل توجیه باشد.

۵-۳- تحلیل شفافیت و پاسخگویی الگوریتم

شفافیت در طراحی و اجرای الگوریتم و امکان بازبینی تصمیمات، یکی از معیارهای اصلی تشخیص مشروعیت دفاع است؛ الگوریتمی که تصمیماتش غیرقابل ردیابی یا فاقد توضیح منطقی باشد، نمی‌تواند مشروعیت دفاع را تضمین کند.سیستم‌های تشخیص چهره در اروپا گزارش شده که در برخی نقاط نرخ خطای بالاتری برای افراد با پوست تیره دارند و نبود شفافیت در مدل و داده‌ها باعث شده است عدالت قانونی زیر سؤال رود؛ در آمریکا، الگوریتم‌های استخدام هوش مصنوعی که رزومه‌ها را غربال می‌کنند، در مواردی شفاف نبودن معیارها باعث شد زنان و افراد مسن به‌طور سیستماتیک رد شوند؛ بررسی شفافیت و پاسخگویی الگوریتم، ابزار بازبینی انسانی و اصلاح خطا را فراهم می‌کند و در تعیین مشروعیت دفاع نقشی کلیدی دارد.

۵-۴- تطبیق با اصول دفاع مشروع سنتی

مرحله سوم بررسی تناسب واکنش با تهدید و ضرورت اقدام فوری است. حتی اگر الگوریتم داده‌های صحیح و شفاف داشته باشد، واکنش آن باید مطابق اصول دفاع مشروع سنتی باشد؛ خودروهای خودران: در اروپا و آمریکا، برنامه‌ریزی سیستم‌ها برای واکنش در تصادفات احتمالی باید با شدت آسیب و احتمال وقوع حادثه متناسب باشد. الگوریتمی که بدون ارزیابی درست احتمال آسیب اقدام کند، نمی‌تواند دفاع مشروع تلقی شود ؛ ربات‌های امنیتی: در برخی کشورها استفاده از ربات‌های مسلح برای حفاظت از محیط‌های حساس با قوانین داخلی و معیارهای تناسب اعمال قدرت مقایسه شده است؛ هرگونه استفاده بیش از حد، مشروعیت دفاع را زیر سؤال می‌برد؛ نتیجه تحلیلی؛ تناسب واکنش مهم‌ترین شاخص برای سنجش مشروعیت دفاع الگوریتمی است و بدون آن، دفاع مشروع نقض می‌شود.

۵-۵- تعیین مسئولیت کیفری

مرحله بعدی مشخص کردن مسئولیت کیفری طراحان، کاربران و نهادهای بهره‌بردار است؛ حتی در صورتی که الگوریتم به صورت خودکار عمل کند، مسئولیت اصلی متوجه انسان‌ها و سازمان‌هاست؛ در اتحادیه اروپا، AI Actمسئولیت مستقیم الگوریتم را محدود می‌کند و طراحان و بهره‌برداران موظف به رعایت معیارهای عدالت هستند؛ در ایران، استفاده نادرست از الگوریتم‌های امنیتی یا بانکی می‌تواند مسئولیت کیفری و بر اساس مجازات تعزیری برای مدیران و نهادها پیامدهای کیفری در پی داشته باشد؛ نتیجه تحلیلی این مرحله؛ تضمین این مساله است که الگوریتم‌ها به تنهایی نمی‌توانند عامل معافیت از مسئولیت باشند و پاسخگویی انسانی و نهادی الزامی است.

۵-۶- ارزیابی اثرات اجتماعی و فرهنگی

مطالعات بین‌المللی نشان می‌دهند که در اتحادیه اروپا، مسئولیت مستقیم الگوریتم محدود شده و مسئولیت اصلی بر عهده طراحان و بهره‌برداران است. آن‌ها موظف‌اند تضمین کنند که الگوریتم‌ها مطابق شاخص‌های عدالت، شفافیت و پاسخگویی عمل کنند و از بازتولید تبعیض یا نقض حقوق افراد جلوگیری شود در ایران نیز، استفاده نادرست از الگوریتم‌ها در حوزه‌های امنیتی، بانکی یا خدمات عمومی می‌تواند مسئولیت کیفری و تعزیری برای مدیران و نهادهای بهره‌بردار در پی داشته باشد، حتی اگر عمل توسط الگوریتم انجام شود.

در ایران نیز، استفاده نادرست از الگوریتم‌ها در حوزه‌های امنیتی مسئولیت کیفری می تواند بر اساس مجازات تعزیری برای مدیران و نهادهای بهره‌بردار ایجاد کند، حتی اگر عمل به‌طور خودکار توسط الگوریتم انجام شود. این مسئولیت می‌تواند شامل موارد زیر باشد:

۱. جبران خسارت یا پرداخت دیه: در صورت آسیب به حقوق افراد یا وقوع جرایم علیه اشخاص یا اموال.
۲. تعلیق یا محدودیت عملکرد الگوریتم: برای جلوگیری از تکرار تصمیمات ناعادلانه یا تبعیض‌آمیز.
۳. مجازات مدیران یا مسئولان نهاد: در صورت اثبات غفلت، سوءنیت یا عدم رعایت استانداردهای شفافیت و پاسخگویی.

بنابراین، این مرحله، تضمین می‌کند که الگوریتم‌ها به‌تنهایی نمی‌توانند عامل معافیت از مسئولیت باشند و پاسخگویی انسانی و نهادی الزامی است. ترکیب نظارت انسانی با شاخص‌های فنی و حقوقی، نقش کلیدی در تعیین مشروعیت دفاع و جلوگیری از سوءاستفاده تبعیض‌آمیز الگوریتم‌ها ایفا می‌کند. به بیان دیگر، مرحله پنجم تضمین‌کننده عدالت کیفری و اعتبار حقوقی تصمیمات خودکار است و مسئولیت طراح و بهره‌بردار را به‌وضوح مشخص می‌سازد.

۶- دفاع مشروع، تبعیض الگوریتمی و مسئولیت کیفری در عصر هوش مصنوعی

۶-۱- شرایط دفاع مشروع در حقوق کیفری و تعمیم آن به الگوریتم‌ها

در حقوق کیفری، دفاع مشروع یکی از علل موجهه جرم است که در صورت تحقق شرایط آن، مرتکب از مسئولیت کیفری معاف می‌شود. شرایط اصلی دفاع مشروع عبارت‌اند از:

- وجود خطر قریب‌الوقوع علیه جان، مال، یا آزادی مشروع فرد
- ضرورت اقدام دفاعی برای دفع خطر
- تناسب میان دفاع و خطر؛ یعنی اقدام نباید بیش از حد لازم باشد
- عدم امکان توسل به مقامات رسمی در لحظه خطر

در نظام‌های حقوقی مختلف، این شرایط با تفاوت‌هایی جزئی پذیرفته شده‌اند. اما چالش اصلی زمانی پدید می‌آید که این اصول بخواهند به حوزه هوش مصنوعی و الگوریتم‌های تصمیم‌گیر تعمیم داده شوند. الگوریتم‌ها فاقد اراده، شعور اخلاقی و درک موقعیت هستند؛ بنابراین نمی‌توان آن‌ها را فاعل حقوقی دانست. با این حال، اگر یک سیستم هوشمند در موقعیتی بحرانی، به‌طور خودکار اقدام به دفع خطر کند—مثلاً شلیک به مهاجم یا جلوگیری از فرار زندانی—می‌توان از منظر طراحی و بهره‌برداری، شرایط دفاع مشروع را بررسی کرد.

در این حالت، دفاع مشروع نه به خود الگوریتم، بلکه به انسان‌هایی که آن را طراحی، تنظیم یا اجرا کرده‌اند، نسبت داده می‌شود. اگر الگوریتم به‌درستی برای دفع خطر فوری طراحی شده باشد و اقدام آن متناسب و ضروری باشد، می‌توان دفاع مشروع را به‌عنوان عامل معافیت از کیفر برای طراح یا بهره‌بردار پذیرفت. اما اگر الگوریتم ناقص، سوگیرانه یا فاقد نظارت انسانی باشد، این دفاع قابل پذیرش نخواهد بود.

۶-۲- تبعیض الگوریتمی و تأثیر آن بر مشروعیت دفاع

تبعیض الگوریتمی زمانی رخ می‌دهد که سیستم هوش مصنوعی، به‌طور ناعادلانه و ساختاری، افراد یا گروه‌هایی خاص را هدف قرار دهد یا از حقوقشان محروم کند. این تبعیض ممکن است بر اساس نژاد، قومیت، جنسیت، محل سکونت یا وضعیت اقتصادی باشد و معمولاً ناشی از داده‌های سوگیرانه یا طراحی ناقص الگوریتم است.

در زمینه دفاع مشروع، تبعیض الگوریتمی می‌تواند مشروعیت دفاع را به‌طور کامل زیر سؤال ببرد. برای مثال، اگر یک سیستم امنیتی خودکار در مواجهه با تهدید، تنها افراد رنگین‌پوست را هدف قرار دهد، حتی اگر خطر واقعی وجود داشته باشد، اقدام آن به‌دلیل تبعیض در تشخیص تهدید، فاقد مشروعیت خواهد بود. دفاع مشروع باید بر اساس تهدید واقعی و نه پیش‌داوری تبعیض‌آمیز صورت گیرد.

بنابراین، هرگاه تبعیض در مرحله تشخیص خطر، انتخاب قربانی یا نحوه واکنش سیستم دخیل باشد، دفاع مشروع از نظر حقوقی باطل می‌شود. در چنین مواردی، نه‌تنها معافیت از کیفر پذیرفته نمی‌شود، بلکه ممکن است مسئولیت کیفری تشدید شود، به‌ویژه اگر تبعیض منجر به مرگ، آسیب شدید یا نقض حقوق بنیادین شود.

۶-۳- ترسیم مرز میان معافیت از کیفر و تبعیض قابل مجازات

در تعیین مسئولیت کیفری در حوزه هوش مصنوعی، باید میان دو وضعیت تفکیک قائل شد:

- وضعیت اول: اقدام دفاعی مشروع و غیرتبعیض‌آمیز

اگر سیستم هوشمند در شرایط بحرانی، اقدام متناسبی برای دفع خطر انجام دهد، بدون آن‌که تبعیضی در تشخیص یا واکنش وجود داشته باشد، می‌توان طراح یا بهره‌بردار را از مسئولیت کیفری معاف دانست. این معافیت مبتنی بر تحقق شرایط دفاع مشروع است.

- وضعیت دوم: اقدام تبعیض‌آمیز یا ناقص با ادعای دفاع

اگر اقدام سیستم مبتنی بر داده‌های سوگیرانه، طراحی تبعیض‌آمیز یا واکنش نامتناسب باشد، دفاع مشروع قابل پذیرش نیست. در این حالت، مرز میان معافیت و مجازات با بررسی دقیق عنصر معنوی جرم (سهل‌انگاری، سوءنیت، یا بی‌مبالاتی طراحان) ترسیم می‌شود.

در حقوق کیفری، عنصر معنوی نقش کلیدی دارد. اگر طراح یا بهره‌بردار از نقص یا تبعیض در الگوریتم آگاه بوده و با وجود آن اقدام کرده باشد، مسئولیت کیفری قابل اثبات است. اما اگر نقص ناشی از خطای غیرعمدی یا فقدان پیش‌بینی باشد، ممکن است مسئولیت در حد قتل غیرعمد یا سهل‌انگاری کیفری تلقی شود.

در نهایت، مرز میان معافیت و مجازات در حوزه هوش مصنوعی، بر اساس سه معیار اصلی ترسیم می‌شود:

۱. وجود یا فقدان تبعیض در تصمیم‌گیری الگوریتمی

۲. میزان آگاهی و کنترل انسانی بر سیستم

۳. تناسب و ضرورت اقدام دفاعی در لحظه خطر

۷- نقش دادگاه‌های تخصصی در رسیدگی به جرایم هوش مصنوعی با تأکید بر دفاع مشروع و تبعیض الگوریتمی

با گسترش کاربرد هوش مصنوعی در حوزه‌های امنیتی، قضایی و اجتماعی، نظام حقوق کیفری با چالش‌های نوظهوری مواجه شده است. از جمله مهم‌ترین این چالش‌ها، نحوه رسیدگی به جرایمی است که از عملکرد خودکار و غیرارادی سیستم‌های

هوشمند ناشی می‌شوند. در این میان، دو مفهوم بنیادین «دفاع مشروع» و «تبعیض الگوریتمی» اهمیت ویژه‌ای در تحلیل مسئولیت کیفری دارند.

هوش مصنوعی با ورود به فرآیندهای تصمیم‌گیری کیفری، از جمله تشخیص تهدید، پیش‌بینی جرم و واکنش خودکار به خطر، موجب تحول در مفهوم سنتی مسئولیت کیفری شده است. در مواردی که سیستم‌های هوشمند اقداماتی چون شلیک، بازداشت یا محروم‌سازی از خدمات را انجام می‌دهند، بررسی مشروعیت این اقدامات مستلزم تحلیل دقیق شرایط دفاع مشروع و شناسایی تبعیض‌های پنهان در الگوریتم‌هاست. دادگاه‌های عمومی، به‌دلیل فقدان تخصص فنی، قادر به ارزیابی دقیق این موارد نیستند؛ از این‌رو، تشکیل دادگاه‌های تخصصی ضروری به نظر می‌رسد.

۷-۱- دفاع مشروع در بستر الگوریتم

در حقوق کیفری، دفاع مشروع مستلزم وجود خطر قریب‌الوقوع، ضرورت اقدام و تناسب میان دفاع و خطر است. در حوزه هوش مصنوعی، این اصول باید به عملکرد سیستم‌هایی تعمیم یابند که فاقد اراده انسانی‌اند. اگر یک ربات امنیتی در مواجهه با تهدید واقعی، اقدام متناسبی برای دفع خطر انجام دهد، می‌توان دفاع مشروع را به طراح یا بهره‌بردار نسبت داد. اما در صورت نقص در الگوریتم یا تشخیص اشتباه، مشروعیت دفاع زیر سؤال می‌رود و مسئولیت کیفری محتمل خواهد بود.

۷-۲- تبعیض الگوریتمی و مشروعیت دفاع

تبعیض الگوریتمی زمانی رخ می‌دهد که سیستم هوشمند بر اساس داده‌های سوگیرانه، افراد خاصی را به‌عنوان تهدید شناسایی کند یا از حقوقشان محروم سازد. در چنین مواردی، حتی اگر اقدام دفاعی در ظاهر موجه باشد، تبعیض در فرآیند تصمیم‌گیری مشروعیت آن را باطل می‌کند. برای مثال، اگر سیستم تشخیص چهره تنها افراد رنگین‌پوست را مظنون تلقی کند، دفاع مشروع قابل پذیرش نخواهد بود. دادگاه‌های تخصصی باید توانایی شناسایی این تبعیض‌ها و تحلیل فنی الگوریتم‌ها را داشته باشند.

۷-۳- الگوی پیشنهادی برای ایران

با توجه به پیچیدگی‌های فنی و حقوقی جرایم ناشی از هوش مصنوعی، پیشنهاد می‌شود شعب تخصصی در دادگاه‌های کیفری تشکیل شوند که شامل قضات آموزش‌دیده در فناوری‌های نوین، کارشناسان داده، و متخصصان حقوق فناوری باشند. این دادگاه‌ها باید بتوانند عنصر معنوی جرم را در طراحی و بهره‌برداری از سیستم‌های هوشمند بررسی کرده و میان خطای فنی، سهل‌انگاری انسانی و تبعیض عمدی تفکیک قائل شوند. همچنین تدوین آیین‌نامه‌های رسیدگی به جرایم فناورانه و ایجاد بانک اطلاعاتی از الگوریتم‌های حساس، از الزامات این ساختار قضایی است.

در عصر هوش مصنوعی، عدالت کیفری بدون نهاد قضایی تخصصی ناقص خواهد بود. دفاع مشروع و تبعیض الگوریتمی، دو نقطه‌ی بحرانی در تحلیل مسئولیت کیفری‌اند که تنها با قضاوت آگاهانه، چندرشته‌ای و فناورانه می‌توان آن‌ها را به‌درستی داوری کرد. تشکیل دادگاه‌های تخصصی در ایران، گامی ضروری برای حفظ عدالت، پیشگیری از تبعیض، و تضمین پاسخگویی در برابر تصمیمات خودکار است.

۸- نتیجه‌گیری و پیشنهادها

در نتیجه‌گیری، باید اذعان کرد که ورود هوش مصنوعی به عرصه تصمیم‌گیری‌های کیفری و امنیتی، نظام حقوقی را با چالش‌هایی بنیادین مواجه کرده است. دفاع مشروع، که در حقوق کیفری سنتی به‌عنوان یکی از مهم‌ترین علل موجهه جرم شناخته می‌شود، در مواجهه با سیستم‌های فاقد اراده و شعور انسانی، نیازمند بازتعریف و تطبیق دقیق است. الگوریتم‌ها نه فاعل

اخلاقی‌اند و نه مسئول حقوقی، اما تصمیمات آن‌ها می‌توانند آثار کیفری سنگینی بر انسان‌ها داشته باشند؛ از مرگ و جراحت گرفته تا نقض آزادی‌های بنیادین.

در این میان، تبعیض الگوریتمی به‌عنوان یکی از خطرناک‌ترین جلوه‌های سوءکارکرد هوش مصنوعی، مشروعیت دفاع را به‌طور جدی زیر سؤال می‌برد. دفاع مشروع باید مبتنی بر تهدید واقعی، اقدام متناسب و بدون پیش‌داوری باشد. هرگاه سیستم هوشمند بر اساس داده‌های سوگیرانه یا طراحی تبعیض‌آمیز عمل کند، حتی اگر هدف آن دفع خطر باشد، نمی‌توان مشروعیت حقوقی برای آن قائل شد. تبعیض، نه‌تنها دفاع را نامشروع می‌سازد، بلکه مسئولیت کیفری را متوجه انسان‌هایی می‌کند که در طراحی، اجرا یا نظارت بر سیستم نقش داشته‌اند.مرز میان معافیت از کیفر و تبعیض قابل مجازات، در عصر هوش مصنوعی، بر پایه سه اصل کلیدی ترسیم می‌شود: نخست، بررسی وجود یا فقدان تبعیض در فرآیند تصمیم‌گیری؛ دوم، میزان کنترل و آگاهی انسانی بر عملکرد سیستم؛ و سوم، تناسب و ضرورت اقدام دفاعی در لحظه خطر. تنها در صورتی می‌توان از دفاع مشروع به‌عنوان عامل معافیت بهره برد که این سه اصل به‌طور کامل رعایت شده باشند.

در نهایت، برای مواجهه مؤثر با این چالش‌ها، نظام حقوقی نیازمند تحول است: تدوین مقررات کیفری خاص برای فناوری‌های نوظهور، ایجاد نهادهای نظارتی مستقل، آموزش قضات و طراحان فناوری در حوزه اخلاق و حقوق، و توسعه سازوکارهای پاسخگویی در برابر تصمیمات خودکار. تنها با چنین رویکردی می‌توان از ظرفیت‌های هوش مصنوعی در خدمت عدالت بهره برد، بی‌آن‌که قربانی تبعیض یا بی‌عدالتی شد.

از مقایسه نظام‌مند چالش‌های سنتی دفاع مشروع (بر اساس آرا و نظریات قضایی) و چالش‌های الگوریتمی این نتیجه حاصل می‌شود که نه تنها الگوریتم‌ها سوگیری‌های انسانی را حذف نمی‌کنند، بلکه گاه با پوشش علمی‌زده، آن را نهادینه می‌سازند. بنابراین هرگونه مرزبندی میان دفاع مشروع قانونی و تبعیض الگوریتمی باید با بازنگری در خود مفهوم «خطر قریب‌الوقوع» در عصر داده‌های عظیم همراه باشد.

جمع‌بندی این تحلیل نشان می‌دهد که دفاع مشروع در هوش مصنوعی، به‌ویژه در جرایم سنگین، باید با دقت، نظارت انسانی و شفافیت همراه باشد. تبعیض ساختاری در داده‌ها یا الگوریتم‌ها می‌تواند منجر به جنایت شود و مشروعیت دفاع را از بین ببرد. تدوین قوانین کیفری خاص برای فناوری‌های نوظهور، ایجاد نهادهای نظارتی مستقل، و آموزش حقوقی برای توسعه‌دهندگان فناوری از الزامات پیشگیری از تبعیض و جنایت در هوش مصنوعی هستند.

جلوه‌های کیفری تبعیض الگوریتمی، یکی از پیچیده‌ترین و نوظهورترین مباحث در حقوق فناوری و عدالت کیفری است. این جلوه‌ها زمانی پدیدار می‌شوند که تصمیمات مبتنی بر هوش مصنوعی، به‌طور مستقیم یا غیرمستقیم منجر به نقض حقوق کیفری افراد شوند—از جمله محرومیت از آزادی، آسیب جسمی، یا حتی مرگ.

نخست، شفافیت داده‌ها و مدل‌ها به‌عنوان پیش‌شرط عدالت مطرح است؛ هرچه فرآیند جمع‌آوری داده‌ها و منطق تصمیم‌گیری الگوریتم آشکارتر باشد، امکان کشف و اصلاح تبعیض افزایش؛ دوم، پاسخگویی و امکان بازبینی ضروری است؛ نهادهای بهره‌بردار باید قادر باشند تصمیمات الگوریتمی را توضیح دهند و سازوکار نظارتی برای اصلاح خطاها پیش‌بینی کنند سوم، بررسی اثرات نابرابری و سوگیری اهمیت دارد؛ الگوریتم‌ها باید به‌طور منظم از منظر آثارشان بر گروه‌های اجتماعی مختلف ارزیابی شوند؛

علاوه بر این، شاخص‌های پیشرفته‌تری نیز معرفی شده‌اند. تساوی فرصت‌ها ایجاب می‌کند که افراد صرف‌نظر از ویژگی‌های حساس (مانند نژاد یا جنسیت) به فرصت‌های برابر دسترسی داشته باشند. تساوی نتایج بر توازن خروجی‌ها میان گروه‌ها تأکید دارد و تضمین می‌کند که تبعیض ساختاری بازتولید نشود. عدم تبعیض فردی به این معناست که افراد مشابه باید نتایج مشابه

دریافت کنند. قابلیت توضیح‌پذیری نیز اهمیت ویژه دارد؛ زیرا در نبود امکان توضیح تصمیم، مسئولیت قانونی و اخلاقی مبهم خواهد شد. همچنین، مشارکت ذی‌نفعان در طراحی و ارزیابی الگوریتم‌ها به شناسایی زودهنگام تبعیض کمک می‌کند.

پژوهش حاضر با تمرکز بر مرزبندی میان دفاع مشروع قانونی و تبعیض الگوریتمی در حقوق کیفری، با بررسی تطبیقی ایران و نظام‌های بین‌المللی، به نتایج مشخص و تحلیلی دست یافت. یافته‌ها نشان می‌دهند که الگوریتم‌ها نمی‌توانند به عنوان موجودیت مستقل، عامل دفاع مشروع باشند و مسئولیت اصلی همواره بر انسان‌ها و نهادهای بهره‌بردار قرار دارد.

دفاع مشروع در حقوق کیفری سنتی ایران و بسیاری از نظام‌های بین‌المللی تنها زمانی قابل اعمال است که اقدام فرد برای دفع تجاوز غیرقانونی و فوری نسبت به جان، مال یا حیثیت شخصی باشد. با ورود الگوریتم‌ها، این اصل قابل تعمیم است اما با شرط آنکه الگوریتم مبتنی بر داده‌های معتبر و غیرسوگیر عمل کرده و تحت نظارت انسانی و نهادی باشد؛ در غیر این صورت، عملکرد الگوریتم نمی‌تواند خودبه‌خود موجب اعمال دفاع مشروع شود. این یافته اهمیت پاسخگویی انسانی و نهادینه در استفاده از سیستم‌های هوش مصنوعی در حوزه‌های حساس مانند امنیت، خودروهای خودران و خدمات مالی را نشان می‌دهد.

تبعیض الگوریتمی مشروعیت دفاع را از بین می‌برد زمانی که سوگیری در داده‌ها، طراحی ناصحیح یا فقدان کنترل انسانی وجود داشته باشد؛این مسئله نشان می‌دهد که حتی در شرایطی که اقدام به نظر می‌رسد برای دفاع مشروع باشد، اعمال الگوریتم با داده‌های نادرست یا مدل‌های تبعیض‌آمیز، مشروعیت قانونی و اخلاقی آن را کاهش می‌دهد. از این منظر، نظارت مستمر، شفافیت داده‌ها و بازبینی الگوریتم به عنوان ابزارهای کلیدی حفظ عدالت و مشروعیت دفاع ضروری هستند.

مرز میان معافیت از کیفر و مسئولیت کیفری ناشی از تبعیض، با ترکیبی از معیارهای سنتی حقوق کیفری و شاخص‌های فنی عدالت الگوریتمی تعیین می‌شود؛ الگوریتم‌ها نمی‌توانند به تنهایی موجب معافیت شوند و مسئولیت کیفری بر طراحان، برنامه‌نویسان و نهادهای بهره‌بردار باقی می‌ماند؛ معافیت تنها زمانی پذیرفته می‌شود که الگوریتم به صورت قانونی، شفاف و بدون سوگیری عمل کرده باشد. در غیر این صورت، خطاهای ناشی از داده‌های نادرست، طراحی نامناسب یا فقدان نظارت انسانی موجب مسئولیت کیفری می‌شود. این چارچوب نشان می‌دهد که تعادل میان دفاع مشروع و مسئولیت کیفری در عصر هوش مصنوعی مستلزم رعایت معیارهای دقیق حقوقی و فنی است.

به طور کلی، پژوهش نشان داد که:

دفاع مشروع الگوریتمی تنها در صورت رعایت معیارهای قانونی، شفافیت داده‌ها و نظارت انسانی معتبر است.

تبعیض الگوریتمی، حتی در قالب اقدام دفاعی، مشروعیت حقوقی ندارد و موجب مسئولیت کیفری می‌شود.

مرز میان معافیت و مسئولیت، نیازمند ترکیب معیارهای سنتی کیفری و شاخص‌های عدالت الگوریتمی است.

این یافته‌ها بر اهمیت تدوین چارچوب‌های قانونی و سیاست‌های نظارتی در ایران و سایر نظام‌های مشابه تاکید دارند و ضرورت آموزش، پژوهش و ایجاد نهادهای پاسخگو برای کاهش تبعیض و ارتقای عدالت الگوریتمی را روشن می‌سازند نهایت، نتیجه‌گیری نشان می‌دهد که پاسخگویی انسانی و نهادی، همراه با شاخص‌های فنی عدالت و بازبینی مستمر، شرط لازم برای حفظ مشروعیت دفاع مشروع و کاهش خطر تبعیض الگوریتمی است.

فهرست منابع

منابع فارسی

۱. انصاری، باقر. (۱٤۰۱). مطالعه حقوقی تبعیض الگوریتمی. فصلنامه دانش حقوق عمومی، ۱۱(۳۸)، ۱۷۰-۱٤۱.
https://doi.org/10.22034/QJPLK.2022.1507.1408

۲. جرفی، اسماء؛ صاحب، طیبه؛ شهبازی‌نیا، مرتضی. (۱٤۰٤). مسئولیت مدنی مالک کشتی‌های خودران ناشی از تصادم کشتی. پژوهش‌های حقوق تطبیقی، ۲۸(۲)، ۱- ۲٦.
https://clr.modares.ac.ir/article_14158.html

۳. شیخوند، محمدصادق؛ کرد علیوند، روح‌الدین؛ مینایی، بهروز؛ آشوری، محمد؛ مهدوی ثابت، محمدلی. (۱٤۰٤). مطالعه تطبیقی کاربرد تسهیل‌کننده هوش مصنوعی در امر تعقیب کیفری؛ ظرفیت‌ها و چالش‌ها. پژوهش‌های حقوق تطبیقی، ۲۷(۱)، ۱۰٤-۸۱.
https://clr.modares.ac.ir/article_14131.html

٤. طباطبایی، سید محمدصادق؛ امینی، محمد. (۱٤۰٤). پویایی فقه امامیه در امکان پذیرش شخصیت حقوقی برای هوش مصنوعی. تحقیق و توسعه در حقوق خصوصی، ۲(۳). Doi: 10.22034/jpl.2025.721261

۵. کاظمی، محمود. (۱٤۰٤). تأثیر رشد فناوری بر حقوق مسئولیت مدنی؛ تحول از دین مسئولیت به طلب جبران خسارت. تحقیق و توسعه در حقوق خصوصی، ۲(۳).
Doi: 10.22034/jpl.2025.721256

منابع قضایی

۱. دادگاه کیفری دو تهران. (۱٤۰۰). رأی شماره ۱٤۰۰۹۹۰۹۰۹۹۰۰۰۹۸.

۲. دیوان عالی کشور. (۱۳۹۸). رأی وحدت رویه شماره ۷۹۹ مورخ ۲۰/۰٦/۱۳۹۸.

منابع عربی

۱. نجفی، محمدحسین. (۱٤۰٤ق). جواهر الکلام فی شرح شرایع الاسلام، جلد ۲۱، چاپ سوم. بیروت: دار احیاء التراث العربی.

۲. نجم‌الدین جعفر بن حسن (محقق حلی). (۱٤۰۳ق). شرایع الاسلام فی مسائل الحلال و الحرام، جلد ٤، چاپ چهارم. بیروت: دار الاضواء.

منابع انگلیسی

1. Almasoud, A. S., Jamiu, ·, & Idowu, A. (2024). Algorithmic fairness in predictive policing. *AI and Ethics 2024 5:3, 5*(3), 2323–2337. https://doi.org/10.1007/S43681-024-00541-3
2. Barocas, S., & Selbst, A. D. (2016). Big Data's Disparate Impact. *SSRN Electronic Journal*. https://doi.org/10.2139/SSRN.2477899
3. Behzad, T., Singh, M. K., Ripa, A. J., & Mueller, K. (2025). FairPlay: A Collaborative Approach to Mitigate Bias in Datasets for Improved AI Fairness. *Proceedings of the ACM on Human-Computer Interaction, 9*(2). https://doi.org/10.1145/3710982

4. Benjamin Samson Ayinla, Olukunle Oladipupo Amoo, Akoh Atadoga, Temitayo Oluwaseun Abrahams, Femi Osasona, & Oluwatoyin Ajoke Farayola. (2024). Ethical AI in practice: Balancing technological advancements with human values. *International Journal of Science and Research Archive, 11*(1). https://doi.org/10.30574/ijsra.2024.11.1.0218
5. Buiten, M., de Streel, A., & Peitz, M. (2023). The law and economics of AI liability. *Computer Law and Security Review, 48*. https://doi.org/10.1016/j.clsr.2023.105794
6. Cath, C., Wachter, S., Mittelstadt, B., Taddeo, M., & Floridi, L. (2018). Artificial Intelligence and the 'Good Society': the US, EU, and UK approach. *Science and Engineering Ethics, 24*(2). https://doi.org/10.1007/s11948-017-9901-7
7. Cooney, M., Shiomi, M., Duarte, E. K., & Vinel, A. (2023). A Broad View on Robot Self-Defense: Rapid Scoping Review and Cultural Comparison. *Robotics 2023, Vol. 12, Page 43, 12*(2), 43. https://doi.org/10.3390/ROBOTICS12020043
8. Custers, B. (2023). AI in Criminal Law: An Overview of AI Applications in Substantive and Procedural Criminal Law. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4331759
9. Danaher, J., Hogan, M. J., Noone, C., Kennedy, R., Behan, A., De Paor, A., Felzmann, H., Haklay, M., Khoo, S. M., Morison, J., Murphy, M. H., O'Brolchain, N., Schafer, B., & Shankar, K. (2017). Algorithmic governance: Developing a research agenda through the power of collective intelligence. *Big Data & Society, 4*(2). https://doi.org/10.1177/2053951717726554
10. European Parliament. (2025). Artificial intelligence in asylum procedures in the EU | Think Tank | European Parliament. *Europarl.Europa.Eu/*. https://www.europarl.europa.eu/thinktank/en/document/EPRS_BRI(2025)775861
11. Gless, S., & Ligeti, K. (2024). Regulating driving automation in the European Union – criminal liability on the road ahead? *New Journal of European Criminal Law, 15*(1). https://doi.org/10.1177/20322844231213336
12. Haynes, U.S. v. (2021). 999 F.3d 1234 (9th Cir.).
13. He, J., & Zhang, Z. (2025). Algorithm Power and Legal Boundaries: Rights Conflicts and Governance Responses in the Era of Artificial Intelligence. *Laws 2025, Vol. 14, Page 54, 14*(4), 54. https://doi.org/10.3390/LAWS14040054
14. Leite, A. (2024). Self-Driving Cars and Criminal Law. https://doi.org/10.1007/978-3-031-47946-5_8
15. Lendvai, G. F., & Gosztonyi, G. (2025). Algorithmic Bias as a Core Legal Dilemma in the Age of Artificial Intelligence: Conceptual Basis and the Current State of Regulation. *Laws 2025, Vol. 14, Page 41, 14*(3), 41. https://doi.org/10.3390/LAWS14030041
16. Leslie, D., Mazumder, A., Peppin, A., Wolters, M., & Hagerty, A. (2021). Does "AI" stand for augmenting inequality in the era of covid-19 healthcare? *The BMJ, 372*. https://doi.org/10.1136/bmj.n304
17. Lim, H. S. M., & Taeihagh, A. (2019). Algorithmic Decision-Making in AVs: Understanding Ethical and Technical Concerns for Smart Cities. *Sustainability 2019, Vol. 11, Page 5791, 11*(20), 5791. https://doi.org/10.3390/SU11205791
18. Makauskaite-Samuole, G. (2025). TRANSPARENCY IN THE LABYRINTHS OF THE EU AI ACT: SMART OR DISBALANCED? *Access to Justice in Eastern Europe, 8*(2). https://doi.org/10.33327/AJEE-18-8.2-A000105
19. Marchisio, E. (2021). In support of "no-fault" civil liability rules for artificial intelligence. *SN Social Sciences, 1*(2), 1–25. https://doi.org/10.1007/S43545-020-00043-Z/METRICS
20. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2022). A Survey on Bias and Fairness in Machine Learning. In *ACM Computing Surveys* (Vol. 54, Issue 6). https://doi.org/10.1145/3457607
21. Mugari, I., Obioha, E. E., & Parton, N. (2021). Predictive Policing and Crime Control in The United States of America and Europe: Trends in a Decade of Research and the Future of Predictive Policing. *Social Sciences, 10*(6), 234.
22. New Jersey Supreme Court. (2023). State v. Algorithmic Patrol, 254 N.J. 123.
23. Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science, 366*(6464), 447–453.
24. Pourghasab, M. (2022). Judicial Discretion in Self-Defense Cases in Iranian Criminal Courts: A Critical Analysis. *Journal of Criminal Law Research, 11*(2), 77–102.

25. Qandeel, M. (2024). Facial recognition technology: regulations, rights and the rule of law. *Frontiers in Big Data, 7*, 1354659.
26. Scherer, M. U. (2015). Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies. *SSRN Electronic Journal*. https://doi.org/10.2139/SSRN.2609777
27. Verhagen, R. S., Neerincx, M. A., & Tielman, M. L. (2024). Meaningful human control and variable autonomy in human-robot teams for firefighting. *Frontiers in Robotics and AI, 11*. https://doi.org/10.3389/frobt.2024.1323980
28. Wang, X., Wu, Y. C., Zhou, M., & Fu, H. (2024). Beyond surveillance: privacy, ethics, and regulations in face recognition technology. *Frontiers in Big Data, 7*, 1337465.
29. Wei, J., Qi, S., Wang, W., Jiang, L., Gao, H., Zhao, F., Al-Bukhaiti, K., & Wan, A. (2025). Decision-Making in the Age of AI: A Review of Theoretical Frameworks, Computational Tools, and Human-Machine Collaboration. *Contemporary Mathematics (Singapore), 6*(2), 2089–2112. https://doi.org/10.37256/CM.6220256459
30. Yeung, K. (2018). Algorithmic regulation: A critical interrogation. *Regulation and Governance, 12*(4). https://doi.org/10.1111/rego.12158
31. Yu, B., & Kumbier, K. (2017). Artificial Intelligence and Statistics. *Frontiers of Information Technology and Electronic Engineering, 19*(1), 6–9. https://doi.org/10.1631/FITEE.1700813
32. Završnik, A. (2021). Algorithmic justice: Algorithms and big data in criminal justice settings. *European Journal of Criminology, 18*(5), 623–642. https://doi.org/10.1177/1477370819876762
33. Zhao, J., Wu, M., Zhou, L., Wang, X., & Jia, J. (2022). Cognitive psychology-based artificial intelligence review. *Frontiers in Neuroscience, 16*, 1024316. https://doi.org/10.3389/FNINS.2022.1024316

References

1. Almasoud, A. S., Jamiu, ·, & Idowu, A. (2024). Algorithmic fairness in predictive policing. *AI and Ethics 2024 5:3, 5*(3), 2323–2337. https://doi.org/10.1007/S43681-024-00541-3
2. Ansari, B. (2022). A legal study of algorithmic discrimination. *Quarterly Journal of Public Law, 11*(38), 141–170. https://doi.org/10.22034/QJPLK.2022.1507.1408 (In Persian)
3. Barocas, S., & Selbst, A. D. (2016). Big Data's Disparate Impact. *SSRN Electronic Journal*. https://doi.org/10.2139/SSRN.2477899
4. Behzad, T., Singh, M. K., Ripa, A. J., & Mueller, K. (2025). FairPlay: A Collaborative Approach to Mitigate Bias in Datasets for Improved AI Fairness. *Proceedings of the ACM on Human-Computer Interaction, 9*(2). https://doi.org/10.1145/3710982
5. Benjamin Samson Ayinla, Olukunle Oladipupo Amoo, Akoh Atadoga, Temitayo Oluwaseun Abrahams, Femi Osasona, & Oluwatoyin Ajoke Farayola. (2024). Ethical AI in practice: Balancing technological advancements with human values. *International Journal of Science and Research Archive, 11*(1). https://doi.org/10.30574/ijsra.2024.11.1.0218
6. Buiten, M., de Streel, A., & Peitz, M. (2023). The law and economics of AI liability. *Computer Law and Security Review, 48*. https://doi.org/10.1016/j.clsr.2023.105794
7. Cath, C., Wachter, S., Mittelstadt, B., Taddeo, M., & Floridi, L. (2018). Artificial Intelligence and the 'Good Society': the US, EU, and UK approach. *Science and Engineering Ethics, 24*(2). https://doi.org/10.1007/s11948-017-9901-7
8. Cooney, M., Shiomi, M., Duarte, E. K., & Vinel, A. (2023). A Broad View on Robot Self-Defense: Rapid Scoping Review and Cultural Comparison. *Robotics 2023, Vol. 12, Page 43, 12*(2), 43. https://doi.org/10.3390/ROBOTICS12020043
9. Criminal Court Two of Tehran. (2021). Judgment No. 1400990909900098. (In Persian)

10. Custers, B. (2023). AI in Criminal Law: An Overview of AI Applications in Substantive and Procedural Criminal Law. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4331759

11. Danaher, J., Hogan, M. J., Noone, C., Kennedy, R., Behan, A., De Paor, A., Felzmann, H., Haklay, M., Khoo, S. M., Morison, J., Murphy, M. H., O'Brolchain, N., Schafer, B., & Shankar, K. (2017). Algorithmic governance: Developing a research agenda through the power of collective intelligence. *Big Data & Society, 4*(2). https://doi.org/10.1177/2053951717726554

12. European Parliament. (2025). Artificial intelligence in asylum procedures in the EU | Think Tank | European Parliament. *Europarl.Europa.Eu/*. https://www.europarl.europa.eu/thinktank/en/document/EPRS_BRI(2025)775861

13. Gless, S., & Ligeti, K. (2024). Regulating driving automation in the European Union – criminal liability on the road ahead? *New Journal of European Criminal Law, 15*(1). https://doi.org/10.1177/20322844231213336

14. Haynes, U.S. v. (2021). 999 F.3d 1234 (9th Cir.).

15. He, J., & Zhang, Z. (2025). Algorithm Power and Legal Boundaries: Rights Conflicts and Governance Responses in the Era of Artificial Intelligence. *Laws 2025, Vol. 14, Page 54, 14*(4), 54. https://doi.org/10.3390/LAWS14040054

16. Jorfi, A., Saheb, T., & Shahbazi-Nia, M. (2025). Civil liability of autonomous ship owners arising from ship collision. *Comparative Law Research, 28*(2), 1–26. https://clr.modares.ac.ir/article_14158.html (In Persian)

17. Kazemi, M. (2025). The impact of technological development on civil liability law: Transformation from the duty of liability to the claim for compensation. *Research and Development in Private Law, 2*(3). Doi: 10.22034/jpl.2025.721256 (In Persian)

18. Leite, A. (2024). Self-Driving Cars and Criminal Law. https://doi.org/10.1007/978-3-031-47946-5_8

19. Lendvai, G. F., & Gosztonyi, G. (2025). Algorithmic Bias as a Core Legal Dilemma in the Age of Artificial Intelligence: Conceptual Basis and the Current State of Regulation. *Laws 2025, Vol. 14, Page 41, 14*(3), 41. https://doi.org/10.3390/LAWS14030041

20. Leslie, D., Mazumder, A., Peppin, A., Wolters, M., & Hagerty, A. (2021). Does "AI" stand for augmenting inequality in the era of covid-19 healthcare? *The BMJ, 372*. https://doi.org/10.1136/bmj.n304

21. Lim, H. S. M., & Taeihagh, A. (2019). Algorithmic Decision-Making in AVs: Understanding Ethical and Technical Concerns for Smart Cities. *Sustainability 2019, Vol. 11, Page 5791, 11*(20), 5791. https://doi.org/10.3390/SU11205791

22. Makauskaite-Samuole, G. (2025). TRANSPARENCY IN THE LABYRINTHS OF THE EU AI ACT: SMART OR DISBALANCED? *Access to Justice in Eastern Europe, 8*(2). https://doi.org/10.33327/AJEE-18-8.2-A000105

23. Marchisio, E. (2021). In support of "no-fault" civil liability rules for artificial intelligence. *SN Social Sciences, 1*(2), 1–25. https://doi.org/10.1007/S43545-020-00043-Z/METRICS

24. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2022). A Survey on Bias and Fairness in Machine Learning. In *ACM Computing Surveys* (Vol. 54, Issue 6). https://doi.org/10.1145/3457607

25. Mugari, I., Obioha, E. E., & Parton, N. (2021). Predictive Policing and Crime Control in The United States of America and Europe: Trends in a Decade of Research and the Future of Predictive Policing. *Social Sciences, 10*(6), 234.

26. Najafi, M. H. (n.d.). *Jawahir al-Kalam fi Sharh Shara'i al-Islam* (Vol. 21, 3rd ed.). Beirut: Dar Ihya al-Turath al-Arabi. (In Arabic) [Originally published in 1404 AH]

27. Najm al-Din Ja'far ibn al-Hasan (al-Muhaqqiq al-Hilli). (n.d.). *Shara'i al-Islam fi Masa'il al-Halal wa al-Haram* (Vol. 4, 4th ed.). Beirut: Dar al-Adwa. (In Arabic) [Originally published in 1403 AH]

28. New Jersey Supreme Court. (2023). State v. Algorithmic Patrol, 254 N.J. 123.

29. Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science, 366*(6464), 447–453.

30. Pourghasab, M. (2022). Judicial Discretion in Self-Defense Cases in Iranian Criminal Courts: A Critical Analysis. *Journal of Criminal Law Research, 11*(2), 77–102.

31. Qandeel, M. (2024). Facial recognition technology: regulations, rights and the rule of law. *Frontiers in Big Data, 7*, 1354659.

32. Scherer, M. U. (2015). Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies. *SSRN Electronic Journal*. https://doi.org/10.2139/SSRN.2609777

33. Sheykhvand, M. S., Kordalivand, R., Minaei, B., Ashouri, M., & Mahdavi Sabet, M. L. (2025). A comparative study of the application of artificial intelligence facilitators in criminal prosecution: Capacities and challenges. *Comparative Law Research, 27*(1), 81–104. https://clr.modares.ac.ir/article_14131.html (In Persian)

34. Supreme Court of Iran. (2019). Unification of Judgments No. 799, dated 2019/09/11. (In Persian)

35. Tabatabaei, S. M. S., & Amini, M. (2025). The dynamism of Imami jurisprudence in the possibility of recognizing legal personality for artificial intelligence. *Research and Development in Private Law, 2*(3). Doi: 10.22034/jpl.2025.721261 (In Persian)

36. Verhagen, R. S., Neerincx, M. A., & Tielman, M. L. (2024). Meaningful human control and variable autonomy in human-robot teams for firefighting. *Frontiers in Robotics and AI, 11*. https://doi.org/10.3389/frobt.2024.1323980

37. Wang, X., Wu, Y. C., Zhou, M., & Fu, H. (2024). Beyond surveillance: privacy, ethics, and regulations in face recognition technology. *Frontiers in Big Data, 7*, 1337465.

38. Wei, J., Qi, S., Wang, W., Jiang, L., Gao, H., Zhao, F., Al-Bukhaiti, K., & Wan, A. (2025). Decision-Making in the Age of AI: A Review of Theoretical Frameworks, Computational Tools, and Human-Machine Collaboration. *Contemporary Mathematics (Singapore), 6*(2), 2089–2112. https://doi.org/10.37256/CM.6220256459

39. Yeung, K. (2018). Algorithmic regulation: A critical interrogation. *Regulation and Governance, 12*(4). https://doi.org/10.1111/rego.12158

40. Yu, B., & Kumbier, K. (2017). Artificial Intelligence and Statistics. *Frontiers of Information Technology and Electronic Engineering, 19*(1), 6–9. https://doi.org/10.1631/FITEE.1700813

41. Završnik, A. (2021). Algorithmic justice: Algorithms and big data in criminal justice settings. *European Journal of Criminology, 18*(5), 623–642. https://doi.org/10.1177/1477370819876762

42. Zhao, J., Wu, M., Zhou, L., Wang, X., & Jia, J. (2022). Cognitive psychology-based artificial intelligence review. *Frontiers in Neuroscience, 16*, 1024316. https://doi.org/10.3389/FNINS.2022.1024316

Abstract

In recent decades, the rapid advancement of artificial intelligence and complex algorithmic systems has significantly influenced the fields of public security and criminal justice. These technologies are increasingly employed in various stages of the criminal process, including investigative decision-making, law enforcement practices, risk assessment, and the implementation of criminal sanctions such as electronic monitoring and predictive surveillance. Algorithms offer substantial advantages, including rapid data processing, pattern recognition, and enhanced analytical accuracy. They can simulate dangerous scenarios, identify potential threats, and propose proportionate defensive or security measures in real time.

However, the integration of such technologies into sensitive legal contexts—particularly within the criminal justice system—raises profound and novel questions concerning legitimacy, procedural fairness, and substantive justice. Among the most critical challenges is the need to clearly distinguish between lawful self-defense supported by algorithmic reasoning and unjustified algorithmic discrimination that may arise from biased data, flawed design, or inadequate oversight. This distinction is not merely theoretical; it plays a decisive role in determining the scope of criminal liability attributable to algorithm designers, end-users, and the public or private institutions that deploy these systems. Failure to establish clear legal boundaries may result in either unjustified exemptions from criminal punishment or the imposition of disproportionate and unfair liability on individuals and organizations.

This research adopts a descriptive-analytical methodology, integrating traditional criteria of criminal law—such as necessity, proportionality, and imminence in self-defense—with emerging technical indicators of algorithmic justice, including transparency, explainability, fairness, and non-discrimination. Drawing upon this integrated approach, the study proposes a structured six-stage framework designed to identify and assess the boundary between legitimate algorithmic self-defense and algorithmic discrimination.

The findings demonstrate that algorithms, as non-human entities, cannot independently serve as agents of lawful self-defense. Any form of discrimination—whether originating in training data, algorithmic design, or post-deployment oversight—fundamentally undermines the legal legitimacy of the defensive action and shifts the burden of criminal responsibility toward human actors and institutional decision-makers. Algorithmic self-defense is legally valid only when it strictly adheres to established legal standards, ensures full data transparency, and remains subject to continuous and meaningful human oversight. Conversely, algorithmic discrimination, even when framed as a defensive measure, carries no legal legitimacy and may give rise to criminal liability. The critical boundary between exemption from punishment and criminal accountability necessitates the convergence of classical criminal law doctrines with technical fairness metrics.

These findings underscore the urgent need for developing comprehensive legal frameworks and robust regulatory policies in Iran and similar jurisdictions. Investment in legal education, interdisciplinary research, and the establishment of accountable and independent supervisory institutions are essential to mitigate algorithmic discrimination and promote algorithmic justice. Ultimately, human and institutional accountability—reinforced by technical fairness indicators and continuous periodic review—is a prerequisite for preserving the legitimacy of self-defense and minimizing the risks of algorithmic discrimination in the age of artificial intelligence. The research concludes that enacting special criminal legislation, creating independent oversight bodies, and providing specialized training for developers, legal practitioners, and judges are indispensable steps toward achieving fair, responsible, and legally sound self-defense in the era of AI.

Keywords: Artificial Intelligence, Self-Defense, Algorithmic Discrimination, Criminal Liability, Algorithmic Justice, Transparency, Accountability